

Internet+ Policies and Issues

B. G. Hultais
Editor

Volume

7

NOVA

INTERNET POLICIES AND ISSUES

**INTERNET POLICIES AND ISSUES,
VOLUME 7**

No part of this digital document may be reproduced, stored in a retrieval system or transmitted in any form or by any means. The publisher has taken reasonable care in the preparation of this digital document, but makes no expressed or implied warranty of any kind and assumes no responsibility for any errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of information contained herein. This digital document is sold with the clear understanding that the publisher is not engaged in rendering legal, medical or any other professional services.

INTERNET POLICIES AND ISSUES

Additional books in this series can be found on Nova's website
under the Series tab.

Additional E-books in this series can be found on Nova's website
under the E-books tab.

INTERNET POLICIES AND ISSUES

**INTERNET POLICIES AND ISSUES,
VOLUME 7**

**B.G. KUTAIS
EDITOR**



Nova Science Publishers, Inc.
New York

Copyright © 2011 by Nova Science Publishers, Inc.

All rights reserved. No part of this book may be reproduced, stored in a retrieval system or transmitted in any form or by any means: electronic, electrostatic, magnetic, tape, mechanical photocopying, recording or otherwise without the written permission of the Publisher.

For permission to use material from this book please contact us:

Telephone 631-231-7269; Fax 631-231-8175

Web Site: <http://www.novapublishers.com>

NOTICE TO THE READER

The Publisher has taken reasonable care in the preparation of this book, but makes no expressed or implied warranty of any kind and assumes no responsibility for any errors or omissions. No liability is assumed for incidental or consequential damages in connection with or arising out of information contained in this book. The Publisher shall not be liable for any special, consequential, or exemplary damages resulting, in whole or in part, from the readers' use of, or reliance upon, this material. Any parts of this book based on government reports are so indicated and copyright is claimed for those parts to the extent applicable to compilations of such works.

Independent verification should be sought for any data, advice or recommendations contained in this book. In addition, no responsibility is assumed by the publisher for any injury and/or damage to persons or property arising from any methods, products, instructions, ideas or otherwise contained in this publication.

This publication is designed to provide accurate and authoritative information with regard to the subject matter covered herein. It is sold with the clear understanding that the Publisher is not engaged in rendering legal or any other professional services. If legal or any other expert assistance is required, the services of a competent person should be sought. FROM A DECLARATION OF PARTICIPANTS JOINTLY ADOPTED BY A COMMITTEE OF THE AMERICAN BAR ASSOCIATION AND A COMMITTEE OF PUBLISHERS.

Additional color graphics may be available in the e-book version of this book.

LIBRARY OF CONGRESS CATALOGING-IN-PUBLICATION DATA

ISSN: 2158-1517

ISBN: 978-1-61728-319-2 (eBook)

Published by Nova Science Publishers, Inc. † New York

CONTENTS

Preface	vii	
Chapter 1	Mobile Middleware Platforms for Content and Service Creation, Management and Delivery: A Technology Classification Framework <i>Antonio Ghezzi</i>	1
Chapter 2	Using Digital Watermarking to Identify Audio and Video Software Products <i>Takaaki Yamada, Yoshiyasu Takahashi, Ryu Ebisawa, Yoshinori Sato and Seiichi Susaki</i>	33
Chapter 3	A Case of Middleware for Information Systems of Public Transport by Road <i>Carmelo R. García, Francisco Alayón and Ricardo Pérez</i>	61
Chapter 4	Instant Messaging in Primary Schools <i>Damian Maher</i>	81
Chapter 5	Search Engine Interfaces <i>Gondy Leroy</i>	97
Chapter 6	Social Motives for Windows Live Messenger Use: Relations to Social Capital Theory and Depressive Personality Styles <i>Craig Ross and Emily S. Orr</i>	111
Chapter 7	Full-Text Search in Electronic Health Records: Challenges and Opportunities <i>David A. Hanauer, Kai Zheng, Qiaozhu Mei and Sung W. Choi</i>	125
Chapter 8	Instant Messaging: Standards, Protocols, Applications, and Research Directions <i>Bazara I.A. Barry and Fatma M. Tom</i>	141

Chapter 9	Instant Messaging Communication: Self-Disclosure, Intimacy, and Disinhibition <i>Joshua Fogel</i>	157
Chapter 10	Digital Watermarking for IPR Protection of Multimedia Contents <i>Sarabjeet Singh Bedi and Shekhar Verma</i>	165
Chapter 11	Pathway Search Engine for Expression Proteomics <i>Consuelo Marín Vicente, David M. Good and Roman A. Zubarev</i>	179
Chapter 12	Biomedical Literature Analysis: Current State and Challenges <i>Maurice H.T. Ling, Christophe Lefevre and Kevin R. Nicholas</i>	189
Chapter 13	Interaction Design Process for Ambient Information in Ubiquitous Space <i>Takuya Yamauchi</i>	229
Chapter 14	Chronologizing Web Pages for Effective Search <i>Sunil Kumar Kopparapu and Arijit De</i>	235
Chapter 15	Towards on Demand IT Service Deployment <i>Jai Dayal, Casey Rathbone, Lizhe Wang and Gregor von Laszewski</i>	249
Chapter 16	A Complete Physical Layer Architecture for High Rate Speech Watermarking on Analog Channels and IP Networks <i>Simone Menci</i>	263
Chapter 17	Broadband Internet Access and the Digital Divide: Federal Assistance Programs <i>Lennard G. Kruger and Angele A. Gilroy</i>	309
Chapter 18	Net Neutrality: Background and Issues <i>Angele A. Gilroy</i>	337
Index		345

PREFACE

This book is part of a series exploring the dynamic universe of the Internet in the 21st century. Collected here are papers discussing a wide range of topics and issues impacting and relating to the Internet such as: text mining in electronic health records; search engine types, interfaces and use in proteomics; computer middleware theory and uses; using digital watermarking classifications and applications in analog channels and IP networks and in identifying audio and video software products; instant messaging in primary schools and social motives; broadband internet access and the digital divide; net neutrality, and others. Such a selection makes this volume important to developing an overview of the key issues in the dynamic and wired world.

The growing complexity of mobile “rich media” digital content and services requires the integration of next generation middleware platform within Mobile Network Operators and Service Providers infrastructural architecture, for supporting the overall process of content creation, management and delivery. The purpose of the research in Chapter 1 is to design a technology classification model for Mobile Content & Services Delivery Platforms – MCSDPs –, the core of Mobile Middleware Technology Providers – MMTPs – value proposition. A three-steps theoretical framework – well grounded on existing literature and gathering information through adopting the multiple case studies research methodology – is provided, which identifies a functional architecture as well as a set of significant classification variables to support the platforms positioning analysis. Afterwards, the model is applied to map the current MCSDP offer presented by a sample of 40 companies, classified as MMTPs, so to test the framework validity and get a valuable insight on the actual “state of the art” for such solutions. The main findings show that existing platforms possess major strengths – e.g. wide content portfolio manageable, integration between mobile and web channels and frequent recourse to SOA and Web Service approach –, while some drawbacks – poor support to context aware and location-based services, verticality and low interoperability of some proprietary products, criticality of content adaptation etc. – are still limiting the solutions effectiveness.

When illegally distributed contents digitally watermarked with the serial numbers of the software products that generated those contents are found, the watermarks help software vendors to determine whether or not their software was illegally re-distributed by licensed users. It is difficult, however, to detect watermarks in content that has been seriously damaged by signal processing. Detection can be improved by using the original content, but the software vendor usually cannot use the original content due to copyright concerns. In

Chapter 2 the authors present two practical software product serialization applications that use a reference signal similar but not identical to the original content: a text-to-speech interface working with audio watermarking and a video encoder working with video watermarking. The authors evaluated them using a previously developed audio and video watermarking prototype based on the patchwork algorithm and by modifying the time and spatial domains of the data. Because the proposed detection method assumes the use of the same embedding method used in the conventional detection methods, it not only detects watermarks without having to use a reference signal but also performs better by using the reference signal.

A practical case of use of the middleware is explained in this chapter. The contents are related with the fields of the ubiquitous computing, and more specifically with the theory and operation principles of the middleware, and the intelligent transport systems. Chapter 3 describes how the theory and operation principles of the middleware can be applied in order to resolve a traditional problem of the information systems of the public transport corporations. Specifically, they explain the design and operation of a middleware, this middleware permits a proper integration of mobile information system of the vehicles of public transport corporation of passengers. In this context, the concept of proper integration means that the all the information related with the vehicles operation is available at time, real time, and the required amount by all the processes of the information system of the transport company. The description will be based on the theoretical and operation principles of the middleware, these principles are: context modeling, spontaneous interaction, context-triggered task management and development system for ubiquitous applications. Other relevant aspect of the middleware consists of that its design rely principles of network administration systems in order to control and administrating on automatic and unsupervised way the mobile information systems. The first point of the chapter is the introduction and it is dedicated to describe the main aspects of the mobile information system in the public transport context. The second point of the chapter is dedicated to describe the technological bases of the middleware, specifically the mobile communication infrastructures and the ubiquitous computing paradigm. The third point is the kernel of the chapter, in this point the formal description of the middleware will be presented. Finally, the authors will explain how to achieve several functionalities, commonly required in the operation of the vehicles of a public transport company, using the middleware.

The use of instant messaging (IM) in primary schools is a recent phenomenon having been around for less than 10 years in most schools with internet access. There is an expectation by Educational authorities, parents, teachers and students that interactive technologies such as IM be included as part of learning experiences. To date there has been very little research examining the use of IM with primary school students.

The use of IM has the potential to change the nature of education by expanding the range of participants with whom students can interact, both while at school and in their homes. Students now have access to experts online and other community members which vastly increases their access to different ideas and opinions. In addition, students can interact with other students who are geographically distant which enables increased cultural awareness. Students are also able to interact with each other, family members and their teachers while at home, which is dissolving the boundaries between school and home.

Access to other participants via IM has brought with it new challenges. In particular, the safety of students online has been a main focus of schools, parents, educational authorities and Governments and is examined in Chapter 4.

More often than not, we turn to the Internet when we need information about products to buy, places to visit, or even doctors to consult. We use search engines to locate information and expect to find answers immediately and without effort. A search engine's interface is a critically important component in this process. Chapter 5 reviews the user query options in general-purpose search engines and the underlying technology used to match that query to web pages. It also describes different query options provided by special-purpose search engines.

General-purpose search engines use a simple interface, a text box, and require only a few keywords to search. Most people use only 2 or 3 keywords, which is very little information, to search among billions of documents. Increasing the number of keywords increases the information available to search and improves the results. There are two approaches to increasing this information: establishing a user profile or using query expansion techniques. User profiles are predominantly used to filter results from a search. Static user profiles are built on information supplied by users about themselves. This information is then used of toward the selection a subset of results. On the whole, it is not very popular with users and too stringent for prolonged use. Dynamically built profiles, requiring no user effort, are continuously updated. However, many users do not like tracking of their behavior. In contrast to filtering results based on a profile, query expansion aims to add additional terms to the original query to make it more precise so that fewer but more precise results are found. A few extra, relevant keywords increase the available information leading to better results. Query expansion, whether it is automated or manual and interactive, generally improves the results and many search engines provide query expansion options as an effortless and dynamic augmentation to their basic search.

Special-purpose and newer search engines provide a different interface. For example, music or image search engines benefit from techniques that use sounds or images in the query. Natural language search engines allow users to type a query in their own words. In their own work, the authors evaluated query diagrams as an input method and found that they are easy to understand and the query itself contains much more information resulting in more precise queries.

The purpose of Chapter 6 is to explore the motivations associated with Windows Live Messenger (Messenger) use within the context of social capital. Furthermore, the present chapter will also explore these motivations in relation to depressive personality styles. Motivations for the use of online technology can be understood within the framework of social capital (Williams, 2006) which proposes that the appropriate use of communication tools can lead to interpersonal and coping benefits for the individual user. The authors' results identified three motivations for Messenger use: Emotion Regulation and Coping, Positive Practicality, and Passing Time. These motivations were consistent with the authors' understanding social capital theory. Further to the authors' understanding of Messenger use from a social capital perspective, the present data supports that these motivations can also be understood within the context of clinically-related personality styles. Blatt (1974) argued that depression stems from one of two personality dimensions: self-criticism or interpersonal dependency. Individuals who exemplify either style are particularly prone to repeated episodes of depression. The relationship between those motives identified above and the personality styles identified by Blatt (1974) will be explored. Taken together, this chapter will explore how the motivations for Messenger use relate to social capital theory as well as within the context of depressive personality.

With the passage of the HITECH Act as part of the American Recovery and Reinvestment Act of 2009, the ubiquitous adoption of electronic health records (EHRs) in the United States is likely to occur in the next few years. However, transforming the storage media of patient data, in itself, does not guarantee desirable quality improvement and cost saving outcomes. The catalyzing effects of EHRs critically rely on the value-augmenting functionalities that could unleash the true power of electronically acquired data.

Unfortunately, while the data are electronic, most EHRs support only rudimentary search capabilities which limits the opportunities for making full use of the data. In Chapter 7, the authors discuss some of the complex issues, and potential solutions, for designing an effective search engine for EHRs. The Commentary is based on the 4 years of experience the authors have had operating a search engine specifically designed for EHRs, referred to as the Electronic Medical Record Search Engine (EMERSE). The authors have found that concepts that work for general-purpose search engines do not necessarily apply to those used for an EHR system, namely [1] ambiguity exists with respect to how to define document ‘match’ when searching patient data; [2] documents should not be retrieved in isolation outside the context of all care episodes for the patient; and [3] ‘stop words’—minor words with little inherent meaning that are often automatically excluded by the search engine software—generally cannot be ignored since many are valid abbreviations (e.g., “AND” = axillary node dissection, “OR” = operating room). Further, medical data are subject to unique privacy restrictions and access to the data is limited based on specific user roles in the health care system.

In this chapter, the authors discuss several use cases for searching data stored in EHRs as well as critical features which they have deployed that have helped to make the authors’ implementation successful—more useful and more usable—at their medical center and beyond. Such features include a vast library of medical synonyms and abbreviations, functionalities accommodating the need to search for a phrase with only a subset of words in a case-sensitive manner (e.g. to distinguish “all” vs. “ALL” = acute lymphoblastic leukemia), and the ability to support basic negation and collaborative search.

Instant messaging has brought an effective and efficient real-time, text-based communication to the Internet community. In addition, most instant messaging applications provide extra functions such as file transfer, contact lists, and the ability to have simultaneous conversations, which strengthens the reliance of wider sectors of users on these applications. In this chapter the authors explore the various attempts to create a unified standard for instant messaging. They show the efforts of organizations such as the Internet Engineering Task Force (IETF) in this regard, in addition to some proprietary solutions. The authors also shed some light on the different types of protocols that are used to implement instant messaging applications. Furthermore, the practical uses of instant messaging are highlighted alongside the benefits that will be reaped by organizations adopting the technology. The authors dedicate some parts of this chapter to review current and future research in the field. Various research trends and directions are discussed to show the impact of instant messaging on users, businesses and the decision making process. Chapter 8 provides an attempt to strengthen the theoretical background behind instant messaging and presents the topic in a systematic way.

Individuals use instant messaging to communicate. Chapter 9 reviews the empirical research from scholarly journals on what is known about the topics of self-disclosure, intimacy, and disinhibition with regard to general instant messaging. The search terms of “(instant message OR instant messaging) AND (self-disclosure OR intimacy OR

disinhibition)" were searched in the databases of Medline, PsycINFO, CINAHL, and Business Source Premier from January 1990 to June 2009. Seven articles were reviewed from studies of adolescents and college students. Instant messaging use is associated with self-disclosure, intimacy, and disinhibition. Research is needed to determine if these associations apply to adults too.

Digital Watermarking technology has been proposed for the implementation of Digital Right Management (DRM) system by establishing ownership right, ensuring authorized access and content authentication to protect the Intellectual Property Rights (IPR). The existing basket of technologies like cryptography secure the multimedia data only during storage or transmission and not while it is being consumed. Digital Watermarking provides an answer to this limitation as the watermark continues to be in the data during its usage. A watermark is designed to permanently reside in the original data, and extraction of this watermark provides the protection of IPR. When the watermark is permanently embedded into digital data at the one hand it may be used for checking whether the data have been modified. On the other hand, the detection of the watermark affects the way it is used in practical application. In watermarking applications, watermark extraction raise security issues and need to be protected from several standard data manipulations and modifications.

The goal of this chapter is to address the theoretical and practical aspects related to watermarking and the issues related to imperceptibility, robustness and security problems in digital contents. The tradeoff between major requirements in watermark embedding is elaborated and examined the existing solutions proposed for the same. Chapter 10 elucidates various aspects of digital watermarking for types of multimedia signals and attacks. The techniques of digital image watermarking in spatial and transform processing domain have also been reviewed in this chapter. The chapter concludes with observations and future directions for researchers to design more robust and secure digital watermarking schemes to address the emerging region of real life applications like medical, telemedicine, insurance, defense, mobile communication and entertainment media, where the need for authentication is often high. This would lead to a basis for design of media security systems for protection of the IPR for digital multimedia contents.

As explained in Chapter 11, proteomics is a high-throughput technology for obtaining information on the identity and the expression levels of proteins in a biological sample. Modern mass spectrometry (MS) combined with liquid chromatography (LC) now routinely yields data on >1000 proteins per hour of analysis. However, interpretation of this high-throughput information has until recently been performed in a reductionist way, with emphasis on a few regulated proteins. Increasingly, expression proteomics data are being interpreted by hypothesis-free analyses of activation levels of signalling pathways. The bioinformatics tool performing this pathway analysis was named Pathway Search Engine (PSE). PSE is a hypothesis-generating tool whose predictions are to be tested and validated by complementary (non-mass-spectrometric) techniques. Typically, the PSE consists of a mass spectrometry data analysis module that converts raw LC-MS data into protein identities and their abundances, and a key node analysis module that maps the proteins onto known signalling pathways, performing an upstream search and assigning each identified regulatory molecule ("key node") a preliminary score, with the pathway score being the sum of key node scores. Finally, the post-processing module performs statistical analysis of the group of identified key nodes or pathways and determines the degree of activation for each, as well as providing statistical significance through computation of the associated p-value.

Advances in molecular biology tools and techniques from the end of the last century had shifted the focus of biomedical research from the study of individual proteins and genes to the interactions within an entire biological systems. At the same time, advanced tools generates large sets of experimental data which required collaborations of groups of biologists to decipher. This resulted in a need to have a diverse research knowledge. However, the amount of published research information in the form of published articles is increasing exponentially, making it difficult to maintain a productive edge. Biomedical literature analysis is seen as a means to manage the increased amount of information – to gather relevant articles and extract relevant information from these articles. In Chapter 12 the authors review the central (information retrieval, information extraction and text mining) and allied (corpus collection, databases and system evaluation methods) domains of computational biomedical literature analysis to present the current state of biomedical literature analysis for protein-protein and protein-gene interactions and the challenges ahead.

The purpose of Chapter 13 is to describe constructing methodology for a distribution system. Ubiquitous system enabled to detect environmental information and human motion from places due to an improvement of sensing technology and the distribution system. Previous ubiquitous technology focused on system configuration for building distribution system related to software engineering. However, current technology for the ubiquitous computing is required to suit to various architectures such as home and public space, and the system expected to be used for art and design. Thus a design process for the system should be flexible for places and usages from the beginning of architecture. This chapter explains the methodology for a middleware in the distributed system.

Search engines have become a part of our e-life. The simplest way to get near to the information that one is looking for is invariably fueled by a search engine. Due to large amount of on-line data, invariably there are multiple pages that satisfy the search criteria and hence ranking is inevitable. Most of the search engines today use some mechanism to rank the result pages and use this to display which search result page goes first. The ranking is based on information, meta or otherwise, that is readily available or easily derivable from the web pages. An important component that the search engine today does not exploit is the "age of the web page" because this temporal information is not available via the web page readily except probably for news type of information which comes usually with a date tag in the text. The "age of the page" dimension can be effectively used by search engines to rank the search results in a chronological order. For example, a search like "Monsoon in Mumbai" in the monsoon period might signify that the user is looking for information on the "current" monsoon situation rather than the highly ranked page, using some criteria, discussing about monsoon. Access to a chronologically ordered display of search results will find definite use. The reason search engines can not provide the chronological rank order is because of the absence of "age of the page" information. In Chapter 14 the authors will elaborate on the need for dimension which helps in ranking web pages in chronological order. The authors investigate and discuss existing and new techniques based on natural language processing which can help in chronologizing web pages.

Complex IT systems allow users to create, organize, and share the users' services and computing resources. As these IT systems become more complex, the more difficult the application deployment process becomes. Deployment, the process of making the application or service available for use, often requires the installation, customization, and configuration of many inter-operable heterogeneous system components. The application developer must

understand the many dependencies and configuration parameters required to enable interoperability between the components.

In Chapter 15, the authors will present the deployment solution for a live complex IT system, the Emergency Services Directory (ESD). Many different technologies exist that attempt automate the application deployment process by allowing a user to describe the dependencies, provide the configuration parameters, and specify the application’s required technologies, such as the operating system, database or Web server technology.

ESD’s deployment solution takes advantage of the benefits provided by virtualisation, and virtual appliances in particular. Each of ESD’s components are wrapped and contained within a virtual appliance image. To automatically and on-demand deploy the virtual appliance image, the authors use the Cyberaide Creative tool, which a tool in the on-going Cyberaide project.

High Rate Speech Watermarking is a simple yet powerful technology, proposed in [1] by K. Hofbauer and G. Kubin, for embedding digital data in speech signals. Its basic working principle is related to two well-known properties of voice signals, usually exploited by vocoders for rate compression and artificial speech synthesis, represented by *linear prediction* and *voicing state*. After the signal undergoes an LPC (Linear Predictive Coding) analysis, the obtained residual (also called voice excitation) is split up into voiced and unvoiced segments by means of a pitch detection algorithm; while the voiced segments cross the system unmodified, the unvoiced ones, thanks to their white noise spectral properties, are easily modified by watermark signal by means of a carefully chosen embedding strategy. The decoding is performed through a very similar scheme, which is advantageous for implementation. Despite this structural simplicity, the system permits us to achieve far higher data rates than previous speech watermarking systems with very limited perceptual impact on voice quality, making possible the implementation of new data channels hidden in voice transporting conversationrelated data services without additional cost for bandwidth. This technology is also interesting for other applications of voice storage and transmission, both analogue and digital.

In [2] the authors showed its efficiency and feasibility for the case study of aircraft authentication in Air Traffic Control (ATC) communications. In Chapter 16 the authors propose a complete architecture for high rate speech watermarking, with simple and efficient solutions to address three main issues, not yet addressed in the first work by original authors, with a satisfying trade-off between performance and complexity: channel coding, synchronization and residual equalization. Still using the ATC scenario as a case study, the authors highlight the advantages of the proposed algorithms and how they can make feasible the implementation of this efficient watermarking principle. The system is also a viable solution for the implementation of added value, analogue communication, high speed side data services or simple transmissions of, e.g., specific information as aircraft position and speed, sensors telemetry, or navigation parameters. In this chapter the authors also show an application example where it is used in conjunction with digital voice encoding and IP-based networks connectivity (Voice over IP) for purposes of authentication or added value services.

As discussed in Chapter 17, the “digital divide” is a term that has been used to characterize a gap between “information haves and have-nots,” or in other words, between those Americans who use or have access to telecommunications technologies (e.g., telephones, computers, the Internet) and those who do not. One important subset of the digital divide debate concerns high-speed Internet access and advanced telecommunications services,

also known as *broadband*. Broadband is provided by a series of technologies (e.g., cable, telephone wire, fiber, satellite, wireless) that give users the ability to send and receive data at volumes and speeds far greater than traditional “dial-up” Internet access over telephone lines.

Broadband technologies are currently being deployed primarily by the private sector throughout the United States. While the numbers of new broadband subscribers continue to grow, studies and data suggest that the rate of broadband deployment in urban and high income areas are outpacing deployment in rural and low-income areas. Some policymakers, believing that disparities in broadband access across American society could have adverse economic and social consequences on those left behind, assert that the federal government should play a more active role to avoid a “digital divide” in broadband access. One approach is for the federal government to provide financial assistance to support broadband deployment in unserved and underserved areas.

Economic stimulus legislation enacted by the 111th Congress includes provisions that provides federal financial assistance for broadband deployment. On February 17, 2009, President Obama signed P.L. 111-5, the American Recovery and Reinvestment Act (ARRA). The ARRA provides a total of \$7.2 billion for broadband, consisting of \$4.7 billion to NTIA/DOC for a newly established Broadband Technology Opportunities Program and \$2.5 billion to existing RUS/USDA broadband programs.

Meanwhile, it is expected that the Obama Administration will ultimately develop a national broadband policy or strategy that will seek to reduce or eliminate the “digital divide” with respect to broadband. It is likely that elements of a national broadband policy, in tandem with broadband investment measures in the American Recovery and Reinvestment Act, will significantly shape and expand federal policies and programs to promote broadband deployment and adoption. A key issue is how to strike a balance between providing federal assistance for unserved and underserved areas where the private sector may not be providing acceptable levels of broadband service, while at the same time minimizing any deleterious effects that government intervention in the marketplace may have on competition and private sector investment.

As congressional policymakers continue to debate telecommunications reform, a major point of contention is the question of whether action is needed to ensure unfettered access to the Internet. The move to place restrictions on the owners of the networks that compose and provide access to the Internet, to ensure equal access and non-discriminatory treatment, is referred to as “net neutrality” and is discussed in Chapter 18. There is no single accepted definition of “net neutrality.” However, most agree that any such definition should include the general principles that owners of the networks that compose and provide access to the Internet should not control how consumers lawfully use that network; and should not be able to discriminate against content provider access to that network. Concern over whether it is necessary to take steps to ensure access to the Internet for content, services, and applications providers, as well as consumers, and if so, what these should be, is a major focus in the debate over telecommunications reform. Some policymakers contend that more specific regulatory guidelines may be necessary to protect the marketplace from potential abuses which could threaten the net neutrality concept. Others contend that existing laws and Federal Communications Commission (FCC) policies are sufficient to deal with potential anti-competitive behavior and that such regulations would have negative effects on the expansion and future development of the Internet.

A consensus on this issue has not yet formed, and the 111th Congress, to date, has not introduced stand-alone legislation to address this issue. However, the net neutrality issue has been narrowly addressed within the context of the economic stimulus package (P.L. 111-5). Provisions in that law require the National Telecommunications and Information Administration (NTIA), in consultation with the FCC, to establish “... nondiscrimination and network interconnection obligations” as a requirement for grant participants in the Broadband Technology Opportunities Program (BTOP).

Chapter 1

MOBILE MIDDLEWARE PLATFORMS FOR CONTENT AND SERVICE CREATION, MANAGEMENT AND DELIVERY: A TECHNOLOGY CLASSIFICATION FRAMEWORK

Antonio Ghezzi^{*}

Politecnico di Milano – Department of Management, Economics and Industrial
Engineering, Piazza Leonardo 32, 20133 Milan, Italy

Abstract

The growing complexity of mobile “rich media” digital content and services requires the integration of next generation middleware platform within Mobile Network Operators and Service Providers infrastructural architecture, for supporting the overall process of content creation, management and delivery. The purpose of the research is to design a technology classification model for Mobile Content & Services Delivery Platforms – MCSDPs –, the core of Mobile Middleware Technology Providers – MMTPs – value proposition. A three-steps theoretical framework – well grounded on existing literature and gathering information through adopting the multiple case studies research methodology – is provided, which identifies a functional architecture as well as a set of significant classification variables to support the platforms positioning analysis. Afterwards, the model is applied to map the current MCSDP offer presented by a sample of 40 companies, classified as MMTPs, so to test the framework validity and get a valuable insight on the actual “state of the art” for such solutions. The main findings show that existing platforms possess major strengths – e.g. wide content portfolio manageable, integration between mobile and web channels and frequent recourse to SOA and Web Service approach –, while some drawbacks – poor support to context aware and location-based services, verticality and low interoperability of some proprietary products, criticality of content adaptation etc. – are still limiting the solutions effectiveness.

Keywords: Design, Theory, Mobile Communications, Mobile Content & Service Delivery Platform, Technology classification model, Multiple case studies, Quality Function Deployment

* E-mail address: antonio1.ghezzi@polimi.it

1. Introduction

In a recent past of the Mobile Content market, when content and services offered by Mobile Network Operators (MNOs) and the first Mobile Content & Service Providers (MCSPs) were quite simple – eg. Short Message Services, monophonic ringtones etc. –, the administration activities were carried out through *ad hoc* “legacy” systems; delivery and billing of services were managed through operators’ SMSCs (Short Message Service Centres). The need of integrated platforms for managing the value added services portfolio was not strongly felt about (ABI, 2006[a], 2006 [b]).

However, the growing complexity and cost of mobile “rich media” digital content , the rise of the off-portal environment, the problems of compatibility with different device models and the necessity of handling articulated billing models forced the MNOs to further develop their legacy systems, thus enhancing their functionalities. Nevertheless, such in-house developed “first generation” Service Delivery Platform proved themselves unable to face the emerging market needs, since they were characterized by a limited number of content formats enabled, few functionalities, a poor management of demand peaks, interoperability with few mobile devices models, low support to application developers and content aggregators and a “vertical silos” approach for each content or service delivered (Karlich et al., 2004; Pavlovsky, Staes-Polet, 2005; Forrester, 2007).

Today, Mobile Content market has evolved to a degree of complexity not any more faceable though unsafe, non-flexible and non-scalable first generation system, and requires the introduction of “second generation” platforms. These solution, offered by Mobile Middleware Technology Providers (MMTPs), are here named “Mobile Content and Service Delivery Platforms” (MCSDPs), and can be defined as middleware platforms combining a wide set of functionalities – consistently aggregated into different modules –, and equipped with network-side and device-side interfaces, thus creating an integrated suite with the purpose of supporting some or each phase of the mobile digital content creation, management & delivery process. Unlike the previous solutions, next generation platforms possess the following characteristics: scalability and flexibility; adoption of open standard and of “best of breed” components; support to multiple relationships with developers and Content Providers (CP); capability of handling a large portfolio of content and services demanded by a wide range of mobile devices; common and reusable interfaces with Business Support Systems and Operation Support Systems (Ericsson, 2006; Forrester, 2007; IRC, 2007; iSuppli, 2007).

The introduction of a MCSDP within operators and service providers’ IT architecture allows to obtain a wide set of benefits, as argued by a vast literature (Sabat, 2002; HP, 2005; Pavlovsky, Staes-Polet, 2005; Brynjolfsson et al., 2006; Ericsson, 2006; Kuo, Yu, 2006; Nordmann, 2006; Peppard, Rylander, 2006; Sur et al., 2006; Forrester, 2007).

Concerning the creation, management and delivery of content and services, second generation platforms grant higher efficiency, higher control on the service lifecycle, lower development costs and shorter time to market; moreover, MCSDP adoption enables the widening of service portfolio, thus leveraging on the “long tail theory” (Anderson, 2004), and making possible to exploit scale and scope economies.

With regard to the operators and service providers technology infrastructure, MCSDP grant some major advantages: the unification of service creation, execution and management environments; the integration of multiple delivery channels; a simplification of interfacing

with third parties; an higher architectural flexibility and scalability; an increased interoperability with legacy systems; and an overall reduction of technological complexity.

Anyway, to grasp the previous benefits, MCSDPs shall be designed according to specific technical concepts and approaches, and shall possess certain key characteristics, that will be later discussed.

Taking from a vast literature review, and from a set of qualitative and quantitative information drawn through a sample of case studies, the purpose of this chapter is to develop a reference model for classifying Mobile Content & Service Delivery Platforms, through the identification of a set of significant technology dimensions or classification variables. The model will have both a descriptive and a normative aim: firstly, it will serve to analyze and describe the MCSDP offer “state-of-the-art”; secondly, it will support the decision making process of platform providers – to drive their choices in terms of platform design and technology elements endowment – and platform customers or external stakeholders – to guide the process of selecting the most suitable product according to their needs –.

Afterwards, the model will be applied to map the current MCSDP offer represented by a sample of 40 companies, classified as MMTP, so to propose a first test to the framework’s validity and get a valuable insight on the actual state of the art for the analyzed solutions.

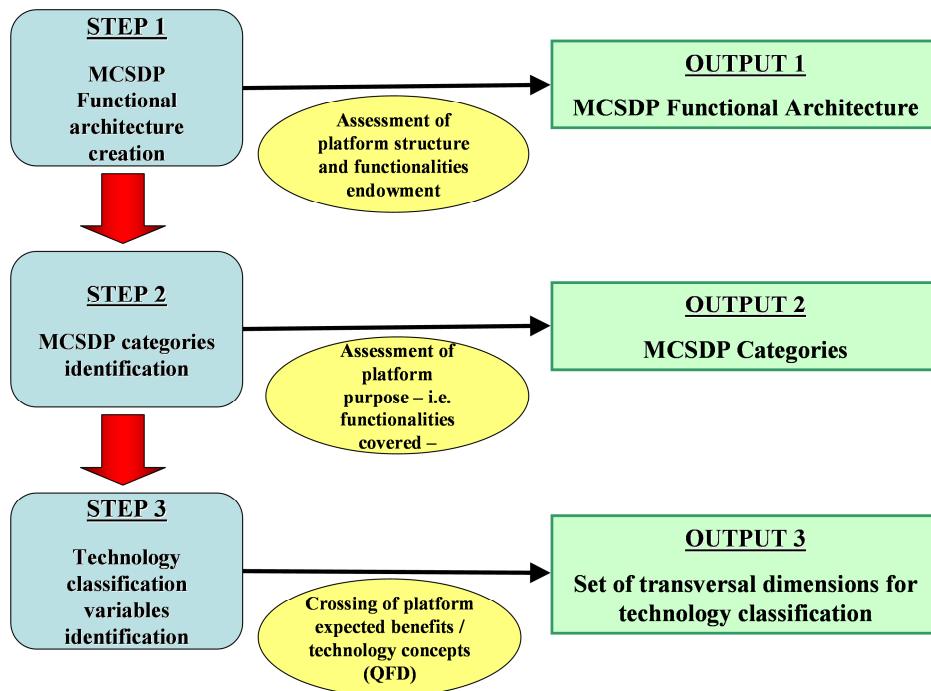


Figure 1. The MCSDP technology classification model main steps and outputs.

2. Methodology

The development of the original MCSDP technology classification model followed three main steps.

At first, taking from a wide literature analysis on the one hand, and from a set of 40 case studies on the other, a thorough MCSDP Functional Architecture was created. Such original architecture will allow to describe the platform's structure in terms of its endowment of functionalities and capabilities, as well as to understand how these functionalities are organized and aggregated into modules.

The second step will employ the functional architecture model to assess the platform's main purposes and objectives, determined by the different combinations of functionalities covered by the platform itself: such approach will lead to the identification of a set of Platform Categories – deducted by the range of functionalities possessed –, which will strongly affect the platform's typology of use.

The third and last step is designed to pair and integrate the previous phases – mainly technology-focused – through the identification of a further set of noteworthy technology dimensions – to be later used as technology classification variables –, complementary and related to the platform functionalities, but directly influencing the platform's performances in terms of benefits achievable by the players introducing the MCSDP within their IT infrastructure.

This fundamental stage addresses the core issue related to linking technology aspects or “Engineering Characteristics” (ECs) to customer desires or “Customer Attributes” (CAs), expressed by the achievable benefits, thus enhancing the model's normative significance by entangling the “voice of the customer” among the dimensions considered, so to guide the Platform Providers towards ensuring customer satisfaction and the consequent product's success on the market.

To integrate this critical stage within the overall classification model, while granting a rigorous approach, the study relies on a technique well grounded in both literature and business practices: the “Quality Function Deployment” (QFD) methodology.

First introduced in Japan in the late 1960s and rapidly adopted in many US industries from the early 1980s (Chan and Wu, 2002), QFD can be defined as “an overall concept that provides a means of translating customer requirements into the appropriate technical requirements for each stage of product development and production (i.e., marketing strategies, planning, product design and engineering, prototype evaluation, production process development, production, sales)” (Sullivan, 1986). Therefore, QFD and the tools it proposes help establishing a clear relationship between marketwise and technology-wise elements.

The use of QFD for the Telecommunications, Electronics and Software Industries is diffused in literature (Haavind, 1989; Wasserman et al., 1989; Brown, 1991; Chang and Lin, 1991; Sharkey, 1991; Bosselman, 1992; Eriksson and McFadden, 1993; LaSala, 1994; Nolle, 1993; Adiano and Roth, 1994; Brown and Harrington, 1994; Philips et al., 1994; Williams, 1994; Sarkis and Liles, 1995; Groenveld, 1997; Kim et al., 1997; Cohen, 1988; Eyob, 1998; Han et al., 1998; Tan et al., 1998; Rosas-Vega and Vokurka, 2000). In the present study, QFD will be employed in the third step of the overall classification model to cross CAs to ECs through the use of a traditional and widely accepted tool: the “House of Quality” (Hauser, Clausing, 1988).

To identify the Customer Attributes representing the expected benefits, a wide literature review was carried out; such review was then integrated with the information gathered through a set of 102 one-to-one, qualitative interviews to platforms' business customers or prospects, performed in a 4 months period (January - April, 2008). The companies sample was made of firms operating in the Mobile Content market or in neighboring areas, which

already owned or were potentially interested in purchasing a MCSDP. The sample covered all the key actor categories whose business can require the adoption of a MCSDP. The main categories the interviewed firms belong to are the following: Mobile Network Operators; Mobile Service Providers; Mobile Content Providers; Media Companies; Web Companies; Web Editors; Device Manufacturers; Software Developers. The sample's completeness and width, as well as the rigor of the process of CAs collection, which relied on a combination of literature analysis and qualitative interviews, grant the significance of the obtained results, dropping the risk of missing relevant attributes.

Though QFD often deals with attributes proposed by end customers, selecting business customers to declare their expected CAs is consistent with previous studies on the subject, since the concept of customer is to be referred to the actor typology who would benefit from the use of the proposed product (Haag et al., 1995), or hold some requirement or need concerning it (Hauser, Clausing, 1988).

One-to-one interviews were preferred to focus group as some evidence in literature (Griffin, Hauser, 1993) show the absence of group synergies expected from focus groups, while their "design cost" is significantly higher. The literature review served as a way to preliminary identify the main expected benefits, but the semi-structure nature of the qualitative interviews allowed the informants to propose their own CAs, thus avoiding to constrain the answers and letting original elements emerge (Yin, 2003).

The obtained CAs were then hierarchically ordered into Primary and Secondary attributes: the Primary level accounted for the macro-benefit expected, while the Secondary level expressed a more detailed version of the primary needs, closer to the customers' own words used to describe them.

To collect both qualitative and quantitative information concerning MMTPs' products, solutions and overall value proposition, meant to feed the first two steps of the model as well as the to support deducting the Engineering Characteristics or technology classification variables that are likely to affect one or more customer need in the third step, the multiple case studies research methodology was employed (Yin, 2003): from January to July, 2008, 40 in-depth case studies – based on 96 both face-to-face and phone semi-structured interviews – on Mobile Middleware Technology Providers were performed, focusing on the set of variables and dimensions identified through the literature analysis. Coherently to the research methodology employed (Pettigrew, 1988), firm sample was not randomly selected, but firms were picked as they conformed to the main requirement of the study, while representing both similarities and differences considered relevant for the data analysis. The main predetermined filters used to discriminate among firms were: the presence of a well-defined line of business – if not the core business – dedicated to the commercialization of Content and Service Delivery Platforms or MCSDP modules; and the presence of an offer directed to the Mobile Telecommunications market. The MMTPs sample can be further divided in International and Local companies, where the former are characterized by a strong cross-country presence, while the latter are prevalently operating in and generating a significant portion of their incomes from a local context (the Italian Mobile Telecommunications market).

The following table provides the full list of the analyzed companies.

A multiple case study approach reinforced the generalizability of results (Meredith, 1998), and allowed to perform a cross analysis on platform characteristics and their combinations – to see which variables changed and which remained constant –, due to the presence of extreme cases, polar types or niche situations within the theoretical sample

(Meredith, 1998). As the validity and reliability of case studies rest heavily on the correctness of the information provided by the interviewees and can be assured by using multiple sources or “looking at data in multiple ways” (Eisenhardt, 1989; Yin, 2003), multiple sources of evidences or research methods were employed: interviews – to be considered the primary data source –, analysis of internal documents, study of secondary sources – research reports, websites, newsletters, white papers, databases, international conferences proceedings –. This combination of sources allowed to obtain “data triangulation”, essential for assuring rigorous results in qualitative research (Bonomo, 1985).

Table 1. Theoretical sample of companies interviewed.

Sample of MMTPs		
Local	International	
Beeweeb	Acotel	Logica CMG
Buongiorno	Aepona – Appium	MBlox
Dylogic	Alcatel Lucent	Microsoft
Interactive Media	Amobee	Nec
Neodata	Bea Systems	Nokia Siemens Networks
Polarix	Comverse	Openwave
Reitek	Drutt	Qualcomm
Reply	Ericsson	Screentonic
Txt Polymedia	First Hop	SPB
Zero9	HP	Sybase 365
	Huawei	Telenity
	IBM	Third Screen Media
	Jamba	Unipier
	Jet Multimedia	Ustream
	Leapstone	Xiam Technologies

As a conclusion, since the technology classification model is developed with the purpose of supporting the classification of MCSDPs offered by MMTPs, an application of the model to the sample of 40 companies analyzed will be hence provided: the platforms will be classified in terms of the functionalities covered and the platform purpose, expressed by the platform category – step 1 and step 2 of the model –, as well as according to the choices made by the platform developers in terms of the further layer of ECs or technology classification variables impacting on the MCSDP performances directly perceived by their users.

3. The MCSDP Technology Classification Model

3.1. The MCSDP Functional Architecture

The integrated assessment of the 40 case studies performed and an academic literature focusing on middleware platforms (Pahlavan, Levesque, 1995; Gaedke et al., 1998; Ma et al.,

2000; Houssos et al., 2000; Kotsopoulos et al., 2001; Metso et al., 2002; Fouial et al., 2002; Chen et al., 2002; Zahariadis et al., 2002; Leavitt, 2003; Li et al., 2003; Moerdijk, Klostermann, 2003; Pailer et al., 2003; Lozinski, 2003; Benali et al., 2004; Karlich et al., 2004; Aioffi et al., 2004; Laakko, Hiltunen, 2005; Capp, Farley, 2005; Barsook, Freedman, 2005; Pavlovsky, Staes-Polet, 2005; Ballon, Van Bossuyt, 2006; Sur et al., 2006; Zhang, 2007; Karlich, 2007), of technical documents elaborated by MMTPs (HP, 2005; Ericsson, 2006; Alcatel, 2006) and of market research reports (ABI, 2006[a]; ABI, 2006[b]; iSuppli, 2007) made possible to build an architectural reference model for a MCSDP.

Such model identifies 7 modules, in turn entangling 48 functionalities or sub-modules, which grant the platform efficiency. In addition, 3 cross-module macro-functions are presented.

The 7 modules and the 48 architectural functionalities can be described as follows.

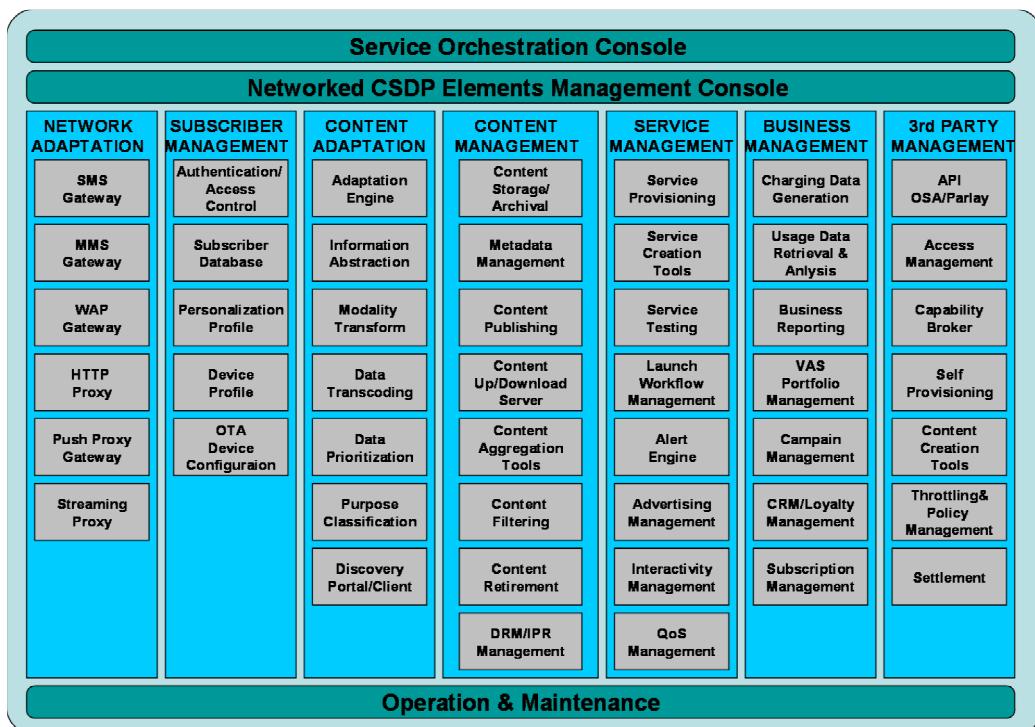


Figure 2. The MCSDP functional architecture.

The “*Network adaptation*” module handles the different activation channels and the main interfaces to different access networks. It is composed by 6 elements.

- SMS Gateway. Gateway to support the distribution of services based on Short Message Service (SMS) delivery technology.
- MMS Gateway. Gateway to support the distribution of services based on Multimedia Message Service (MMS) delivery technology.
- WAP Gateway. Gateway to support the distribution of services based Wireless Application Protocol (WAP).

- HTTP Proxy. Proxy to link the platform to IP networks based on hypertext languages.
- Push Proxy Gateway. Gateway to support the delivery of push services.
- Streaming Proxy. Gateway to support the delivery of streaming services

The “*Subscriber management*” module supports the process of managing the customer base, containing information on preferences and device potentialities – representing the key inputs of content adaptation activities –. It is further divided into 5 components.

- Authentication and Access Control. Functionality enabling the user authentication and access control to the services published. It differs from the Access Management functionality found in the “Third Party Management” module as it is introduced to govern end user accesses to the platform.
- Subscriber Database. Archive where all generic information on the end user pool is stocked.
- Personalization Profile. User profiles, where information and data pertaining to expressed or unexpressed – that is, automatically collected – user preferences and usage data are contained.
- Device Profile. Device profiles, describing and collecting capabilities and characteristics of different end user device models, in order to orient the Content Adaptation phase.
- Over the air (OTA) Device Configuration. Functionality enabling the remote configuration of end user devices, allowing the download of applications and updates.

The “*Content adaptation*” module performs the functionalities of content/services adaptation according to user profiles, device profiles and network capabilities. Content adaptation is widely recognized in literature as one of the key macro-functionalities performed by a MCSDP, so to allow the adaptive content delivery for heterogeneous networks, platforms and devices (Gaedke et al, 1998; Ma et al., 2000; Foial et al., 2002, Chen et al., 2002; Metso et al., 2002; Laakko, Hiltunen, 2005; Zhang, 2007). The module is made of 7 functionalities or sub-modules.

- Adaptation Engine. Engine meant to perform the activity of selecting the proper adaptation algorithm on the basis of the content characteristics and of the device/user profiles, through the application of a set of pre-determined rules and criterion.
- Information Abstraction. Functionality performing the process of content abstraction and compression, whose goal is to reduce the content requirements in terms of scarce resources – e.g. bandwidth – through a data compression, though preserving the most valuable information for the end user.
- Modality Transform. Functionality performing the process of transforming the content media type – e.g. text-to-speech, video-to-image etc. – in order to enable its processing by a given receiver device.

- Data Transcoding. Functionality performing the process of converting data formats according to the receiver device capabilities, to allow a satisfactory content fruition notwithstanding the software or hardware device limitations.
- Data Prioritization. Functionality performing the process of attributing different delivery priorities to different parts of the service, privileging the most relevant information and delaying or deleting the least significant ones.
- Purpose Classification. Functionality aimed at enabling the process of classifying the various elements composing the whole service in terms of their purpose – e.g. advertising banners, links, logos, images etc. –, and consequently eliminating the redundant or useless elements so to increase the service delivery efficiency.
- Discovery Portal – Discovery Client. Functionality meant to identify the capabilities characterizing: the client installed on the user device; or the capabilities of the Portal the content is to be published on.

The “*Content management*” module is dedicated to the end to end handling of digital content published on the platform. It is further divided into 6 functionalities.

- Content Storage/Archival. Sub-module for content storage.
- Metadata Management. Functionality for managing metadata, i.e. information detailing the content published on the platform in terms of content typology, content format, size etc.
- Content Publishing. Functionality for enabling the content provisioning on the platform.
- Content Download Server. Server for uploading and downloading the content, thus supporting Content Publishing.
- Content Aggregation Tools. Tools supporting the content aggregation and bundling activities, so to create an integrated service.
- Content Filtering. Functionality of content control, meant to restrict their accessibility and modification by the platform users.
- Content Retirement. Functionality for retiring and/or deleting obsolete or no longer available content. The content obsolescence can be due several reasons, e.g. its dependence from specific marketing campaigns, or the introduction of updated and substitutive versions; while the unavailability can be related to the expiry of a commercial relation with the Content Provider, which leads to the retirement of its digital products.
- DRM/IPR Management. Sub-module for managing Digital Rights Management and Intellectual Property Rights policies; it allows to control the distribution and duplication of digital products protected by royalties or intellectual property.

The “*Service management*” module deals with the management of value added services – comprising a combination of digital content and/or other applications – offered. It is further composed by 8 functionalities.

- Service Provisioning. Sub-module for supporting VAS provisioning to the end user, through the link to the “Content delivery layer” later described.

- Service Creation Tools. Tools for integrating digital content and software applications in a service to sell to the end customer.
- Service Testing. Functionalities for service testing, which come before the selling phase. Together with the tools supporting the creation of content and services on the one hand, and the launch workflow management on the other, this component has a positive impact on the fast introduction of innovative services, and consequently on their Time to Market (T2M).
- Launch Workflow Management. Sub-module for managing the workflow related to the launch of new value added services.
- Alert Engine. Engine for enabling the delivery of relevant messages and alerts to the services' subscribers.
- Advertising Management. According to the platform's aim and potential, such sub-module is either able to bundle advertising messages externally produced by third parties to the services delivered, or to support the process of advertising content creation and integration as a whole. In the latter case, the platform owner also takes on the role of Mobile Advertising Service Provider (MASP), and handles the advertising creation system (Komulainen et al., 2004).
- Interactivity Management. Functionality for enabling and managing the interaction between the platform and the end user, it handles the two-way channels and grants the possibility for the end user to upload and publish "user-generated content" on the platform itself; these content can be either final products and therefore be published as such, or they can be enriched by other digital content coming from different sources, thus becoming part of a more complex service bundle.
- QoS Management. Sub-module for managing Quality of Service (QoS) and Quality of Experience (QoE) delivered to the end user. It interacts with SLAs contained in the "Settlement" sub-module, as well as with the user profiles memorized on the "Subscriber Management" module, so to grant the full respect of contractual terms, and of the preferences expressed – implicitly or explicitly – by the end customer. QoS is to be considered a noteworthy differentiation element for MCSDP owners, since ensuring a high QoS in dynamics and heterogeneous environments is far from being an easy task, as it requires the coordination of multiple aspects, such as network resources, content and services and user profiles (Koutsopoulou, 2001; Karlich et al., 2004). These considerations enhance the importance of QoS Management sub-module within the MCSDP architecture

The "*Business management*" module encompasses the functionalities related to managing the activities of mobile digital content & services business as a whole. This is made of 5 functionalities:

- Charging Data Generation. Sub-module for processing the information related to service selling transactions and to the subsequent generation of charging reports; such reports are then dispatched towards the Mobile Network Operators' billing systems, to determine the actual payment of the content or service bought by the end customer.

- Usage Data Retrieval & Analysis. Functionality of retrieval and analysis of content or service usage data. It produces the main input for Business Reporting, and proves itself extremely useful as a feedback to design the service offer.
- Business Reporting. Sub-module for managing business reporting to support the decisional process.
- VAS Portfolio Management. Sub-module for managing the overall portfolio of value added service created and delivered through the platform.
- Campaign Management. Sub-module for managing and monitoring content and services distribution campaigns.
- CRM/Loyalty Management. Sub-module providing capabilities for managing the relationship with the customer; it is based on and backed by the legacy Customer Relationship Management (CRM) system.
- Subscription Management. Sub-module for managing subscriptions to content or services.

The “*Third party management*” module supports the relationships with third parties – MNOs, MCSPs and CPs – cooperating with the platform owner. This is further divided in:

- OSA/Parlay API. Set of application interfaces allowing the connection to third parties (Content Providers and developers) systems.
- Access Management. Functionality supporting the platform authentication and access management from business partners in a safe and reliable way.
- Capability Broker. Sub-module for the intermediation and sharing of a set of platform functionalities with partners, according to the level of trust and the intensity of the business interaction.
- Self Provisioning. Functionality to allow partners to deliver themselves content and services created or published on the platform, to be used in other contexts.
- Content Creation Tools. Set of tools to support the content creation activity.
- Throttling and Policy Management. Sub-module meant to handle the allocation of the platform resources and associated capabilities – bandwidth etc. – to partners, as well as the policies and rules regulating the relationships established with third parties.
- Settlement. Sub-module related to all the contractual terms and rules with direct impact on content and service creation and distribution, such as the Service Level Agreements (SLAs) and the Service Quality Cards.

The 3 cross-module functions enable the creation of an integrated and common environment.

1. “Service orchestration console”, leveraging on the concepts of Service Oriented Architecture (SOA), Web Services and IP Multimedia Subsystem (IMS), allows to enhance the efficiency and effectiveness of service management, through the reuse of service components and applications etc.
2. “Platform management & capabilities connectors” handles the processes of service creation, execution and management, and coordinates the interconnections between

the different modules, thus working as an integration layer of the platform's core functionalities.

3. “Operations & maintenance” supports the platform efficiency and maintenance.

The MCSDPs are not operating as separate, monolithic entities, but are obviously part of an overall system. On their Third Party Management side, the platforms interfaces itself to Content Providers and to the wide community of mobile digital content developers; on the Network Adaptation side, the platform relates to the external fixed, mobile and IP network environment, to deliver the content or service to the end customer's device. Moreover, the MCSDP is based on its owner's legacy infrastructure, and interacts with the following key elements.

- Billing and Accounting Systems: they receive the charging reports generated by the platform, and in turn execute the Billing – invoicing and enablement of the actual payment of the digital good purchased by the end customer – and Accounting – division of the revenues generated from selling the digital goods among all the involved players, according to the revenue sharing models in place – activities.
- Customer Relationship Management Systems: they specifically support the integrated management of each and every customer-related activity.
- Enterprise Resource Planning Systems: they support the overall business management.
- Data Base and Data Warehouse: they store transactional or multidimensional information businesswise information.

Within the overall technology classification model, the above described functional architecture will serve as a first tool for providing a MCSDP classification in terms of modules and functionalities covered by the existing platforms: different combinations of functionalities endowments will give rise to a list of different platform categories.

3.2. The MCSDP Functional Architecture

Within the overall technology classification model, the above described functional architecture will serve as a first tool for providing a MCSDP classification in terms of modules and functionalities covered by the existing platforms: different combinations of functionalities endowments will give rise to a list of different platform categories.

As stated earlier, the second step of the MCSDP technology classification model seeks to employ the MCSDP functional architecture to discriminate between the platforms offered by the MMTPs under scrutiny in terms of their main purposes, starting from the assumption that such purposes can be inferred from an evaluation of the modules and functionalities covered – which, in fact, enable the execution of the platforms tasks -. According to the key functionalities offered, it was possible to identify 5 distinct MCSDP categories, characterized by different purposes.

1. *Content Creation platforms.* These MCSDPs' main functionalities are prevalently related to the activities of concept, development and production of the digital content

or service. They offer tools for service creation, workflow management, service testing, as well as for aggregation of internally produced and third parties uploaded content. Anticipating the model's application to the MMTPs sample later illustrated, the figure below shows and example of the functionalities endowment proper of a platform with a prevailing content creation purpose: the product considered is offered by the company Acotel, and the sub-modules and functionalities it covers – deducted from the case studies, with reference to the general Functional Architecture – are evidenced in red.

2. *Content Management platforms.* Such platforms mainly cover the activities spanning from content publishing to content delivery, offering several functionalities: content storage, publishing, aggregation, filtering, retirement; metadata management; digital rights and intellectual property rights management; content adaptation; authentication and access control; user & device profiles management; over-the-air configuration; third parties relationship management. The figure below presents the functional configuration of Xiam Technology's Content Management platform.
3. *Business Management platforms.* The platforms are meant to handle digital content in a wider business perspective, ensuring the integration between the specific VAS business and legacy systems – e.g. BSS/OSS, database and data warehouse, Customer Relationship Management, Enterprise Resource Planning, billing & accounting system. The key functionalities are related to service orchestration, reporting, portfolio and campaigns management, subscribers management. As a real example, Figure 5 shows Ericsson's Business Management platform.

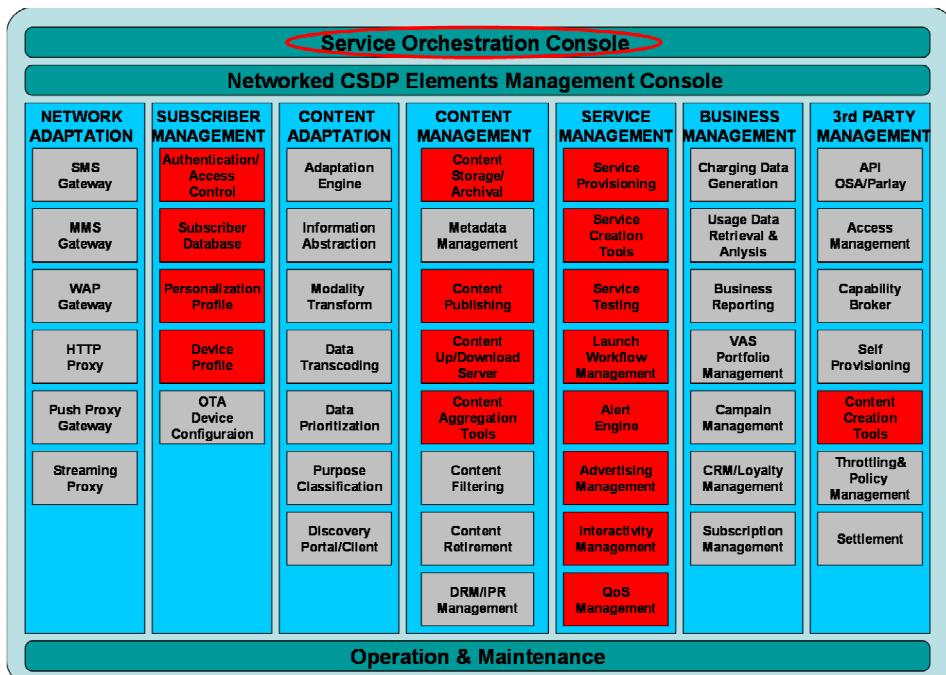


Figure 3. An example of Content Creation Platform.

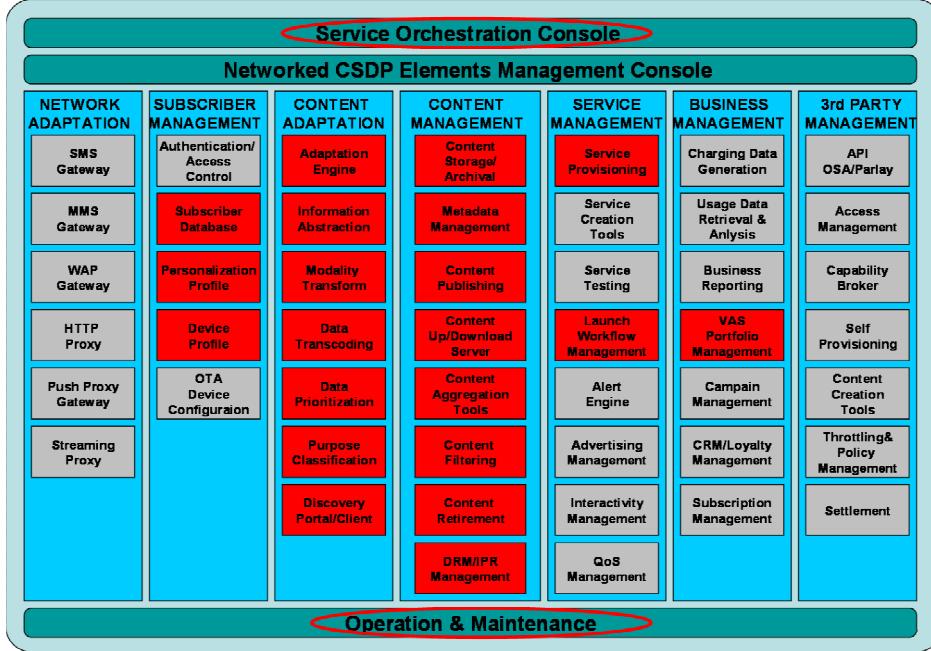


Figure 4. An example of Content Management Platform.

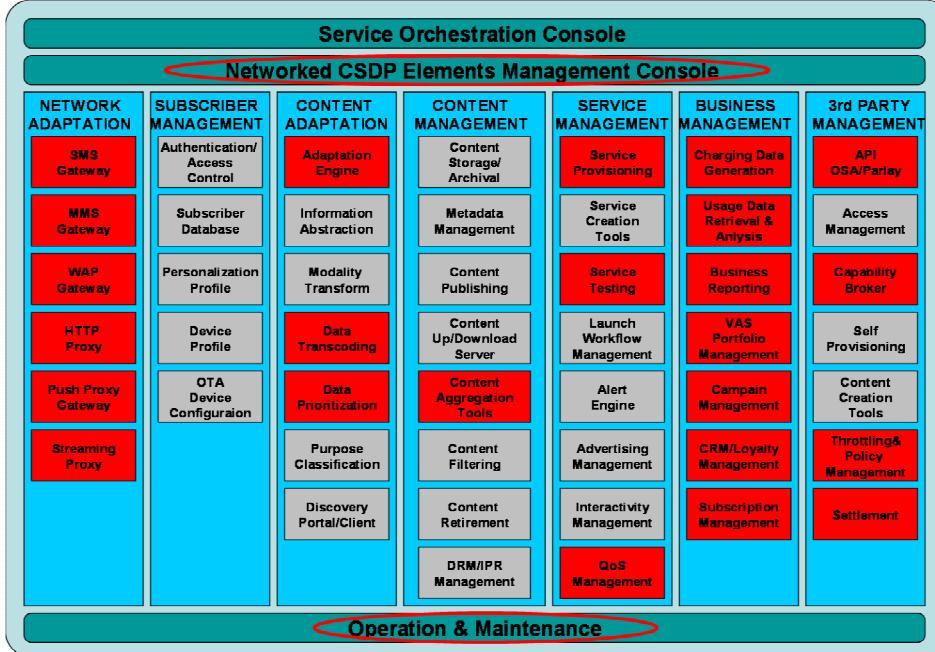


Figure 5. An example of Business Management Platform.

4. *Transactional platforms.* These solutions are interconnected to MNOs' systems, and support the activities related to the so called "CBA process": content charging, content billing, and revenues accounting among the involved parties. These MCSDPs

commonly possess some functionalities of SMS/MMS/WAP-based service delivery. Figure 6 shows MBlox's Transactional platform.

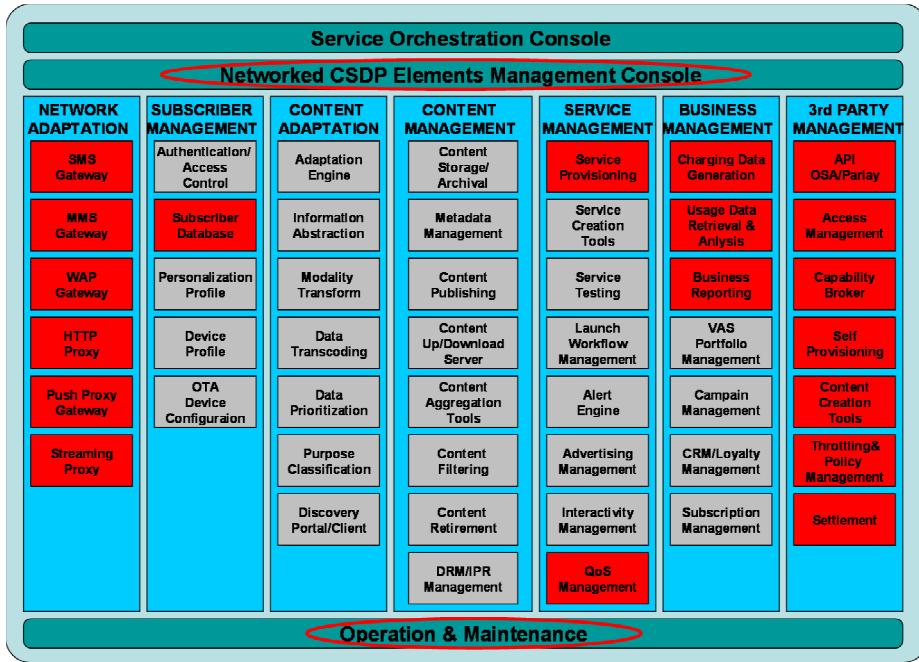


Figure 6. An example of Transactional Platform.

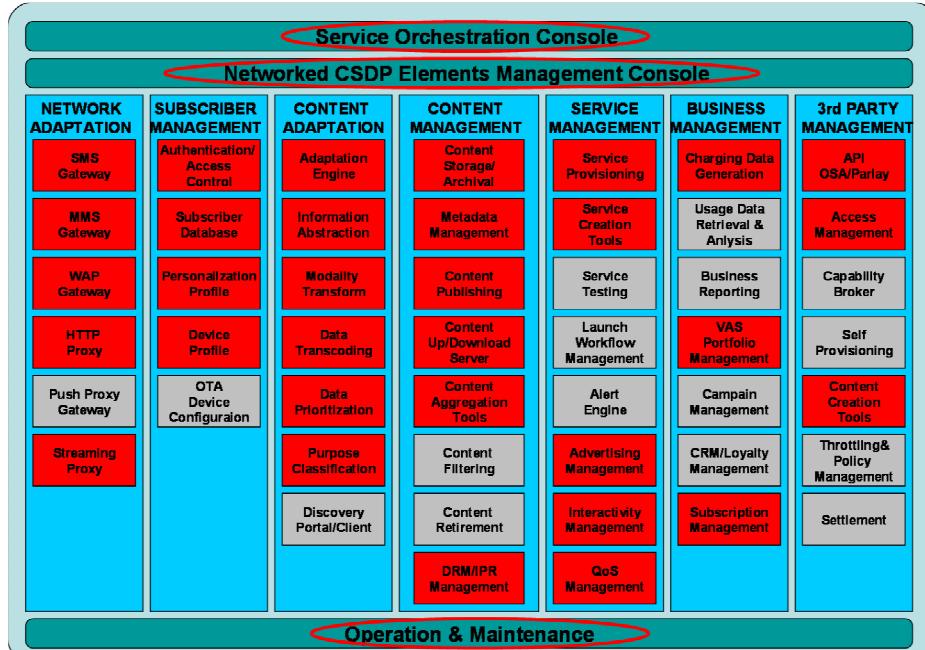


Figure 7. An example of Transactional Platform.

5. *Transversal platforms.* Such MCSDPs show transversal coverage of modules and functionalities, that makes it difficult to identify a prevalent purpose. Figure 7 shows Reitek’s Transversal and multi-purpose platform, with its cross-module functionalities endowment.

3.3. The Identification of the Model’s Technology Classification Variables

The third and last step of the model concerns the identification of a further layer of MCSDP technology dimensions or classification variables that directly influence the platform’s performances, in turn strictly related to the benefits achievable from the platform adoption. The identification of such key technology variables or dimensions is essential to provide the basis for the MCSDP classification and benchmarking processes, as it makes possible to discriminate the technical origins of different platform performances.

Basically, the *rationale* followed to judge a dimension’s significance was its impact on the achievable benefits. A technology concept is relevant if its presence or absence influences, to some or to a large extent, the attainability of expected benefits deriving from the MCSDP introduction, constituting a plus or a drawback for the platform itself – conditions being equal in terms of functionalities endowment –.

In order to grant the rigor of this stage, a classical tool borrowed from Quality Function Deployment literature will be employed: the so called “House of Quality”.

This model allows to cross a product’s Customer Attributes – representing the expected benefits – to a set of Engineering Characteristics – or technology dimensions – which are likely to affect one or more of the CAs (Hauser, Clausing, 1988).

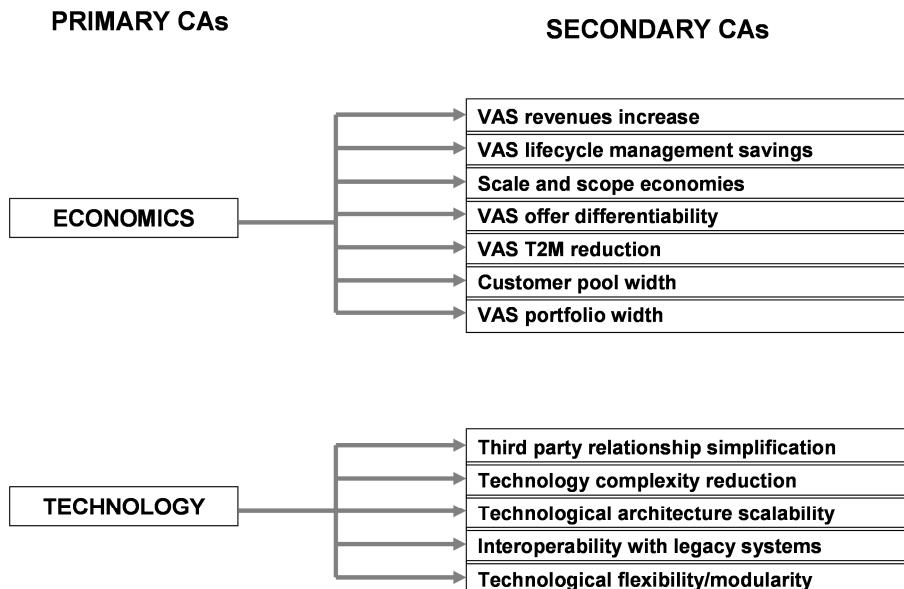


Figure 8. The 2 levels of CAs bundle for a MCSDP.

As widely explained in the methodology section, the MCSDP CAs were derived from a combination of a literature review and of qualitative interviews to platform's business customers. As a whole, the customer requirements were brought back to 2 primary levels:

1. the Economics level, aggregating all the economics-wise and financial-wise benefits;
2. the Technology level, considering all the technology-wise benefits related to the player's technological infrastructure.

In turn, the economics benefits could be further disaggregated into 7 second-level benefits or requirements: Value Added Services revenues increase; Value Added Services lifecycle management savings; scale and scope economies; Value Added Services offer differentiability; Value Added Services Time to Market reduction; customer pool width increase; Value Added Services portfolio width.

At the same time, the benefits bundled in the Technology primary level could be detailed in 5 second-level requirements: third party relationship simplification; technology complexity reduction; technological architecture scalability; interoperability with legacy systems; technological flexibility and modularity.

After the CAs were identified, in order to build the House, the Engineering Characteristics influencing the CAs synthesizing the expected benefits from the MCSDP introduction needed to be listed. Taking into account these adoption benefits pinpointed through the qualitative interviews, and leveraging on a wide technical literature integrated with the data gathered through the 40 case studies, a set of 12 ECs to be used as technology classification dimensions was identified.

1. *Delivery channels available.* Through the “Network adaptation” modules and external network elements like Media Gateways, Media Switch and Media Routers (Li et al., 2003; Aioffi et al., 2004), MCSDP are able to deliver content on multiple channels: fixed, mobile, mobile broadcast (DVBH), IP, wireless, satellite, digital terrestrial tv. The availability of a wide range of delivery channels allows to reach a larger customer base, thus increasing the content selling revenues; moreover, it reduces technology complexity thanks to the unified delivery environment, and makes possible to exploit scale and scope economies in distribution.
2. *Content types treated.* The main rich media digital content categories the platform can manage are: mobile games; video; music; infotainment – micro-browsing, SMS, MMS –; Personalization – logos, wallpapers, ringtones, ring-back tones – (Bertelè et al., 2008). The platform capability of treating the lifecycle of different content types impacts on several benefits, like the enhancement of services management efficacy and efficiency, the widening of service portfolio and potential customers and the increase of revenues coming from Mobile Content.
3. *Media types and formats supported.* Strictly related to the “content types” variable, this dimension assumes great relevance because of the growing multimediality of content, embedding audio, video, images, graphics and messaging (MEIC, 2003). Though the support to multiple media types and formats increases the platform complexity, it positively impacts on the width of the content & services portfolio.

4. *Proprietary vs. Open Source* technology employed. The trade-off here presented is between vertical, end to end proprietary platforms and open standards-based solutions. While the former option is related unique products, hardly replicable by competitors and potentially generating lock-in effects as regard to customers – MNOs and MCSPs –, the latter option makes the platform more flexible and easily interoperable with legacy and third parties systems (Blind, 2005).
5. *Service Oriented Architecture and Web Services adoption.* The introduction of SOA allows to depart from a point-style approach in platform design, ensuring a full connection between BSS/OSS and the platform itself, also allowing the integration of different applications and the reusability of service components, through transversal orchestration functions (Gartner, 2002; Forrester, 2005; IRC, 2007). In addition to this, Web Services grant interoperability between distributed applicative components, representing a service layer the SOA leverages on to access to different content and services and to combine them so to create new applications (Capp, Farley, 2005; Pavlovsky , Staes-Polet, 2005; Sur et al., 2006). Therefore, the adoption of a pervasive SOA and Web services approach impacts on several potential benefits: the increase of efficiency and automation of value added services (VAS) lifecycle management; the services time to market reduction; the widening of offer portfolio; the ability of exploiting scale and scope economies; the reduction of technology complexity, and the architectural flexibility and scalability.
6. *IP Multimedia Subsystem adoption.* IMS can constitute the standard on which to create dedicated architectures for IP multimedia services distribution to end customers (3GPP,2006). The IMS key concepts are close to those proposed by the SOA and Web Services approach (Sur et al., 2006), pushing towards the reuse of applicative components and the creation of a common “control layer” to centralize the management of services published on the MCSDP. This increases efficacy and efficiency in the VAS portfolio management, making the architectural solution more flexible and scalable.
7. *OSA/Parlay Interfaces integration.* OSA/Parlay Application Program Interfaces offer an abstraction of core network functionalities, supporting the interfacing between the platform and third parties systems (Zahariadis et al., 2002; Moerdijk, Klostermann, 2003; ETSI, 2005[a]; Karlich, 2007). Specifically, API Parlay X leverage on Web Services technologies, letting the emergent developers community to easily access network functions and capabilities (ETSI, 2005[b]). Therefore, API influence new services’ time to market reduction and the simplification of the relationships with business partners.
8. *Interactivity and two-way channels availability.* Making interactivity and two-way communication available to end users can increase the service perceived “quality of experience”, also allowing the appealing upload of “user generated content” on the platform – thus making the end customer become a content provider of his own (Pavlosky, Staes-Polet, 2005) -. This can differentiate the platform from competitors’ offers.

9. *Context aware & location-based services enablement.* The possibility of delivering forefront services based on the context of fruition – determined by network capabilities, device profile and user profile – and on the end user geo-spatial location rests on the platform equipment of technologies for “network discovery”, user & device profiles storing and GPS localization (GSMA, 2003). This all enhances the innovativity of the offer, with a potential positive influence on revenues generated – depending on the services uptake –.
10. *Out of the box vs. taylor made solution.* As Porter (2001) asserted, strongly standardized and poorly customizable products bring down both the technology complexity and the offer differentiability; on the contrary, taylor made solutions imply higher development costs, but grant offer uniqueness.
11. *Application Development Platforms supported.* Concerning software technologies allowing the creation and consequent fruition of mobile applications – Sun Microsystems’ J2ME, Qualcomm’s Brew, Macromedia’s Flash Lite, W3C’s SVGT, Streamezzo’s Laser etc. (Barsook, Freedman, 2005) –, it can be argued that supporting a wide range of ADPs positively impacts on the range of content deliverable; however, the proprietary nature of some ADP solutions can make interoperability and third parties relationships more complex.
12. *Mark-up languages supported.* Within the MCSDPs perimeter, mark-up programming languages belonging to the HTML family (IETF, 1995) – XML, XHTML, XSL, SGML, WSDL, PML, SMIL, VXML, SALT, SAML – have two main purposes: first, they represent the codes used for platform development, and ensure the overall technology infrastructure governability; second, they support the creation of multimedia applications. Relying on such languages increases the efficiency in the VAS lifecycle management, making the architectural solution more flexible, scalable and interoperable.

At this stage, the 12 secondary CAs were to be crossed to the 14 ECs – proprietary technology and open source technology were considered as two separate technology characteristics, since their impact on the achievable benefits is dual; the same approach was taken when considering the ECs of out-of-the-box and taylor made solutions – in order to develop the “relationship matrix”, where each and every existing relationship between customer attributes and engineering characteristics are made explicit, and the strength of these relationship is also addressed.

As shown in Figure 9, a “green” crossing of a CA and a EC indicates a positive relationship between the two, implying that the adoption of the technology variable under consideration makes it easier to obtain a given performance: for instance, on the economics benefit side, the availability of many delivery channels allows to reach a wider customer pool, thus gaining higher revenues and benefiting from scale and scope economies on VAS delivery; on the technology side, a Service Oriented Architecture and Web Service based approach grants higher flexibility, scalability, modularity and interoperability with the legacy environment. On the contrary, a “red” crossing stands for a negative relationship, as the technology dimension has a negative influence on that specific customer perceived benefits’

attainability: for instance, an open source-based platform will not catalyze the VAS offer differentiability, as it will become imitable by competitor more easily than a proprietary solution. In some cases, the relationship between CAs and ECs can also be conflicting – represented by a “yellow” crossing –, as the technology dimension affects both positively and negatively the given attribute, therefore making it hard to claim which of the two effects is stronger: this is the case of the variable ADPs supported, when considered in relation to third party relationship simplification and interoperability with legacy.

The House of Quality also serves as a powerful tool to analyze the existing relationships and trade-offs between ECs: the “roof matrix” on top of the relationship matrix allows to highlight how the ECs collaterally affect each other. Again, positive, negative or conflicting relationships are evidenced through green, red and yellow points respectfully. For example, adopting the interactivity dimension is positively related to the platform’s technological capability of delivering innovative context-aware and location based services, while usually it is negatively associated to the development of simple, generic out-of-the-box solutions.

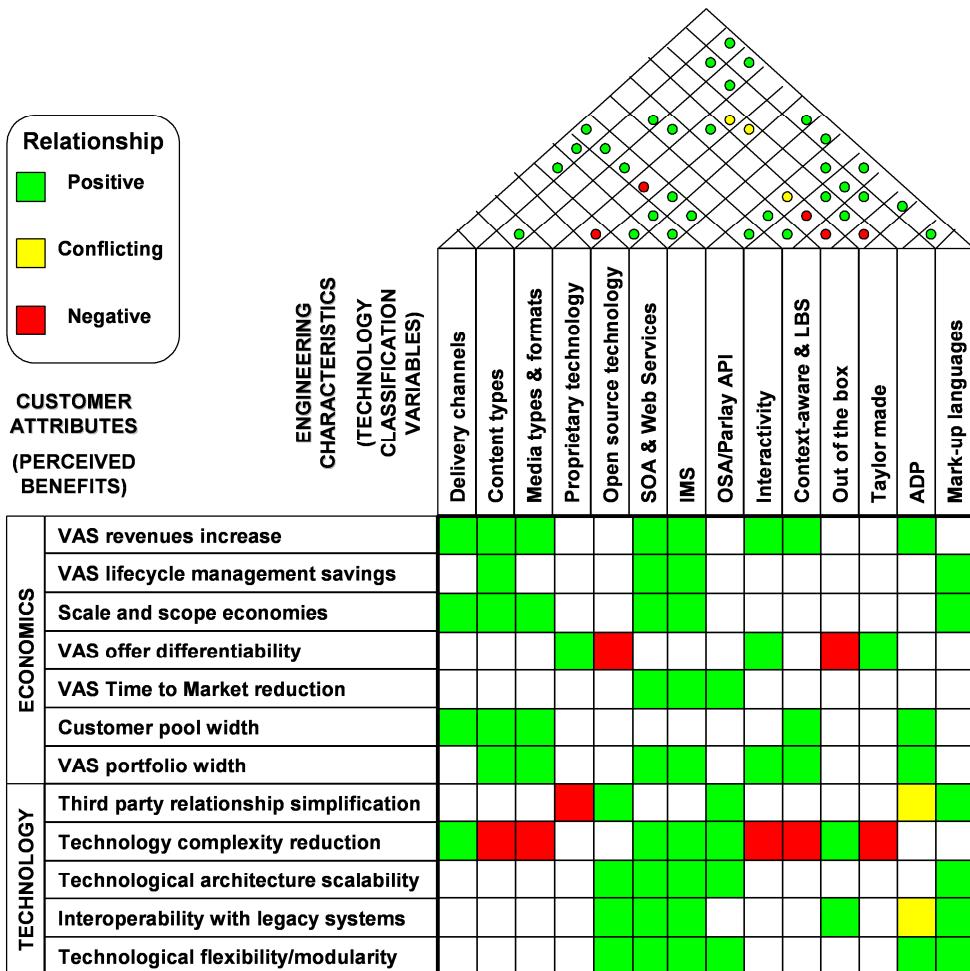


Figure 9. The House of Quality applied to a MCSDP CAs and ECs.

As a whole, the third and last step of the model, through the use of a rigorous QFD tool – the House of Quality – allowed to identify a further set of technology variables or ECs, directly influencing the achievable benefits, to be used as an additional layer for enabling the MCSDP classification.

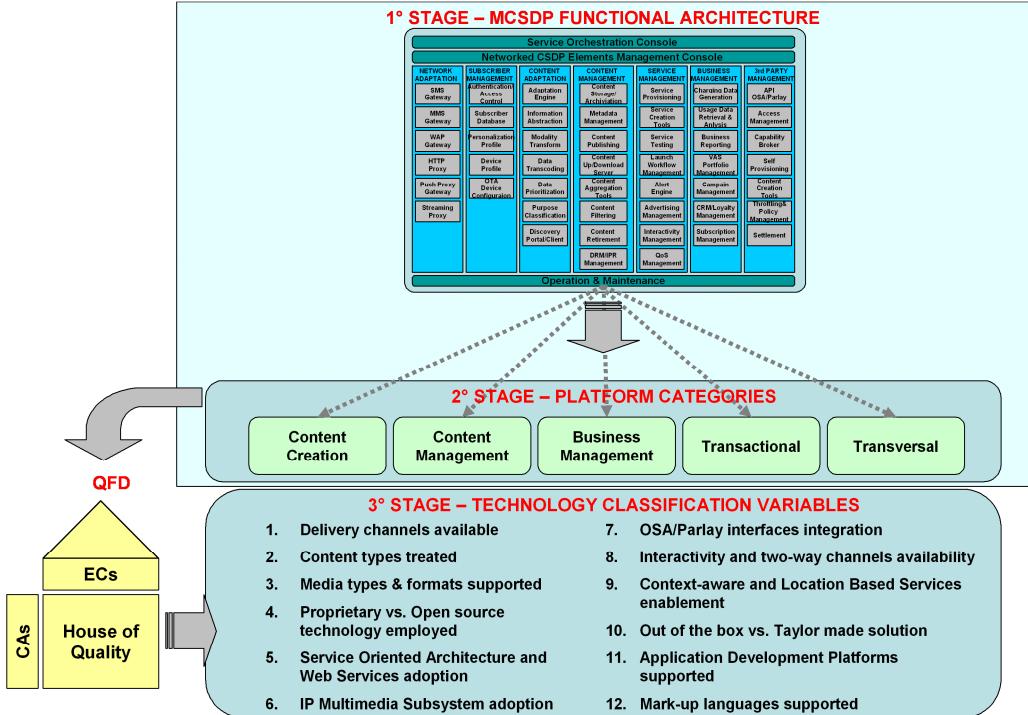


Figure 10. The overall MCSDP Technology Classification Model.

3.4. The Overall MCSDP Technology Classification Model

The overall MCSDP Technology Classification model is synthesized Figure 10. Ontologically, the 3 model's building blocks are strongly connected. In the first stage, a general model of a MCSDP Functional Architecture is provided, portraying the platform endowment in terms of functionalities. The second stage, whose goal is to offer a classification in terms of the platform purposes, directly descends from the first one, as the creation of the 5 key platform categories emerges from a combination of the previously identified architectural functionalities: the platform purpose depends heavily on the functionalities it covers, which allow to operationally perform the required tasks. In the third stage, through the adoption of the House of Quality tool borrowed from QFD theory, the main Customer Attributes expressing the voice of MCSDP customers were employed to define a further set of Engineering Characteristics or technology classification variables: such ECs represent a sub-layer of platform design levers directly affecting the platform's performances in terms of customer's attainable benefits – again, conditions being equal in terms of functionalities endowment –.

After having provided an in-depth description of the model's constitutive building block and core stages, the original framework created will be employed to classify MCSDP solutions, as shown in the next section, concerning the model's application to the MMTPs sample.

4. The Model Application to the Companies Sample

In order to prove the model's validity and descriptive effectiveness, also accomplishing the a major research objective, i.e. to provide a classification of the current MCSDP offer presented by MMTPs, the theoretical model developed will be applied to a real context, represented by the sample of firms analyzed through case studies.

The functional architecture and the platform categories of stages 1 and 2 will serve to categorize the existing MMTP solutions in terms of their main purposes; consequently, the platforms offer will be mapped with reference to the additional sub-layer of technology dimensions, so to assess its coverage of these CAs-impacting variables.

4.1. The Offer Distribution on the Five MCSDP Categories

Through the analysis of the solutions repartition in the 5 platform categories according to their prevailing purpose – in turn expressed by their architectural design, as shown in section 3.2 –, some considerations can be made concerning the emerging clusters and the players populating them.

Table 2. The application of Step 1 and 2 to the MMTPs sample

CONTENT CREATION PLATFORMS	CONTENT MANAGEMENT PLATFORMS	BUSINESS MANAGEMENT PLATFORMS	TRANSACTION PLATFORMS	TRANSVERSAL PLATFORMS
1. ACOTEL 2. OPENWAVE 3. SPB 4. USTREAM 5. ZERO9	6. AMOBEE 7. BEEWEB 8. DRUTT 9. JAMBA 10. JET MULTIMEDIA 11. LEAPSTONE 12. NEC 13. NEODATA 14. NOKIA SIEMENS 15. POLARIX 16. SCREENTONIC 17. THIRD SCREEN MEDIA 18. UNIPIER 19. XIAM TECH	20. ALCATEL LUCENT 21. BEA SYSTEMS 22. ERICSSON 23. IBM 24. IM 25. LOGICA CMG 26. MICROSOFT 27. REPLY	28. MBLOX 29. SYBASE 365	30. AEPPONA - APPIUM 31. BUONGIORNO 32. COMVERSE 33. DYLOGIC 34. FIRST HOP 35. HUAWEI 36. HP 37. QUALCOMM 38. REITEK 39. TELENITY 40. TXT POLYMEDIA

Within the theoretical sample of companies analyzed through case studies, only 2 platforms on 40 – 5% of the overall sample – under consideration are Transactional platforms: such platforms, whose core task is to support the Charging – Billing – Accounting process, need their providers to create a net of both contractual relationships and

infrastructural connections with MNOs' system. This element makes them hard to replicate, but also to develop: therefore, only a few players operate globally on the transactional services business area.

The second less-populated cluster is the Content Creation one, containing 5 solutions – representing the 12.5% of the overall sample –: these solutions are mainly implemented by players internalizing the content creation core business, like CPs.

In the companies sample, 8 – 20% of the overall sample – solutions could be labeled as business management platforms: these platforms are mainly offered by big companies coming from System Integration – e.g. Logica CMG –, IT Platforms provisioning – e.g. IBM, Microsoft – or Network Equipment provisioning – e.g. Alcatel Lucent, Ericsson – markets, who leveraged on their core resources and assets like technology know-how, products portfolio, competitive positioning and financial solidity, to penetrate the MCSDP segment.

The Transversal platform category is populated by 11 solutions – 27,5% of the overall sample –: these all-purpose, end-to-end platforms are provided by Mobile incumbent players – e.g. Buongiorno, Comverse, Dylogic – or other quite heterogeneous companies attempting to diversify their IT offer – e.g. HP, Qualcomm, Reitek, Txt Polymedia –.

The most populated cluster is the one grouping Content Management platforms: 14 solutions on 40 – 35% – are classified as focused on content management: again, these products, covering the core module related to managing rich media digital content, are commercialized by quite a diverse range of players, encompassing: pure MMTPs, like Beeweeb, Nokia-Siemens Networks, Nec, Polarix and Xiam Technologies; Mobile incumbent companies that pair their MCSDP offer to the commercialization of content and services published on the platform, like Jamba and Jet Multimedia; MASP focusing on Campaign Management and Mobile Advertising, like Amobee, ScreenTonic and Third Screen Media; and providers of “best of breed” products covering specific platform modules, e.g. Drutt for content lifecycle management, Neodata for business intelligence and reporting, and Unipier for DRM and IPR.

4.2.The Mmtps Choices Concerning the Technology Classification Variables

The last step of the model's application consists in mapping the MCSDPs offered by the analyzed players in terms of their adoption of the technological design dimensions previously identified through the House of Quality. Through such analysis, it will be possible to conclude the platform classification, thus supporting the technical benchmarking of the solutions currently available on the market, pinning down and interpreting their main positive and negative elements, and in the end drawing insightful conclusions on the offer state of the art.

The Table below shows, for each of the 40 analyzed players, the adoption choices made at the transversal technological design dimensions level. A heavy adoption – i.e. a strong implementation of the given technology element – is evidenced in green, while a low adoption is evidenced in yellow; the absence of the given technology lever in the solution under scrutiny is conveyed by a red cell in the table.

Analyzing each line of the table, one can grasp a single player's overall positioning, while a column analysis shows a technology dimension's pervasiveness throughout the companies sample.

Table IV. The application of Step 3 to the MMTPs sample.

	Delivery Channels	Content Types	Media Types & formats	Proprietary Technology	Open Source Technology	SOA & Web Services	IMS	OSA/Parlay API	Interactivity	Context-aware & LBS	Out of the Box	Taylor Made	ADP	Mark-up languages
ACOTEL	Y	G	G	R	R	R	R	R	R	G	R	R	Y	G
AEPONA	G	G	G	R	G	G	G	G	G	R	G	G	G	G
ALCATEL	G	G	G	R	R	G	G	G	G	G	G	Y	G	G
AMOBEE	Y	Y	Y	G	R	R	R	R	G	R	G	R	Y	G
BEA SYST.	Y	Y	R	R	G	G	R	R	R	R	G	R	Y	G
BEEWEEB	Y	Y	R	R	G	G	G	G	R	R	R	R	Y	G
B!	G	G	G	G	G	G	R	R	G	G	G	Y	G	G
COMVERSE	Y	G	G	G	R	R	G	G	G	G	G	G	G	G
DYLOGIC	Y	G	G	R	R	G	G	G	G	G	R	Y	G	G
DRUTT	Y	Y	R	R	R	G	R	R	R	G	G	Y	R	R
ERICSSON	G	G	G	R	R	G	G	G	G	G	R	Y	R	R
FIRST HOP	G	G	R	R	G	G	G	G	G	R	R	R	R	R
HUAWEI	G	G	G	R	R	R	G	R	R	R	R	Y	R	R
HP	G	G	G	R	R	R	R	R	R	R	R	Y	R	R
IBM	G	G	G	R	R	R	R	R	R	R	R	Y	R	R
IM	G	Y	G	R	R	R	R	R	R	R	R	Y	R	R
JAMBA	G	G	G	R	R	R	R	R	R	R	R	Y	R	R
JET M.	Y	G	G	R	R	R	R	R	R	R	R	Y	R	R
LEAPSTONE	G	G	G	R	R	G	G	G	R	R	R	R	R	R
LOGICA	G	G	G	R	R	R	R	R	R	R	R	Y	R	R
MBLOX	Y	Y	Y	R	R	R	R	R	R	R	R	Y	R	R
MICROSOFT	G	G	G	R	R	R	R	R	R	R	R	Y	R	R
NEC	G	G	Y	R	R	R	R	R	R	R	R	Y	R	R
NEODATA	Y	G	Y	R	R	R	R	R	R	R	R	Y	R	R
NOKIA	G	G	G	R	R	R	R	R	R	R	R	Y	R	R
OPENWAVE	Y	Y	Y	R	R	R	R	R	R	R	R	Y	R	R
POLARIX	Y	Y	G	R	R	R	R	R	R	R	R	Y	R	R
QUALCOMM	G	G	G	R	R	R	R	R	R	R	R	Y	R	R
REITEK	Y	G	G	R	R	R	R	R	R	R	R	Y	R	R

Table IV. Continued

REPLY	Heavy	Low	Heavy	Absent	Heavy	Heavy	Heavy	Absent	Heavy	Low	Heavy	Heavy
SCREENT.	Low	Low	Low	Heavy	Absent	Absent	Absent	Absent	Heavy	Low	Absent	Heavy
SPB	Low	Low	Low	Heavy	Absent	Absent	Absent	Absent	Heavy	Low	Absent	Low
SYBASE	Low	Low	Low	Heavy	Absent	Absent	Absent	Absent	Heavy	Low	Absent	Heavy
TELENITY	Heavy	Heavy	Heavy	Absent	Heavy	Heavy	Heavy	Absent	Heavy	Low	Absent	Heavy
3RD SCREEN	Low	Low	Low	Absent	Absent	Absent	Absent	Absent	Heavy	Low	Absent	Heavy
TXT	Heavy	Heavy	Heavy	Absent	Heavy	Heavy	Heavy	Absent	Absent	Heavy	Heavy	Heavy
UNIPIER	Heavy	Heavy	Low	Heavy	Heavy	Heavy	Heavy	Absent	Absent	Low	Absent	Heavy
USTREAM	Low	Low	Low	Absent	Heavy	Heavy	Heavy	Absent	Absent	Low	Absent	Heavy
XIAM TECH	Low	Heavy	Low	Absent	Absent	Absent	Absent	Absent	Heavy	Low	Absent	Low
ZERO 9	Low	Heavy	Low	Absent	Heavy	Heavy	Heavy	Absent	Heavy	Low	Absent	Heavy

Adoption

Heavy
Low
Absent

By observing the MCSDPs positioning, the prevalence of platforms capable of delivering a wide range of content clearly emerges. Moreover, it is possible to argue that the multichannel option is followed exclusively by platforms offering a large content portfolio: this finding can be explained by considering that the investments required for the integration of different delivery channels are only justifiable if high revenues coming from a wide VAS offer are expected. The interactivity feature is also quite diffused in the sample, being present on 27 platforms out of 24, while context-aware and LBS are not so common yet, being supported by only 15 platforms.

When addressing the correlation between open or proprietary technologies, multiple channels, SOA and IMS, other intriguing elements emerge. The SOA approach is adopted in 24 products out of 40, demonstrating the validity of this architectural paradigm coming from the IT enterprise platforms environment, and quickly diffusing in the Telecommunications context. IMS is employed by approximately half of the of products – 21 solutions –, testifying the service layer evolutions towards an “all IP” approach. Proprietary technologies are preferred to open source ones – 27 vs. 13 solutions –, as MMTPs struggle to make their offers unique, and potentially lock in their business customers.

The products distribution in the table shows that while SOA approach is more common along the open source axis – proving the positive correlation existing between the two elements, as evidenced in the “roof” of the House of Quality –, the IMS adoption is frequent in the multichannel alternatives – again, consistently with the considerations made when crossing the two ECs in the House –, regardless of the “technology” variable, in the light of the growing significance of IP in the process of integrating different delivery technologies

Considering the combined effect of the dimensions mostly impacting on the platform interoperability with legacy or third parties systems – i.e. technology employed; customizability level; OSA/Parlay API availability; mark-up languages support; ADP support, ranging from narrow to wide support –, a consistent fragmentation becomes evident. APIs are widely used, as well as mark-up languages; on the other hand, the support to

different ADPs is still narrow, because of some “standard wars” between proprietary technologies the regulators or the international consortiums will be asked to settle. Concerning customizability, in 20 cases the MCSDPs are standard, out-of-the-box products, with little or no personalization or parameterization features; in 11 cases, however, standardization and customizability coexist, and the platforms are characterized by some degrees of flexibility in terms of design and implementation.

The picture obtained through this technology classification allows to make insightful inferences on the MMTPs offer state of the art, in terms of both strengths and weaknesses.

The main pluses characterizing the offer can be synthesized as follows:

- wide portfolio of content and services deliverable;
- widespread support to interactivity;
- integration between mobile and web channels;
- wide support to media types and formats;
- frequent SOA and IMS adoption;
- significant modularity, flexibility and scalability;
- frequent OSA/Parlay API adoption;
- common use of mark-up languages.

On the other side of the coin, the current MCSDP offer is characterized by some significant drawbacks:

scarce support to context aware and location-based services;

- verticality and poor interoperability of some proprietary products;
- criticality of content adaptation processes;
- low products customizability;
- limited horizontal support to ADP.

5. Conclusion

The research provided an original reference model for supporting a technology classification of mobile Content & Service Delivery Platforms.

Such original framework was built on two pillars: a literature analysis well grounded on the existing body of knowledge concerning middleware platforms and IT systems, and an empirical analysis based on case studies on 40 Mobile Middleware Technology Providers, the platform vendors.

Quality Function Deployment methodology also played an important role within the reference model: the House of Quality tool was employed for translating the platform customers’ expected benefits in engineering characteristics. Integrating QFD methodology has a number of relevant positive effects on the model developed. First, it creates a strong link between the company and the “outer world” represented by the market, as it forces to clearly identify the product’s benefits – i.e. the customer requirements – and establish a direct, structured and quantifiable relationship to design and manufacturing characteristics, thus making the engineers meet the customers voice. Second, it potentially enhances the internal

relationship and communication among the company's departments, as CAs should be evaluated and analyzed by inter-functional teams so to individuate the right ECs to match them, without losing important information. Third, it strengthens the normative role of the model, supporting core activities of the decision making process, such as: product strategic positioning assessment; competitors benchmarking; strategic opportunities discovery; and target setting.

The model was hence applied to the sample of platforms currently marketed by MMTPs, so to test its validity, and obtain a valuable insight on the MCSDP offer state of the art. The findings show the real world offer of middleware platforms possesses some interesting features – ranging from the width of service offered and delivery channels supported, to the adoption of SOA and IMS approaches –; nevertheless, other significant drawbacks – e.g. insufficient support to context aware and location-based services, poor coverage to application development platforms etc. – are limiting the solutions effectiveness. Short term market trends will most likely see the coexistence of end-to-end transversal platforms, and of niche solutions focused on few modules or functionalities.

Concerning the model's properties, internal validity is ensured, for the platform positioning – dependent variable – is perfectly explained by the identified dimensions of classification – independent variables –; in terms of external validity, the model can be generalized to different populations, thanks to the width and significance of the sample under scrutiny; moreover, the rigorous qualitative research methodology employed should grant the reliability and replicability of the model's results. A limitation that needs to be addressed is that of tautological validity of the obtained results, since the model is applied to the very same sample of companies analyzed to gather the information employed to create it. However, the approach followed does not affect in any way the methodological soundness proper of the model development process, and it may exclusively impact on the relevance of the conclusions on the MCSDP offer state of the art drawn in section 4: nonetheless, the significant number of case studies performed, and the fact that the model's variables are also derived from a wide literature review which pairs the empirical analysis, contribute in attenuating this limitation. Future research will need to apply the model to a different sample of MMTPs, in order to test its validity outside the first sample which originated it, so to definitely solve the issue of tautological validity the model' conclusions may be burdened with.

The chapter's value for researchers can be brought back to the creation of a reference framework capable of rigorously modeling the emergent phenomenon related to the rise of middleware platform providers within the Mobile Content market. The chapter also contributes to extending the existing QFD literature, since it demonstrates the House of Quality tool's usefulness in a new context of application.

The value for practitioners lies in the provisioning of a tool with powerful descriptive and normative value.

On the descriptive side, the model can be used for mapping existing and future MCSDP offer, in terms of technological strengths and weaknesses. On the normative side, it supports the decision making process of a wide set of stakeholders. Customer firms attempting to find out what they should look for in middleware solutions and what they should implement according to their needs can employ it to set guidelines for platform adoption. Platform vendors themselves can look at the model to guide their offer positioning at a functionalities endowment level, and thanks to the creation of strong ties between platform capabilities and

associated benefits, to drive the choices made at a technological design dimensions level, which heavily affects the MCSDP market attractiveness

Though representing a significant step towards the study of MMTPs through the evaluation of the core element of their value proposition, the research does not specifically assess the strategic and competitive implications of choosing a given MCSDP technology positioning. The framework essentially constitutes an interesting technology-based support for internal strategic analysis: future research will need to focus on integrating the present model within a thorough external strategy analysis framework for mobile middleware platform providers.

References

- 3rd Generation Partnership Project (3GPP), 2006. *IP Multimedia Subsystem (IMS)*. TS23.228, Stage 2.
- ABI Research, (2006 [a]). *Mobile Content Delivery Platforms Enable Revenue Growth for Video, Games and music*. www.abiresearch.com.
- ABI Research, (2006 [b]). *Next Generation Service Delivery Platforms*. www.abiresearch.com.
- Adiano, C. & Roth, A. V. (1994). Beyond the house of quality: Dynamic QFD. *Benchmarking: An International Journal*, **1**, (1), 25-37.
- Ahmed, D. T. & Shirmohammadi, S. (2007). A framework for provisioning overlay network based multimedia distribution service. *IEEE ICME*, Beijing, China.
- Aioffi, W., Almeida, J., Mateus, G. & Mendes, D. (2004). Mobile Dynamic Content Distribution Networks. *MSWiM*, Venezia, Italy.
- Alcatel, (2006). *Managed Communications Services Architecture Framework*. Technology white paper.
- Anderson, C. (2004). *The Long Tail*. Wired.
- Ballon, P. & Van Bossuyt, M. (2006). Comparing business models for multimedia content distribution platforms. *MCDP project*, Institute for Broadband Technology.
- Balocco, R., Bonometti, G., Ghezzi, A. & Renga, F. (2008). Mobile Payment Applications: an Exploratory Analysis of the Italian Diffusion Process. *Proceedings of the 7th International Conference on Mobile Business (ICMB 2008)*, Barcelona, July 7-8, 2008.
- Barsook, J. & Freedman, E. (2005). Mobile Content Delivery Technologies. *IEEE*.
- Benali, O. et al. (2004). A Framework for an Evolutionary Path toward 4G by Means of Cooperation of Networks. *IEEE Communication Magazine*, **42**(5), 82-89.
- Bertelè, U., Rangone, A. & Renga, F. (2008). Mobile goes Web. Web goes Mobile. *Osservatorio Mobile Content report*. www.osservatori.net.
- Blind, K. (2005). Interoperability of software: demand and solutions. In Panetto, H. (Ed.). *Interoperability of Enterprise Software and Applications*. Hermes Science, London, 199-210.
- Bonomi, T. V. (1985). Case research in marketing: opportunities, problems, and a process. *Journal of Marketing Research*, **22**, 199-208.
- Bosserman, S. (1992). Quality Function Deployment: The Competitive Advantage. *Privated Trunked Systems Division*, Motorola white paper.

- Brown, P. G. (1991). QFD: Echoing the voice of the customer. *AT&T Technical Journal*, **70** (2), 18-32.
- Brown, P. G. & Harrington, P. V. (1994). Defining network capabilities using the voice of the customer. *IEEE Journal on Selected Areas in Communications*, **12**(2), 228-233.
- Brynjolfsson, E., Hu, Y. & Smith M. D. (2006). From Niches to Riches: Anatomy of the Long Tail. *Mit Sloan Management Review*.
- Capp, M. & Farley, P. (2005). Mobile Web Services. *BT Technology Journal*, **23**(2), 202-213.
- Chan, L. K. & Wu, M. L. (2002). Quality Function Deployment: a literature review. *European Journal of Operational Research*, **143**, 463-497
- Chang, C. H. & Lin, J. T. (1991). Data flow model of a total service quality management system. *Computers and Industrial Engineering*, **21**(1-4), 117-121.
- Chen, Y. et al. (2002). Personalized Multimedia Services Using a Mobile Service Platform. *IEEE*.
- Cohen, L. (1988). Quality function deployment: An application perspective from digital equipment corporation. *National Productivity Review*, **7**(3), 197-208.
- Eisenhardt, K. M., (1989). Building theories from case study research. *Academy of Management Review*, **14**(4), 532-550.
- Ecklund, D., Goebel, V., Plagemann, T., Ecklund, E. F. Jr., Griwodz, C., Aagedal, J. Ø., Lund, K. & Berre A. J. (2001). *QoS Management Middleware: A Separable, Reusable Solution*. Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 124-137.
- Ericsson, (2006). Service Delivery Platforms: efficient deployment of services. Ericsson *White paper*.
- Eriksson, I. & McFadden, F. (1993). Quality function deployment: A tool to improve software quality. *Information and Software Technology*, **35**(9), 491-498.
- European Telecommunications Standards Institute (ESTI) Standard (2005 [a]). Open Service Access; Application Programming Interface (API); Part 1: Overview (Parlay 5)", ES 203 915-1 v1.11.
- European Telecommunications Standards Institute (ESTI) Standard, (2005 [b]). Open Service Access (OSA); Parlay X Web Services; Part 1: Common", ES 202 391-1 v1.11.
- Yob, E. (1998). Quality function deployment in management information systems. *Journal of International Information Management*, **7**(2), 95-100.
- Hermann, F. & Heidmann, F. (2002). *User Requirement Analysis and Interface Conception for a Mobile, Location-Based Fair Guide*. Proceedings of Mobile HCI 2002, *Lecture Notes in Computer Science*, **2411**, 388-392.
- Forrester Research (2005). Real-world SOA: SOA Platform case studies. *Tech Choices*.
- Forrester Research (2007). Service Delivery Platform success requires a strategic vision and corporate collaboration. *Trends*.
- Fouial, O. et al. (2002). Adaptive Service Provision in Mobile Computing Environments. *IEEE*.
- Gaedke M. et al. (1998). *Web Content Delivery to Heterogeneous Platforms*. *LNCS*, **1552**, 205-217.
- Gartner Research (2002). Simplify Your Business Processes With an SOA Approach. *Research report*.
- Georganas, N. D. (1997). Multimedia Applications Development: Experiences. *Multimedia Tools and Applications*, **4**, 313-332.
- Griffin, A. & Hauser, J. R. (1993). The voice of the customer. *Marketing Science*, **12**, 1, 1-27.

- Groenveld, P. (1997). Roadmapping integrates business and technology. *Research Technology Management*, **40**(5), 48-55.
- GSM Association (2003). Location based services. *Permanent reference document SE.23*.
- Haag, S., Raja, M. K. & Schkade, L. L. (1995). Quality function deployment usage in software development. *Communications of the ACM*, **39**, 1, 42-49.
- Haavind, R., (1989). Hewlett-Packard unravels the mysteries of quality. *Electronic Business*, **15** (20), 101-105.
- Han, C. H., Kim, J. K., Choi, S. H. & Kim, S. H. (1998). Determination of information system development priority using quality function deployment. *Computers and Industrial Engineering*, **35**(1-2), 241-244.
- Hauser, R. & Clausing, D. (1988). The house of quality. *Harvard Business Review*, May-June 1988.
- Hewlett-Packard, (2005). HP Service Delivery Platform. *Hewlett-Packard white paper*.
- Houssos, N., Koutsopoulos, M. & Schaller, S. (2000). A VHE architecture for advanced value-added service provision in 3rd generation mobile communication networks. *IEEE*.
- IETF (1995). *Hiper Text Markup Language*.
- iSupply Corp (2007). *Mobile Content Enablement Platforms: Software Platforms Monetize and Deliver Mobile Music, Games and Video*. www.isuppli.com.
- Karlich S. et al. (2004). A self-adaptive service provisioning framework for 3G+/4G mobile applications. *IEEE Wireless Communications*.
- Karlich, S. (2007). An approach to mobile service delivery platforms. *IST OPIUM Project*, blue paper.
- Kim, K., Park, K. & Seo, S. (1997). A matrix approach for telecommunications technology selection. *Computers and Industrial Engineering* **33**(3-4), 833-836.
- Komulainen, H., Mainela, T., Sinisalo, J., Tahtinen, J. & Ulkuniemi, P. (2004). Business Models in the emerging context of Mobile Advertising. *ROTUAARI research project*.
- Kotsopoulos, M., Alonistioti, N., Gazis, E. & Kaloxyllos, A. (2001). Adaptive Charging Accounting and Billing system for the support of advanced business models for VAS provision in 3G systems. *IEEE MOBIVAS Project*.
- Kuo, Y. & Yu C, (2006). 3G Telecommunication operators' challenges and roles: a perspective of mobile commerce value chain. *Technovation*, 1347-1356.
- Laakko, T. & Hiltunen T. (2005). Adapting Web Content to Mobile User Agents. *IEEE Internet Computing*, 46-53.
- LaSala, K. (1994). Identifying profiling system requirements with quality function deployment. *Proceedings of the Fourth Annual International Symposium of the National Council on Systems Engineering*, August 10-12, San Jose, CA, 1, 249-254.
- Leavitt, N. (2003), Two technologies vie for recognition in speech market. *Computer*, 13-16.
- H. Li, M. Li, B. Prabhakaran (2006). Middleware for streaming 3D progressive meshes over lossy networks. *ACM Transactions on Multimedia Computing, Communications, and Applications*, **2**, 4, 282-317.
- Li, J., Zhang, J., Verma, S. & Ramaswamy, K. (2003). Mobile Content Delivery Through Heterogeneous Access Networks. *IEEE*.
- Lozinski, Z. (2003). Parlay/OSA – a New Way to create Wireless Services. *Mobile Wireless Data, International Engineering Consortium*.

- Ma, W., Bedner, I., Chang, G., Kuchinsky A. & Zhang, H. (2000). A framework for adaptive content delivery in heterogeneous network environments. *Hewlett-Packard laboratories*, White paper.
- Meredith, J. (1998). Building operations management theory through case and field research. *Journal of Operations Management*, **16**, 441-454.
- Metso, M. et al., (2002). Mobile Multimedia Services – Content Adaptation. *IEEE*.
- Mobile Entertainment Industry and Culture (2003). WP4 – Mobile Technologies Deliverable D4.1.1: *Existing and imminent Mobile Entertainment Technologies*. IST-2001-38846.
- Moerdijk, A. J. & Klostermann, L. (2003). Opening the Networks with Parlay/OSA: Standards and Aspects Behind the APIs. *IEEE Network*, **3**, 58-64.
- Muschamp, P. (2004). An introduction to Web Services. *BT Technology Journal*, **22**(1), 9-18.
- Nolle, T. (1993). ATM must clothe itself in cost justification, not naked hype. *Network World*, **10**(11), 27.
- Noordman, M. (2006). Squeezing the guy in the middle. *Ericsson Business Review*, **24**, 26-29
- Pahlavan, K. & Levesque, A. (1995). *Wireless Information Networks*. Wiley.
- Pailer, R., Stadler, J. & Miladinovic, I. (2003). Using PARLAY APIs Over a SIP System in a Distributed Service Platform for Carrier Grade Multimedia Services. *Wireless Networks*, **9**(4), 353-363.
- Pavlovsky C. J. & Staes-Polet, Q. (2005). Digital Media and Entertainment Service Delivery Platform. *IBM MSC*.
- Peppard, J. & Rylander, A. (2006). From Value Chain to Value Network: an Insight for Mobile Operators. *European Management Journal*, **24**(2).
- Pettigrew, A. (1988). *The management of strategic change*. Blackwell, Oxford.
- Philips, M., Sander, P. & Govers, C. (1994). Policy formulation by use of QFD techniques: A case study. *International Journal of Quality and Reliability Management*, **11**(5), 46-58.
- Porter, M. (2001). Strategy and the Internet. *Harvard Business Review*, 62-78.
- Rosas-Vega, R. & Vokurka, R. J. (2000). New product introduction delays in the computer industry. *Industrial Management and Data Systems*, **100**(4), 157-163.
- Sabat H. K. (2002). The evolving mobile wireless value chain and market structure. *Telecommunications Policy*, **26**, 505-535
- Sarkis, J. & Liles, D. H. (1995). Using IDEF and QFD to develop an organizational decision support methodology for the strategic justification of computer-integrated technologies. *International Journal of Project Management* **13**(3), 177-185.
- Sharkey, A. I. (1991). Generalized approach to adapting QFD for software. *Transactions of the Third Symposium on Quality Function Deployment*, June 24-25, Novi, MI, 380-416.
- Sullivan, L. P. (1986). Quality function deployment. *Quality Progress*, **19**, 6, 39-50
- Sur, A., Skidmore, D. & Chakravarty, S. (2006). Web Services based SOA for Next Generation Telecom Networks. *IEEE International Conference on Service Computing*.
- Tan, K. C., Xie, M. & Chia, E. (1998). Quality function deployment and its use in designing information technology systems. *International Journal of Quality and Reliability Management*, **15**(6), 634-645.
- The Insight Research Corporation, (2007). IMS, SIP and Service Delivery Platforms: Telecom adoption of SOA and Enterprise applications 2007-2001. *Report*.
- Wasserman, G. S., Gavoort, M. & Adams, R. (1989). Integrated system quality through quality function deployment. *Proceedings of the 1989 IIE Integrated Systems Conference*, Atlanta, GA, pp. 229-234.

- Williams, R. A. (1994). Delivering the promise. *World Class Design to Manufacture* **1**(1), 33-38.
- Yin, R. (2003). *Case study research: Design and methods*. Thousand Oaks, CA: Sage Publishing.
- Zahariadis T. et al. (2002). Global Roaming in Next-Generation Networks. *IEEE Communications Magazine*, **40**(2), 145-51.
- Zhang, D. (2007). Web Content Adaptation for Mobile Handheld Devices. *Communications of the ACM*, **50**(2), 75-79.
- Zhijun Lei and Georganas, N. D. (2001). Context-based Media Adaptation in Pervasive Computing. *Proceedings of Can. Conf. on Electr. and Comp. Eng.*

Chapter 2

USING DIGITAL WATERMARKING TO IDENTIFY AUDIO AND VIDEO SOFTWARE PRODUCTS

***Takaaki Yamada, Yoshiyasu Takahashi, Ryu Ebisawa,
Yoshinori Sato and Seiichi Susaki***

Systems Development Laboratory, Hitachi, Ltd., Yokohama, Japan

Abstract

When illegally distributed contents digitally watermarked with the serial numbers of the software products that generated those contents are found, the watermarks help software vendors to determine whether or not their software was illegally re-distributed by licensed users. It is difficult, however, to detect watermarks in content that has been seriously damaged by signal processing. Detection can be improved by using the original content, but the software vendor usually cannot use the original content due to copyright concerns. We present two practical software product serialization applications that use a reference signal similar but not identical to the original content: a text-to-speech interface working with audio watermarking and a video encoder working with video watermarking. We evaluated them using a previously developed audio and video watermarking prototype based on the patchwork algorithm and by modifying the time and spatial domains of the data. Because the proposed detection method assumes the use of the same embedding method used in the conventional detection methods, it not only detects watermarks without having to use a reference signal but also performs better by using the reference signal.

Keywords: software product serialization, software security, video watermark, audio watermark, reference signal

1. Introduction

The threat of piracy starts when a digital asset such as a commercial software product is released for sale. Software pirates may find a way to duplicate the software and spread it without the permission of the software product vendor. Copy protection schemes for early systems were designed to defeat the casual copiers, several types of copy protection methods

were developed [1][2]. Some password-based encryption, some authenticate special data patterns hidden in distributed media, some used dedicated devices such as dongles, and some authenticate products or users through the network. Password-based copy protection could prevent uses of illegal copies if the subsequent users are not told the password. However, if the licensed first users were not reliable, illegal re-distribution of the encrypted data along with the password could not be prevented. Protection depended on the morals of the users. Once such information was provided freely, the increasing number of secondary users damaged legal sales of the software. Therefore, password-based copy protection does not provide sufficient protection. If more stringent security methods used to protect the software, legal users would face usability problems that might reduce software sales. For instance, authentication through the internet can prevent casual software copying. However, it is inconvenient for not only end users but also software vendors to maintain the authentication server systems. Moreover, such strict security can be difficult to apply depending on the software application. Once a software product is cracked and illegally copied, it is difficult to prevent illegal behaviors by users such as distributing illegal copies. Therefore, it is important to identify illegally distributed copies in order to limit the extent of the damage and/or obtain compensation for the damage.

Here, we consider software products that process digital content such as audio and video. One way to protect the copyright of a software product is to use a digital watermarking technique (typically used for digital content). One way to identify illegal copies of software is to use software watermarking [3]. That is, a user ID for the software product is embedded in the software itself by the code. Detection of this ID in the illegal copies of the software can lead to identification of the source of the copying. Similarly, detection of embedded watermarks can help auditors identify the source of copying and/or prove theft. The problem with most software watermarking methods is that software such as video encoders is primarily used in-house. Therefore, auditors have difficulty determining whether illegal copies of software are used or not. However, the digital content processed by the software product may be disclosed in an open space where auditors can survey it. Therefore, digital content processed by a software product may give auditors a chance to detect illegal copies of the software, if the user ID is contained in the content. However, such additional information as user ID in metadata or file headers is apt to be erased by format conversion or attacks. A digital watermarking technique [4] can be used to tightly bond identifier information to the digital content processed by the software product. When watermarks indicating a specific software user are detected in content not consistent with the user's regular operations, auditors (and/or the software vendors) should be suspicious of illegal copying and use. Auditors thus can start investigating the original software by hearing to the identified user, and so on.

Embedded watermarks in digital content are almost imperceptible to people. Embedded watermarks can survive the various signal processing that are performed slightly on digital content. However, it is difficult to detect watermarks in content that has been seriously damaged by signal processing. For instance, some service systems for sharing digital content require that the content be encoded at a low bit rate. Watermark robustness against signal processing is thus as essential requirement. There are many ways to improve watermark robustness such as by repeating the watermark pattern, strongly embedding a watermark in a specific part of the content, and using a specific watermark pattern [4-7]. Previous work includes, for example, improving error correction coding [16] and using a probability model for error detection [20]. A conventional approach to improving watermark detection is using

the original content [17]. However, the original content is usually protected by copyright, making it difficult to use the original content for detecting watermarks without the cooperation of the copyright holders.

We present two practical watermarking applications for serializing software products using a reference signal similar but not identical to the original content: a text-to-speech interface working with audio watermarking and a video encoder working with video watermarking. We describe methods for software product serialization using digital watermarking and their problems in section 2. Our text-to-speech application working with audio watermarking is described in section 3, and our video encoder application working with video watermarking is described in section 4. We conclude in section 5 with a summary of the key points and a look at future work.

2. Software Product Serialization

2.1. Effectiveness of Identifying Software Product

The primary purpose of detecting illegal behaviors is to help software vendors limit the extent of damage or obtain compensation for the damage. The objective is to both detect illegal user behavior (making illegal copies and so on) and obtain information identifying the persons responsible.

Consider a case in which auditors or software vendors detect only the illegal behavior and not the person responsible. Illegal behaviors to be detected are making illegal copies of software licensed to a legal user, redistributing the illegal copies, and using the illegal copies. If the auditors are able to identify the area where the behavior occurred, they might be able to reduce the extent of the damages, by implementing countermeasures such as changing sales strategies. For instance, if the price of the software is reduced down in an area where the sales of illegal copies have been detected, the illegal sales might be reduced. However, it is difficult for auditors to detect the behaviors of illegal users within a company. They may be able to detect sales of illegal copies of software by means of surveillance on the internet and in shops. Even so, it is hard to detect illegal sales through underground networks such as those using P2P communication. The detection usually occurs only after the damages have become substantial. When illegal copies of software are found on an open web site by chance, auditors can ask the site manager to delete the illegal copies from the server. However, because the responsibility for the damages is limited in accordance with the law for internet providers, auditors generally have difficulty making a claim for damage compensation against the site manager. Moreover, the site manager may not reveal who subscribed to the illegal copies because the anonymity of subscribers to such open web sites is usually respected. Although deletion of illegal copies from a specific site would help reduce the extent of damages, such illegal behavior detected by chance is only the visible part of the iceberg. The distribution of copied software is rapid, and the distribution area is apt to be worldwide. Therefore, if auditors detect only the illegal behavior and not the persons responsible, the effectiveness of such detection is limited. That is, it is important to also identify the people responsible.

Consider another case in which auditors or software vendors not only detect the behaviors of illegal users but also identify them. There are three kinds of persons to be

identified: the ones who make illegal copies, the ones who redistribute the illegal copies, and the ones who use the illegal copies. Even if auditors identify the illegal users, it may be difficult to press a claim against each one for damages because, for instance, some of them may be bona fide third persons. Moreover, it is difficult to recover all damages when there are many illegal users. Even if auditors identify illegal redistributors, the responsibilities of site managers are so restricted that the auditors may have difficulty recovering damages.

We focus on identifying the persons responsible for causing the damage by, for example, making illegal copies. Assume that identifying the software product is equivalent to identifying the legal user of the original software. For instance, a software package is typically activated by entering the product serial number when the software is installed. If the product serial number is redistributed illegally, it can be used to identify the user if the relationships between product serial numbers and purchasers have been adequately recorded. Such information would help auditors begin surveying the distribution route from the identified purchaser. Thus, identifying the purchaser would help auditors limit the extent of damages and recover damages.

2.2. Software Product Serialization Using Digital Watermarking Techniques

Software product serialization can be accomplished by making each instance of software identifiable. The assumed application of software product serialization using watermarking technique is achieved by (a) making software installed in a user's PC after the product serial number is input, (b) making the software customized as the product serial number must be embedded in content that processed by the software, and (c) permitting users to use the software. Digital watermarking tightly bonds such identifier information to digital content processed by the software product. The software would embed identifier information in the original content and output watermarked content in place of the original content.

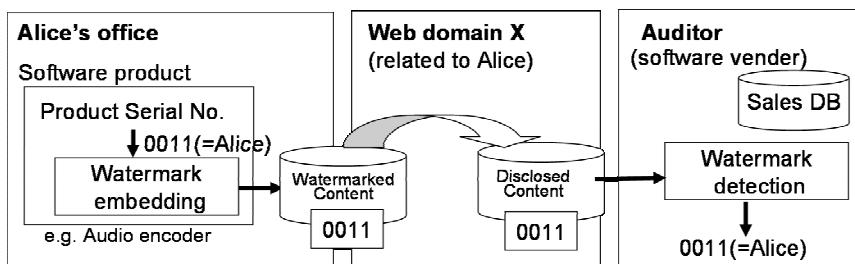


Figure 1. Software product serialization using digital watermarking (regular process).

Consider a regular case in which a software product, an audio encoder, is used legally. The software vendor (or auditor) identifies each instance of the software product.

- (1) The vendor receives an order for the product, say from Alice.
- (2) The vendor stores specific information (typically the package serial number, for instance, 0011) along with information about Alice in its sales database.

- (3) It releases the software package to Alice and notifies her of the product serial number.
- (4) Alice installs and activates the software by inputting the product serial number during the set-up process. Then, Alice uses the software to produce digital content files in which the product serial number is embedded by digital watermarking, as shown in figure 1. The watermarked files are disclosed by Alice where is known as related to Alice.
- (5) The auditor looks for suspicious files by surveying web sites.
- (6) If a suspicious file is found, the auditor identifies the software that processed it by detecting the watermark (containing the product serial number) in the file, as shown in figure 1.
- (7) The auditor identifies the licensed user, Alice, by checking the sales database.
- (8) The auditor estimates whether the identified software product would have been regularly used by Alice by considering the relationships among Alice, the web domain where the file was found, and the file content.

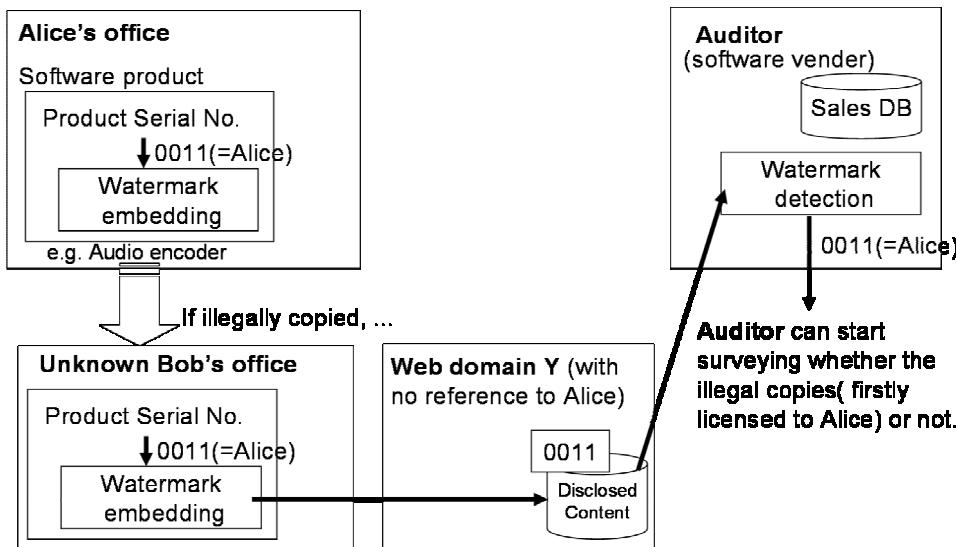


Figure 2. Software product serialization using digital watermarking (irregular process).

Consider an irregular case in which a software product (licensed to Alice) is illegally copied and used in Bob's office, as shown in figure 2. Bob is an illegal secondary user unknown to the software vendor or auditor. The first three steps by the vendor are the same as steps (1), (2), and (3) described above. In step (4), Bob uses the illegally copied software product, which embeds a watermark (0011, indicating Alice) in the digital content files processed by the software. Then, Bob discloses those files at a web domain Y, which can be estimated out of relation to Alice. The next three steps by the auditor are the same as step (5), (6), and (7) described above. When the auditor estimates the relationship between Alice and web domain Y in step (8), the auditor might notice something unusual. For instance, if the language used in web domain Y differs from that used in web domain X, it is likely that the software has been illegally copied and used somewhere in the other country. Only Alice should be using the software. The auditors can then start investigating whether the software

package has been illegally copied by talking to Alice or the manager of web domain Y. The auditor should be careful because the identified legal user, Alice, may not be the person who made the illegal copies. If the auditor can identify the person responsible for the copying, the auditor could make a claim for damages. Alice should be held accountable for the copying in accordance with the license agreement. The auditor thus might find Bob. Responsibility for damages due to the illegal copying could lie with Alice, with Bob or with both of them. The auditor should make a claim accordingly.

2.3. Underlying Watermarking Algorithm

Many digital watermarking algorithms for audio and video have been developed [5-13] that work by modifying signal values such as the audio amplitude values or pixel values of video frames. The patchwork algorithm [5] is one such algorithm, and we use it in our proposed systems for audio and video, which are described below. First, consider a one-bit scheme for embedding one bit in the signal values.

Embedding process: Signal value set in the original signal is $\mathbf{s} = \{s_i \mid 1 \leq i \leq N\}$ where N is the number of the signal values into which a watermark is to be embedded and i is an index identifying a specific signal value in the signal value set. Two sets of index numbers, $\mathbf{A} = \{\alpha_k \mid 1 \leq k \leq n\}$ and $\mathbf{B} = \{\beta_k \mid 1 \leq k \leq n\}$ (where $\mathbf{A} \cap \mathbf{B} = \emptyset$ and $2n \leq N$), are selected at random. Watermark pattern, $\mathbf{m} = \{m_i \mid 1 \leq i \leq N\}$, is given by

$$m_i = \begin{cases} \pm \delta & \text{if } i \in \mathbf{A} \\ \mp \delta & \text{if } i \in \mathbf{B} \\ 0 & \text{otherwise} \end{cases},$$

where $\delta (> 0)$ is watermark strength. The bit value of the embedded information is $b \in \{1, 0\}$. The " \pm " and " \mp " respectively indicate "+" and "-" if $b = 1$ and "-" and "+" if $b = 0$.

The signal value set of the watermarked signal, $\mathbf{s}' = \{s'_i \mid 1 \leq i \leq N\}$, is given by following formula.

$$\mathbf{s}' = \mathbf{s} + \mathbf{m} \quad (1)$$

Detection process: An estimation value v is used for deciding whether result extracted from watermarked signal \mathbf{s}' is a bit value or not. It is given by

$$v = \frac{1}{n} \sum_k (s'_{\alpha_k} - s'_{\beta_k}) \quad (2)$$

Bit value b is obtained by comparing v and positive threshold value T :

$$b = \begin{cases} 1 & \text{if } v \geq T \\ 0 & \text{if } v < -T \end{cases}. \quad (3)$$

In the multiple-bit scheme, signal data is divided into regions, and the one-bit process is applied to each region. The multiple-bit process can be applied repeatedly to signal data. The patchwork technique has some degree of robustness against signal processing such as encoding. Note that the detection method itself uses only watermarked signal s' , and does not necessarily use the original signal s .

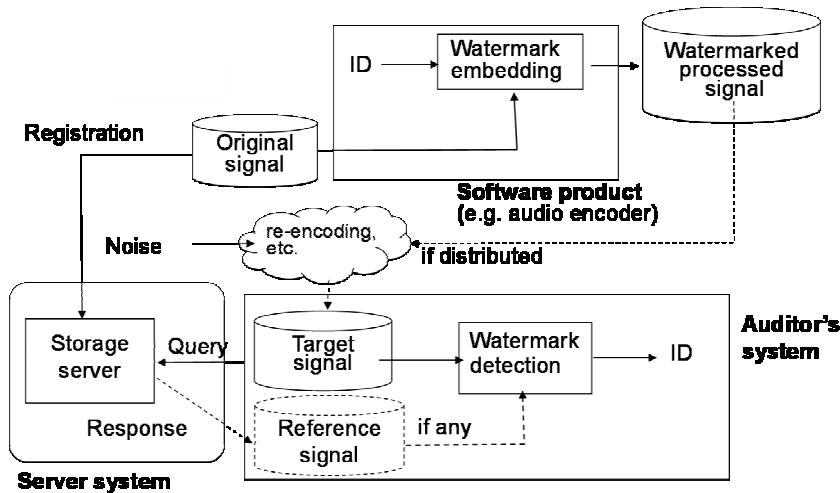


Figure 3. Conventional method for improving watermark detection performance using original signal. Watermarks can be detected with or without using a reference signal obtained by querying the server storing the original signal.

2.4. Problems with Conventional Watermark Detection Improvement Methods

It is difficult to detect watermarks in content that has been seriously damaged by signal processing. Some content sharing services require content to be encoded at quite a low bit rate, meaning that content can be seriously damaged when it is uploaded to the service site. One conventional method for improving detection performance is using the original signal. The target signal, the one from which watermarks should be detected, includes a weak watermark signal and a content-specific signal, which is similar to the original signal. When an auditor detects watermarks from the target signal, the content-specific signal is too noisy for watermark detection because the watermark signal is weak. The difference between the target signal and the original signal would approximately compensate for content-specific signal and help improve watermark detection by strengthening the watermark signal. Therefore, using the original signal could improve watermark detection performance. Some detecting schemes use the original signal [10][11] and some do not [5][8][9].

A conventional watermark application system is shown in figure 3. Software products have a watermark embedding function that embeds identifier information such as the user ID into the original signal and outputs a file containing a watermarked signal. If the watermarked files are distributed and an auditor comes across a suspicious data file (a “target signal”), the auditor tries to detect the watermarks in the target signal. The watermarks should be

detectable even if the original signal is not used. If a distributed signal is substantially changed by signal processing, the auditor's system may fail to detect the watermarks embedded within it. However, if the original signal was previously registered in a storage server from which the auditor could obtain it as a reference signal, watermark detection performance could be improved,

From the viewpoint of the copyright of digital content processed by a software product, there are three stakeholders: (a) the copyright holder for the original content, (b) the software user who processes the original content using the software product, and (c) the software vendor who provides the software. The software vendor is also an auditor who detects watermarks in the target content. Although software users own the original content, they cannot typically rent the original content to the software vendor due to the contract between the copyright holders and the software users. Since the original content is valuable, a contract is needed between the auditor and copyright holder. However, unless the copyright of the digital content will not be infringed, there is no reason for the copyright holder to cooperate with the auditor to detect watermarks. That is, auditors have difficulty obtaining permission to use the original content or original signal for detecting watermarks.

Therefore, a method is needed for improving detection performance without using the original signal. Here we present two approaches to doing this. They are practical application systems for software product serialization using a reference signal similar but not identical to the original content: a text-to-speech interface working with audio watermarking and a video encoder working with video watermarking.

3. Audio Watermark Detection Using Reference Signal

3.1. Research Approach

Many audio watermarking methods have been developed for protecting the copyright of digital audio content by embedding copyright information into it [6-9]. However, there has been little discussion of protecting the copyright of audio software products. Although digital watermarking techniques can be applied to digital content in both cases, the technical requirements differ. One reason for this difference is that auditors working to protect the copyright of software products have difficulty in using the original audio content for watermark detection, as described in section 2.4.

Audio watermarks should be inaudible to end users, invisible to unauthorized parties trying to detect or remove them, and robust enough to withstand the standard signal processing operations performed by audio encoders [6]. One watermarking method that potentially satisfies these requirements is the patchwork algorithm. We apply the original patchwork algorithm [5] to spatial-domain (or, equivalently, time-domain in audio) data. It embeds watermarks by creating a statistically meaningful difference between the average pixel values. Several techniques for applying the patchwork algorithm to audio data have been developed [8][9], and most work within the frequency domain of audio data. Most audio watermarking techniques use relatively short time frames for embedding, typically less than one second [7]. Although the size of the ideal detection frame can be the same as that of the frame used for embedding, many previous evaluations have shown that stable watermark detection takes several seconds or more, e.g., 30 seconds.

We found that re-synthesized voice can help watermark detection by serving as a reference signal in place of the original signal because synthesized voice can be easily re-synthesized. Synthesized voice data contain various imperceptible areas that could be slightly modified for watermark embedding. Synthesized voice has unvoiced parts of the time domain [12] and frequency domain [13] into which watermarks can be embedded. Such watermarks, however, would be potentially visible by analysis. Moreover, that might be so perceptible that attackers could erase them by simply using noise filters. Nevertheless, the embedding of watermarks in the voiced parts of synthesized voice has not been sufficiently investigated. The use of a reference signal should shorten the watermark detection time frame even if watermarks are embedded in the time-domain of the voiced parts.

3.2. Proposed System for Audio Watermark Detection Using Re-synthesized Voice

As shown in figure 4, the original voice data is made by a text-to-speech software user. The text-to-speech function uses synthesizing parameters set by the user before the original voice is synthesized. If the synthesizing parameters and the original text data are provided to auditors, they could re-synthesize the voice, and it would be the same as the original voice. However, auditors typically have difficulty not only obtaining the original text data but also obtaining the synthesizing parameters. Even if the original text data is not obtained, the auditor could re-synthesize a reference signal from the text data obtained by simply listening to the target signal. If the intonation for specific parts in the original signal was adjusted or tuned slightly by the user of the synthesizing software, the reference signal would differ somewhat from the original signal. Nevertheless, the unchanged parts in the original signal should be the same as the corresponding parts in the re-synthesized signal. Therefore, the re-synthesized voice should work well as a reference signal in place of the original signal.

We implemented a prototype audio watermark application system based on the patchwork algorithm [5]. As shown in figure 5, the embedded watermark patterns are superimposed on the amplitude of the original audio signal. The original audio data is partially masked to modify only the voiced part of the audio data. (That is opposite to previous work [12] in which the unvoiced part is used). The embedded information can be detected without the reference signal, but the detection performance should be improved if a reference signal is used [10][11]. The target signal might be changed from the watermarked signal due to signal processing. For instance, a pitch or time shift in the target signal is often caused by audio encoders. The target and reference signals should thus be synchronized in the watermark detection time frame.

This approach is a systematic combination of previous algorithms [5, 10-12], but its evaluation from the viewpoint of software serialization applications has not been discussed yet [15]. A method that modifies the time-domain of audio data has been thought to be disadvantageous for watermark robustness [6], but the proposed method using a reference signal could improve watermark detection performance by strengthening the embedded watermark patterns. This approach thus enables a watermark to be detected in a relatively short detection time frame (e.g., 2 s), which is shorter than the time frame needed when using another method without a reference signal. The shorter the detection time frame, the greater the number of detection chances.

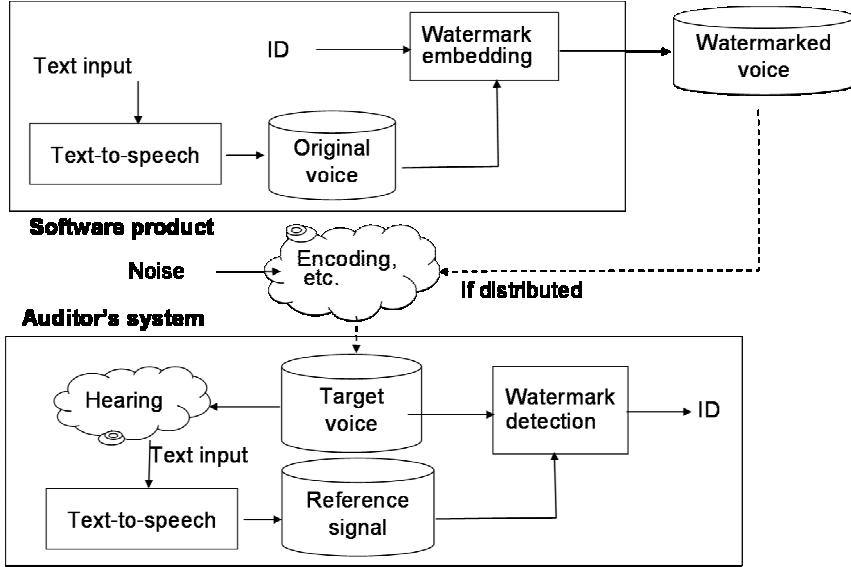


Figure 4. Proposed watermark application system for Text-to-speech software.

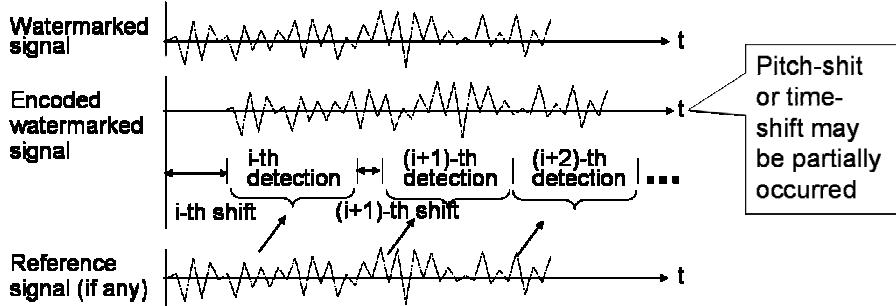


Figure 5. Waveforms in detection processes.

3.3. Process Flow of Audio Watermark Embedding and Detection

Consider a one-bit-scheme based on the patchwork algorithm for embedding one bit in amplitude values of audio data, as described in section 2.3. The set of amplitude values in the original audio is $\mathbf{x} = \{x_i | 1 \leq i \leq N\}$ where N is the number of amplitude values into which watermark is to be embedded and i is an index identifying a specific amplitude value in the amplitude value set. Two sets of index numbers, $\mathbf{A} = \{\alpha_k | 1 \leq k \leq n\}$ and $\mathbf{B} = \{\beta_k | 1 \leq k \leq n\}$ (where $\mathbf{A} \cap \mathbf{B} = \emptyset$ and $2n \leq N$), are selected at random.

The embedding steps are as follows:

- [E1]. Select n pairs (\mathbf{A} , \mathbf{B}) in voiced part of the original audio data.
- [E2]. Add a small positive value, δ , to the amplitude value, x_{α_k} . (δ is watermark strength.)

[E3]. Subtract the same value from the amplitude value, x_{β_k} .

[E4]. Repeat the step E2 and step E3 for all k .

Each bit of information is thus embedded by slightly modifying the original waveform values, and multiple bits can be embedded by repeating the process. Furthermore, the same information is repeatedly embedded as long as the input data remain.

The detection steps are as follows:

[D1]. Find the watermark pattern (the n pairs) in the pulse code modulation (PCM) data read from a specific address of the target signal, starting with the same location as selected in the embedding process.

[D2]. Calculate estimation value v by subtracting the mean amplitude of x_{α_k} from that of x_{β_k} .

[D3]. Decide whether or not one bit is embedded by comparing the difference with the previously defined threshold value, calculated using equation (3).

[D4]. If the detection is not complete, change the offset value for synchronization and repeat steps D1 through D3.

A multiple-bit-scheme is accomplished by repeating the above steps as described in section 2.3. When a reference signal is used, the difference signal between the target signal and reference signal should be likely watermark pattern. Detection step [D2] is thus modified to use the difference signal.

[D2]. Calculate difference signal $\text{difx} = \{ \text{difx}_i \mid 1 \leq i \leq N \}$ by subtracting the amplitude values of the reference signal from those of the target signal. Then, calculate estimation value v by subtracting the mean value of difx_{α_k} from that of difx_{β_k} .

3.4. Evaluation for Synthesized Voice

3.4.1. Experimental Conditions

Synthesized voice data files were generated using a commercial product, "Voice Sommelier" [14]. "Voice Sommelier" is a registered trademark of Hitachi Business Solution Co., Ltd., Japan. Thirty-second files of a voice reciting the preamble to the Japanese Constitution were used. Classical music files (the Gymnopédies) were also used. The sampling frequency was 22 kHz, and the sampling bit size was 8 bits. The original WAVE files in linear PCM format were watermarked. The original and watermarked files were encoded into MPEG audio layer 3 (MP3) files at a bit rate of 96 kbps and then subjectively evaluated using an audibility test with headphones. The data flow for the test is illustrated in figure 6.

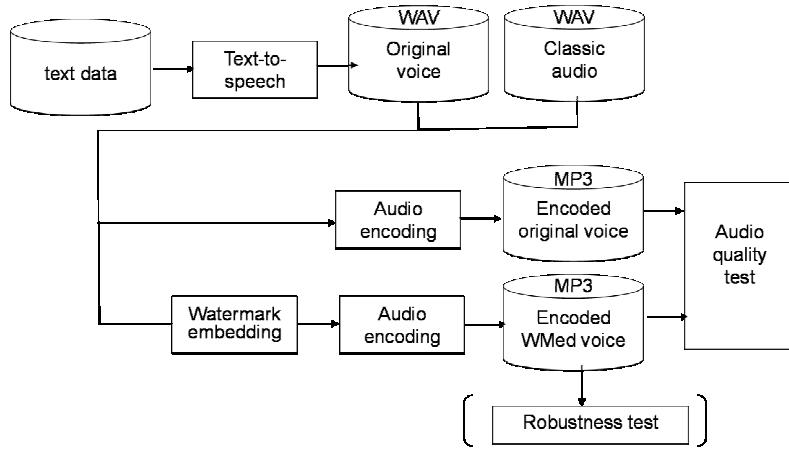


Figure 6. Data flow for audibility test.

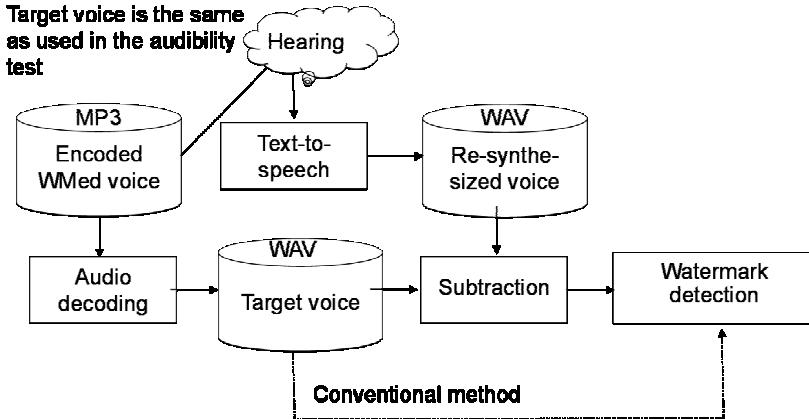


Figure 7. Data flow for robustness test.

The watermark payload was 64 bits, and the embedded information was scrambled, expanded, and repeated in time lengths of a specific length. The watermarked files were also encoded at several bit rates from 32 to 96 kbps and used in a robustness test. Re-synthesized voice can be generated through a text-to-speech function by inputting text data that by listening. Such re-synthesized voice is always the same as the original voice unless the synthesizing parameters are changed. Although illegal users may change the synthesizing parameters in a partial time line by, for example, synthesizing parameters for intonation at the end of a word, the rest of the synthesized voice should be the same as the default output, that is, the original voice. The watermarked MP3 files are decoded as target files. Then, each amplitude value in the target files is subtracted from the corresponding value in the reference signal file. Watermarks can be detected not only in the difference data by using the proposed method but also directly in the decoded target file by using a conventional method. The data flow for the test is illustrated in figure 7.

3.4.2. Results of Audibility Test

The quality of the watermarked audio was evaluated in a subjective manner by five persons listening to the watermarked MP3 files at 96 kbps and the original MP3 files through headphones. They rated watermark disturbance on a scale of 1 to 5, as shown in table 1. Watermarks in synthesized voice were found to be more transparent than watermarks in classical music, as shown in figure 8. Watermark strength 1 and 2 are adequate for assumed applications. Stronger watermarks would have better performance for watermark robustness.

Table 1. Level of disturbance and rating scale

Disturbance	Score
Imperceptible	5
Perceptible but not annoying	4
Slightly annoying	3
Annoying	2
Very annoying	1

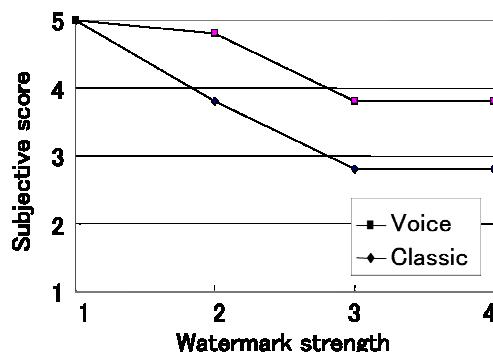


Figure 8. Audio quality of watermarked samples.

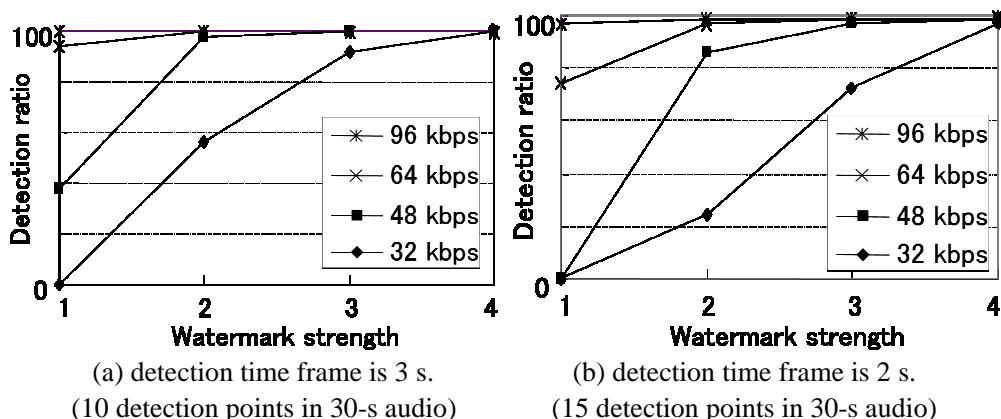


Figure 9. Watermark robustness for synthesized voice signals.

3.4.3. Results of Robustness Test

We used the watermark detection ratio, which is the ratio of the number of points for which the embedded 64 bits were correctly detected to the total number of detection points. As shown in figure 9, the detection ratio for synthesized voice can be improved by embedding stronger watermarks (i.e., by modifying the amplitude of the audio data more). Furthermore, the longer the detection time frame, the better the detection ratio. When the detection time frame was 2 s and the bit rate was 32 kbps, as shown in figure 9(a), the detection ratio was 22%, and there were 15 detection chances (determined by dividing 30 s by 2 s). In contrast, when the detection time frame was 3 s, as shown in figure 9(b), the detection ratio was 56% and there were 10 detection chances. The detection chances increased by a factor of 1.5 when the detection time frame was shortened from 3 to 2 s.

Pitch-shift, time-shift, and some filtering effect caused by the audio encoder were mostly unpredictable. They often affected the embedded watermark patterns embedded in the amplitude values of the audio data, resulting in detection errors. Even if watermarks were not detected for a long detection time frame, a less-affected small part of the audio data could offer give other chances to detect the watermarks for a short detection time frame. This approach should be effective for synthesized voice.

3.5. Summary of Evaluation for Audio Watermark Application

Copyright protection for text-to-speech software using digital watermarking techniques was discussed in this section. We developed a prototype audio watermarking system by combining the results of previous studies using the patchwork algorithm, modifying the amplitude in the time-domain of voiced audio data, and using a reference signal. Evaluation results showed successful detection of 64-bit watermarks in a 2-s time frame.

The reference signal was re-synthesized voice, the same as the default output of text-to-speech. Re-synthesized voice is different from the original voice because users may change their default voice by partially adjusting intonation and so on. However, since synthesized voice is made of many orally voiced word elements, some synthesized words are used the same as the original element data. Some parts of the re-synthesized voice should thus be the same as those of the original one. Therefore, using re-synthesized voice as a reference signal is an effective way to improve watermark detection performance in place of using the original synthesized voice.

4. Video Watermark Detection Using Reference Signal

4.1. Use Case of Illegal Copies of Video Encoder

Here we consider a use case of illegal copies of video encoder software product. As illustrated in figure 10, a software product (a video encoder) is sold by the software product vendor to an unreliable customer. If the encoder is illegally copied and used by a secondary user to distribute video content in a closed network, an auditor, or the software vendor, will have difficulty detecting the illegal copies because the copy of the encoder is used in-house.

Moreover, video content distribution services often protect the videos they create by using access control techniques. However, users of the video content distribution service using an illegal copy of the encoder can access the video content. If they download the videos and then upload them to an open video sharing service, the auditor can freely detect them as target images.

Since the video encoder works with watermarking, identifier information such as license ID or product serial number should be embedded in each video processed by the video encoder and its illegal copies. Watermarks in the content can survive the various image processes applied to the content. The target images should thus have watermarks identifying the encoder software that encoded the target images. If watermarks are in fact detected, the auditor should be able to identify not only the encoder but also the licensed user of the encoder. The auditor can then estimate whether the identified encoder would have been regularly used by the identified user by considering the relationship among the identified user, the web domain where the target images were found, and the disclosed content itself.

However, it is difficult to detect watermarks in content severely degraded by the image processing procedures used by video sharing services. Watermarks often have to be detected from videos that have experienced severe image processing such as a combination of re-encoding and resizing. From the viewpoint of the effect of image processing on watermarks, the use case shown in figure 10 is considered. As an actual threat in this use case, we use the video content redistribution model summarized in figure 11.

- (1) Before video content is distributed, the product serial number is embedded in the uncompressed original images. The watermarked original images are encoded into a compressed video format. The compressed video, comprising what we call “distributed images”, is distributed to users. DVDs are often used for such distribution.
- (2) Before the distributed images are secondarily redistributed, they are decoded, scaled down, and re-encoded in another compressed video format. The recompressed video, comprising what we call “target images”, is uploaded to a video sharing site. QVGA-size MPEG4-based format is often used by such sites.
- (3) If the target images are found by chance, the auditor tries to detect watermarks. If the same embedding information used in the watermark embedding process is detected, the licensed user can be identified.

4.2. Problems with Conventional Methods

As mentioned above, a conventional way to improve detection performance is to use the original images. Since the pixel values of the original image constitute a content-specific signal much stronger than the almost imperceptible signal of a watermark, the content-specific signal creates much noise in the watermark detection process. Since embedding watermarks adds a signal pattern to the original signal, the watermark pattern can be approximately calculated by subtracting the original image from the target image. That is, the signal of the watermark pattern can be identified by balancing the content-specific signal of the target image with that of the original image. However, the content-specific signal has been changed by the image processing applied to the watermarked image. The original image

should thus be adequately subtracted from the target image in accordance with the process of the image processing. Otherwise, much of the content-specific signal would remain in the difference image and disturb watermark detection.

Let's assume that the watermarking technique described in section 2.3 is applied to the luminance value of images. Watermark detection performance could be improved by using difference image \mathbf{D} between the luminance set of the target image and that of the original image.

$$\mathbf{D} = f(\mathbf{y}') - \mathbf{y}, \quad (4)$$

where $f()$ is the image processing function applied to the luminance set of watermarked image \mathbf{y}' .

The \mathbf{D} can be used for detection in place of \mathbf{s}' (equation (2)):

$$v = \frac{1}{n} \sum_k (D_{\alpha_k} - D_{\beta_k}) \quad (5)$$

However, it is still difficult to detect watermarks in content severely degraded by the image processing processes in the redistribution model, as described in section 4.1. For instance, a video encoder often causes geometric movement of the pixel data in video content. When pixel data in a watermarked image are moved geometrically by image processing, the pixel data in a specific location in the target image do not correspond to those in the same location in the original image. Therefore, the difference image between the original image and target image does not necessarily indicate a watermark pattern. However, the conventional method simply creates a difference image between the original and target images. As a result, it does not necessarily improve detection performance more than expected when watermarked content is severely degraded by image processing.

Moreover, it is difficult for auditors to use the original video content to detect watermarks due to copyright concerns, as described in section 2.4. However, if target images uploaded to a video sharing site violate the copyright, the copyright holders might cooperate with auditor in detecting watermarks. Copyright holders may allow use of the original images. If not, they may allow used of compressed original images in place of the original ones. We consider this possibility.

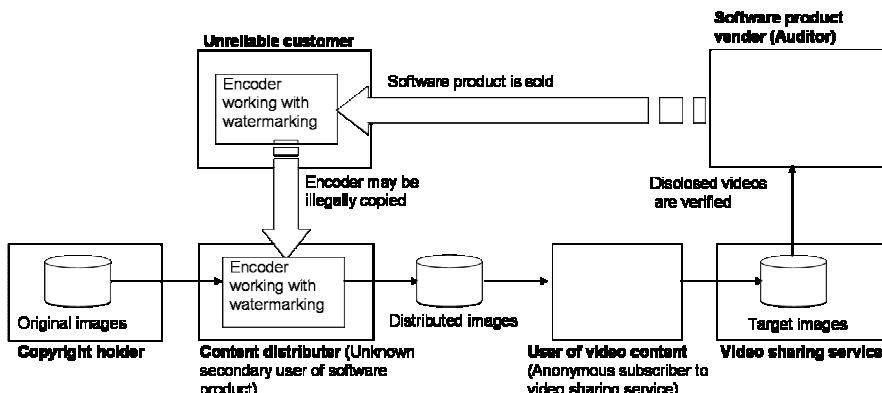


Figure 10. Use case of illegally copied video encoder.

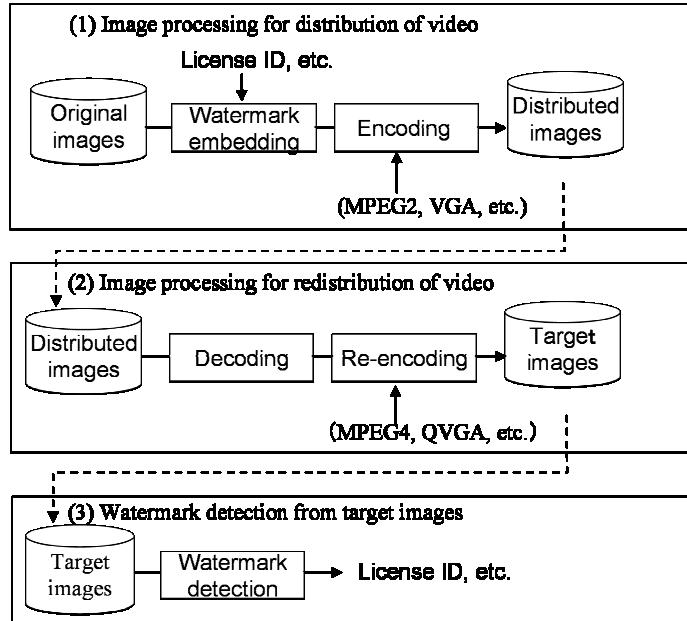


Figure 11. Video content re-distribution model.

In the case illustrated in figure 12, target images have been accessed by an anonymous user. The user is one of several users of a video content distribution service that uses an illegally copied encoder. One of the users has uploaded the target images to the site of a video sharing service. If an auditor finds the target images by chance on the video sharing service, the auditor does not know who initially distributed the content because content distributors often operate in a closed network open to only specific video users. Moreover, the auditor does not know whether the target images had been processed with an illegally copied video encoder. If the target images violate the copyright of the video content, the copyright holder might help the auditor detect the watermarks and find the illegal re-distributor of software product. The auditor could co-operate with a reliable content distributor who is a licensed user of an encoder that does not have a watermarking function. In cooperation with the copyright holder, the reliable content distributor could provide compressed original images to the auditor.

4.3. Proposed System for Video Watermark Detection Using Degraded Original Images

4.3.1. Research Approach

The content-specific signal is changed by the image processing applied to watermarked images. However, the changed content-specific signal can be calculated approximately by applying the same image processing to the original content. The approximated content-specific signal is called the "degraded original image". We previously reported an early concept of watermark detection using degraded original images [11]. Here we describe a

video watermark detection procedure that uses degraded original images and improves performance. There are four steps:

- (1) Estimate the image processing applied to the watermarked images.
- (2) Apply the estimated image processing to the original images.
- (3) Subtract the degraded original images from the target images.
- (4) Detect watermarks from the difference images.

Improved difference image \mathbf{D}' between the luminance set of the target image and that of the degraded original image is given by

$$\mathbf{D}' = f(\mathbf{y}') - f'(\mathbf{y}) \quad (6)$$

$$v = \frac{1}{n} \sum_k (D'_{\alpha_k} - D'_{\beta_k}) \quad (7)$$

where $f'()$ is an estimated function of the image processing likely applied to the luminance set of watermarked image \mathbf{y}' .

The difference image can be used for detection in place of \mathbf{s}' (equation (2)), as shown in equation (7). Experiments demonstrated that our proposed watermark detection procedure is effective for the various image-processing procedures in the redistribution model (figure 11).

4.3.2. Design of Proposed System

The proposed system, illustrated in Figure 13, is based on the redistribution model (figure 11). Encoding module Ep is a simulation of Bob's illegal copy of the encoder that was licensed to Alice (figure 2). Content processed using Bob's encoder is distributed (figure 11(1)). Note that Ep provides the same encoding function as Bob's encoder, but it does not work with watermarking. Therefore, compressed original images created by Ep do not have watermarks. \mathbf{P} is a set of encoding parameters used in Ep (MPEG2, VGA, and so on). Encoding parameter set \mathbf{P} should be adequately estimated by an auditor who detects the watermarks or by an entity cooperating with the auditor. The encoding styles used by various content distribution services are similar because they are concerned about the performance of end-user devices. This means that the default set of the parameters for encoder Ep can be estimated as \mathbf{P} . Otherwise, the auditor or cooperating entity could select a commonly used set of encoding parameters as \mathbf{P} .

Encoding module Eq is a simulation of the encoder used in the redistribution process (figure 11(2)). Estimated image processing $f'()$ of equation (6) is designed as consecutive processes of Ep , Dp , Eq , Dq , and scaling. Two videos (uncompressed degraded original images and uncompressed target images) that have been decoded by Dq should be resized to the size of the watermarked images before the subtraction process of equation (6). Encoding parameter set \mathbf{P} , including such size information, should be used again for this scaling. The process labeled "Difference" in figure 13 is the subtraction process. Watermark detection from the difference images is done by the "Watermark detection" process. Since the watermark embedding process of the proposed system is the same as that of a conventional

method, the watermark detection module (based on equations (6) and (3)) of the proposed method is the same as that of a conventional method (based on equations (2) and (3)), as shown by the dashed line in figure 13. That is, this system can detect watermarks from target images without using degraded original images.

Encoding parameter \mathbf{Q} should be adequately estimated by the auditor. Some of the encoding parameters given to the target images can be recognized by reading the attributes of the file headers of the target images. The readable attributes are file format, video codec, data size, playback time, number of streams, video frame size, color depth, bit rate, frame rate, interlace, aspect ratio, and comments (for instance, name and version of encoder software). Those parameters can be set automatically as parameter \mathbf{Q} in the proposed system. However, some parameters such as image quality need to be estimated by the auditor. For instance, the cheap encoders often used by hobbyists simply provide a useful graphic user interface that enables the user to adjust the encoded image quality simply by choosing alternative buttons (high and low). Video encoders compress video data by reducing redundant parts of the data. Determining which parts of the video are redundant depends on the encoder's implementation. Therefore, the auditor should estimate the parameters, except for the recognized ones. Although there may be other parameters in the encoder, the default settings were likely used, making the auditor's job a bit easier.

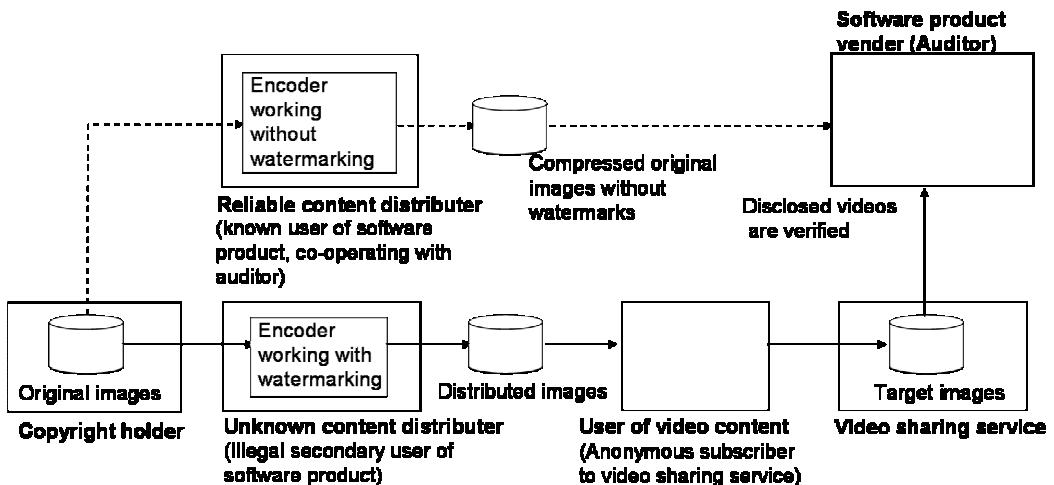


Figure 12. Watermark detection scheme using compressed original images.

4.3.3. Prototype

The watermarking algorithm in the prototype is based on the patchwork method (described in section 2.3). We developed a watermark-embedding filter and a watermark detection filter in accordance with the standard video interface provided by the operating system (OS). Various image-processing filters working for the same video interface are also provided by the OS. Users can combine the watermark embedding and detection filters with various image-processing filters such as encoders or decoders by using a graphical user interface (GUI).

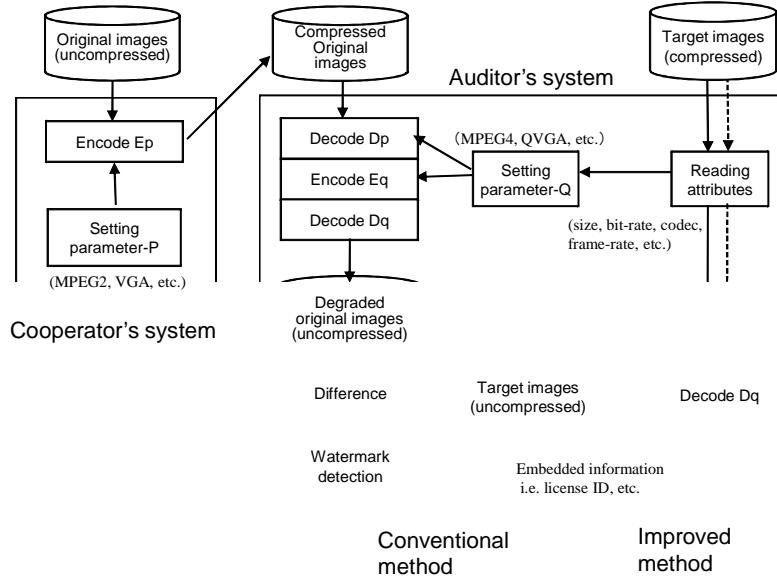


Figure 13. Proposed watermark detection system using degraded original images.

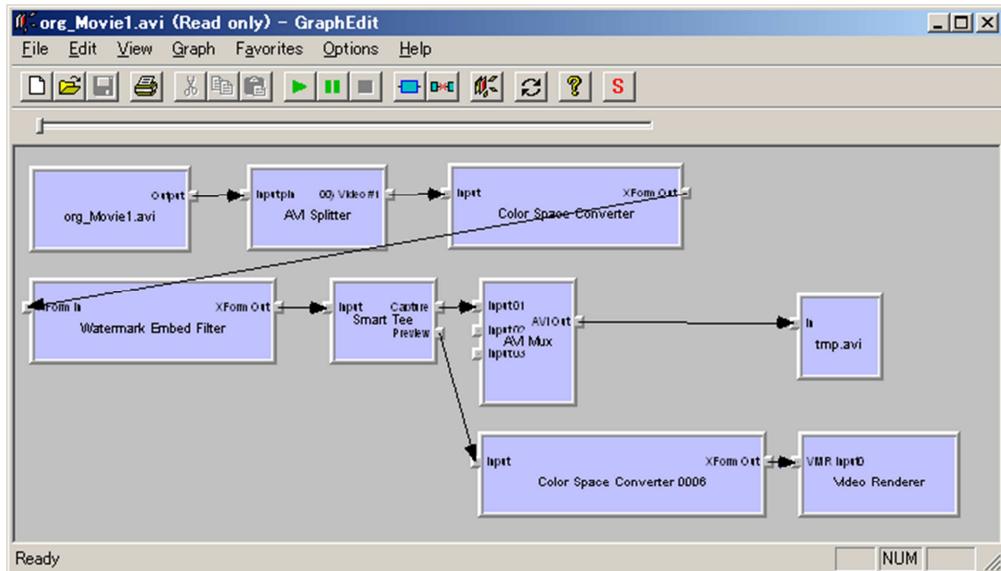


Figure 14. Graphical user interface sample for embedding watermarks.

The watermark-embedding filter works by repeating three steps as long as the video stream continues: (a) receive video frame from the filter that previously accessed the original video file, (b) embed watermarks into the video frame, and (c) send the watermarked video frame to the next process. The output video frames can be converted into a video stream file on a hard disk through several filters such as an encoding filter. Moreover, the output video frames can be simultaneously displayed by making a tee-branch to a rendering filter at the mid point of the video output stream. A sample screen shot of this procedure is shown in figure 14.

The watermark detection filter also works with several pipelined filters in a way similar to the usage of the watermark-embedding filter. For instance, a typical video player has consecutive processes to open a file, de-multiplex the video and audio, decode the bit-stream, and render the frame images. A video player working with watermark detection can be composed by inserting a watermark detection filter between the decoding and rendering filters. Various kinds of video files can be decoded and processed with the watermark detection filter if the video codec interface is the same as that of the developed prototype.



Figure 15. Sample images: EntranceHall (left) and WalkThroughTheSquare (right).

4.4. System Evaluations

4.4.1. Experimental Conditions

Two samples (shown in figure 15) were selected from standard videos [18] used for image quality testing (15 seconds, 450 frames, VGA). EntranceHall is a fixed-camera shot in which people in the foreground are slowly moving. WalkThroughTheSquare is a dolly shot of people walking.

Watermark detection was done for each frame and was considered successful if all 64 embedded bits were extracted without any bit errors. The watermark detection ratio is the number of successful detections divided by the number of detection chances. The image processing procedures were, in order, encode (E_p), decode (D_p), re-encode (E_q), and re-decode (D_q) the original and watermarked images. Encoding parameter sets **P** and **Q** were

Parameter **P**: VGA, MPEG2, 6Mbps

Parameter **Q**: QVGA, MPEG4, 500kbps.

Since encoder E_p simulated the illegal copied encoder (figure 11(1)), the auditor knew E_p but not **P**. The effects of changing **P** on watermark detection performance were less than those of changing **Q** in a preliminary test. Therefore, to simplify the testing, we assumed that encoding parameter set **P** used by the auditor (or cooperating entity) was the same as that used for the illegal copying. Since encoder E_q simulates the encoder used by the video subscriber in the anonymous user environment (figure 11(2)), E_q was unknown to the auditor. If the header information in the target images gave the auditor information related to

Eq , the auditor might correctly estimate the encoder used in the video subscription. To simplify the testing, we assumed that Eq was correctly estimated.

Encoding parameter set **Q** for Eq was unknown to auditor, and some of the parameters in the set often conflicted with each other. For instance, high image quality and high compression ratio generally have a trade-off relationship. The priority of each parameter to be processed by the encoder depends on the encoder's implementation and the content. Encoders often neglect some encoding parameters when the compression ratio is as high as it was in our testing. In preliminary tests of an encoder, the encoding parameter for image quality significantly affected watermark detection performance while others such as GOP (group of pictures)-length did not. Therefore, we focused on the image-quality encoding parameter as an example of an estimated parameter in set **Q**. We used only two image qualities (high or low). An auditor should be able to estimate whether the image-quality encoding parameter was set to high or low. Some parameters can be automatically recognized from the file headers. The other estimated encoding parameters would be correct if both the auditor and illegal user used the default settings for them. We assumed that the encoding parameters besides that for image quality were estimated correctly.

The image quality of videos encoded using Ep was evaluated from the viewpoint of practical distribution (figure 11(1)). The robustness was evaluated by detecting watermarks in video re-encoded by Eq (figure 11(3)). The effect of estimated parameter set **Q** on watermark detection performance was also evaluated.

Table 2. Evaluated image quality

Sample	Score
EntranceHall	4.8
WalkThroughTheSquare	4.6

4.4.2. Image Quality

We embedded watermarks into the original video files and then encoded the files using encoding parameter set **P** (VGA, MPEG2, 6 Mbps). The original images were also encoded using **P**. Those MPEG-2 video files, the original and watermarked ones, were assumed to be the distributed images (figure 11(1)). We subjectively evaluated the image quality of the MPEG-2 videos using a procedure based on Recommendation ITU-R BT.500-11 [19]. The encoded original images and encoded watermarked images were displayed on a monitor and evaluated by ten participants, who rated the image quality using the scale shown in table 1 (5 for imperceptible, 4 for perceptible but not annoying, 3 for slightly annoying, 2 for annoying, and 1 for very annoying).

As shown in table 2, the average scores for the two samples exceeded 4.5. A 4.5 is equivalent to the score for the case in which half the participants did not perceive any watermarks and the other half rated the quality to be 4 (perceptible but not annoying). A score of 4.5 is generally considered to mean that the evaluated images are practically useful. Therefore, the image quality of the watermarked video files (the distributed images in figure 11) was maintained for practical use.

4.4.3. Robustness Compared with Conventional Methods

We define following three methods:

- M1: A conventional method based on equation (2) for detecting watermarks in target images without using the original images [5]
- M2: A conventional method based on equation (4) for detecting watermarks using the difference between the target image and the original image (not degraded original image) [17].
- M3: The proposed method (based on equation (6)) for detecting watermarks using the difference between the target image and a degraded original image.

All three methods assume that embedding is done in the same manner, as described in section 2.3. In an experiment using these methods, the same watermark detection filter (based on equations (2), (5), and (7)) was used to decide whether the result extracted from the watermarked image was a bit value or not.

The same images used in the image quality test (described in section 4.4.2) were processed in accordance with each method, M1, M2, and M3. The auditor was assumed to correctly estimate the image quality parameter in encoding parameter set \mathbf{Q} . As shown in table 3, for the EntranceHall image, the detection ratio with the proposed method (M3) was 73.1%, an improvement of 34.4% compared to the conventional method without using the original image (M1). The detection ratio of conventional method M2 was 2.0%. For the WalkThroughTheSquare image, the improvement was 8.9% (M3 vs. M1).

Table 3. Detection ratio

DetectionMethod	M1	M2	M3
EntranceHall	38.7	2.0	73.1
WalkThroughTheSquare	18.9	0.0	27.8

M1: Conventional method without the original image [5]

M2: Conventional method with the original image [8]

M3: Improved method with degraded original image

Since M2 (using the original image) uses more information for detecting watermarks than M1 (without using the original image), the potential detection performance of M2 should be no less than that of M1. However, our experimental results showed that it was less than expected. Why? If the pixel values of original image were not in accordance with those of the target images, the difference data would be noisy and interfere with watermark detection, thus degrading detection performance. One reason it was not better is that the encoder might have caused complex changes (such as location shifts or strains) in the encoded images. It is difficult for an auditor to cancel the effect of such changes in images, even with manual adjustment. Therefore, it is difficult to achieve stable watermark detection using conventional method M2 in the research scope of this work.

These problems should also affect the performance of our proposed method. It is preferable not only for M2 but also for our proposed method to adequately match the original image and target images. With M3, an auditor estimates the image processing applied to the target image and applies it to the original image. If the estimation is adequate, the changes

made to the original image are almost the same as those made to the target image. Therefore, changes such as location-shift in the target image are likely to also be in the degraded original image. With M3, the location gaps should thus be automatically adjusted.

4.4.4. Effect of Estimated Parameter on Detection Performance

The image quality parameter in encoding parameter set \mathbf{Q} should be estimated by the auditor as described in section 4.1. Watermark detection performance using M1 (conventional method without using original image) and M3 (proposed method) was evaluated in terms of the adequacy of estimating image quality. The encoded watermarked video (the same one used in the subjective image quality test) was decoded and re-encoded using each parameter for image quality (high or low). The re-encoded videos were the target videos. The degraded original video was also processed in the same manner.

Table 4. Watermark detection performance for Entrance Hall

Image quality in target images	M1	Improved method (M3)	
	Detection ratio (%)	Estimated image quality in degraded original images	Detection ratio (%)
Low	38.7	Low	73.1
		High	61.3
High	27.3	Low	59.6
		High	60.4

M1 and M3 are the same as those in table 3.

As shown in table 4, when the parameter for image quality used in creating the target image was 'Low' and the parameter was correctly estimated, the watermark detection ratio using M3 was 73.1% and using M1 was 38.7%. That is, if the auditor estimates the parameter correctly, the watermark detection ratio is improved by 34.4%. When the parameter used in creating the target image was 'Low' and the parameter was incorrectly estimated, the detection ratio using M3 was 61.3%. That is, even if the auditor estimates the parameter incorrectly, the detection ratio is still improved by 22.6%. The effect of estimated parameter set \mathbf{Q} on watermark detection performance is 11.8%, which is the difference between 73.1% and 61.3%. When the image quality parameter in the target image is 'High', the effect of estimated parameter set \mathbf{Q} on watermark detection performance is negligibly small. However, with the proposed method, the watermark detection ratio is improved by more than 30% whether or not the parameter is estimated correctly.

Watermark detection performance for encoded images with high quality is typically better than that with low quality. However, in our experiment, the watermark detection ratio for target images that were encoded using 'Low' image quality was larger than that using 'High' image quality. This was apparently due to the encoder we used. From the viewpoint of the encoder specifications, images with high quality should be produced when the encoding parameter for image quality is set to 'High'. However, the actual encoded images were so severely degraded from the original images that the evaluators had difficulty determining subjectively which image had better quality. Therefore, it is difficult for an auditor who views only the target images to estimate the image-quality encoding parameter used for encoding

the target images. Nevertheless, the proposed method improves watermark detection performance even if the estimate is not correct.

4.4.5. Summary of Evaluation for Video Watermark Application

Implementation of the proposed detection system does not require updating the existing watermark embedding system because it assumes that the conventional embedding method is used. Moreover, previously watermarked content does not need to be watermarked again. The proposed system is easily installed by appending a new detection apparatus to the existing system.

If the watermarked images are not severely degraded, the conventional detection method M1 is still effective, as described in section 4. Moreover, the usability of the conventional method is better because its processing time is shorter than that of the proposed detection method. Therefore, using a combination of the conventional and the proposed detection methods should be effective. If watermarked images are severely degraded and if the watermarks cannot be detected using the conventional detection method, an auditor can try to detect them using our proposed method. Experimental results showed that our proposed method improves the watermark detection ratio by more than 30% whether or not the image quality parameter is estimated correctly, as shown in table 4.

5. Conclusion

When distributed contents digitally watermarked with the serial number of the software product that generated them are found, the watermarks can be used to determine whether the software product was illegally copied by a licensed user. It is difficult, however, to detect watermarks in content that has been severely degraded by signal processing. We developed two practical digital watermarking applications using a reference signal similar but not identical to the original content: a text-to-speech interface working with audio watermarking and a video encoder working with video watermarking.

Synthesized voice can be easily re-synthesized from the target watermarked voice data by listening, but the re-synthesized data may differ from the original data. Re-synthesized voice can improve detection performance by making detection time frames shorten. Even if a watermark is not detectable using a long detection time frame, a less-affected small part of the audio data could offer other chances to detect the watermarks for a short detection time frame. Experimental results obtained using re-synthesized voice as a reference signal showed that the proposed method can detect watermarks in MP3 data better with a 2-second time frame than a conventional method does with a 30-second time frame.

Compressed original video can be used for watermark detection. Detection performance can be improved by (a) estimating the image processing applied to the target video, (b) applying the estimated processing to the compressed original video, (c) subtracting the degraded original video from the target video, and (d) detecting watermarks from the difference video. Experimental evaluation using a re-distribution model, in which a watermark survives MPEG-2 encoding and MPEG-4 re-encoding and downscaling from VGA to QVGA, showed that the proposed method works well, i.e., the detection ratio was improved by up to 34.4%.

Because the proposed detection methods assume the embedding method is the same as that used in conventional detection methods, it can not only detect watermarks without using a reference signal but can perform better by using a reference signal. Therefore, if watermarked content is severely degraded so that the watermarks cannot be detected using a conventional detection method, an auditor may be able to detect them using the proposed method. Identifying the software product by detecting the watermarks will help the auditor identify the licensed user. The auditor can then determine whether the software was illegally copied. If it was, the auditor can make a claim for damages. Therefore, identifying the software product will help auditors limit the extent of damages due to illegal copying and help them recover damages due to illegal copying.

Future work includes development of enhanced systems that maintain the audibility and visibility of digitally watermarked content. We believe that our idea of using a reference signal is effective even if another watermark algorithm is used. Moreover, the idea can also be applied to other types of content (other than audio and video) by using a method similar to the proposed one. We thank Hitachi Business Solution Co., Ltd. for lending us their high-quality text-to-speech product, "Voice Sommelier" [14].

References

- [1] Cohen, F. B. (1993). "Operating system protection through program evolution", *Computers and Security*, Vol. 12, No. 6, 565-584.
- [2] Rosenblatt, B., Trippe, B. & Mooney, S. (2001). "*Digital Rights Management, Business and Technology*", M&T Books.
- [3] Nagra, J., Thomborson, C. & Collberg, C. (2002). "A functional taxonomy for software watermarking", in Proc. of *Australasian Computer Science Conference (ACSC2002)*, Vol.4, 177-186.
- [4] Cox, I. J., Miller, M. L. & Bloom, J. A. (2001). "*Digital Watermarking*", Morgan Kaufmann Publishers.
- [5] Bender, W., Gruhl, D., Morimoto, N. & Lu, A. (1996). "Techniques for datahiding", *IBM Systems Journal*, Vol.35, No. 3&4, 313-336.
- [6] Boney, L., Tewfik, A. H. & Hamdy, K. N. (1996). "Digital watermarks for audio signals", in Proc. of IEEE Int'l Conf. on Multimedia Computing and Systems (ICMCS96), 473-480.
- [7] Nishimura, A. (2008). "Data Hiding for Audio Signals that are Robust with respect to Air Transmission and a Speech Codec", in Proc. of Int'l Conf. on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP08), 601-604.
- [8] Yeo, I. K. & Kim, H. J. (2003). "Modified patchwork algorithm: a novel audio watermarking scheme", *IEEE Transactions on Speech and Audio Processing*, Vol. 11, No. 4, 381-386.
- [9] Hiratsuka, K., Kondo, K. & Nakagawa, K. (2008). "On the Accuracy of Estimated Synchronization Positions for Audio Digital Watermarks using the Modified Patchwork Algorithm on Analog Channels", in Proc. of Int'l Conf. on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP08), 628-631.

- [10] Hamada, D. & Unoki, M. (2008). "A Study on Audio Watermarking Method based on the Cochlear Delay Characteristics", *Journal of Signal Processing*, Vol. 12, No. 4, 315-318.
- [11] Yamada, T. & Sato, Y. (2008). "A Case Study of Digital Watermark Detection Errors for Video Integrity Verification", *Information System, Technical Report of the Institute of Electrical Engineers of Japan(IEE)*, IS-08-51, 85-88, (Japanese).
- [12] Hofbauer, K. & Kubin, G. (2006). "*High-Rate Data Embedding in Unvoiced Speech*", in Proc. of Int'l Conf. on Spoken Language Processing (ICSLP/Interspeech2006), 241-244.
- [13] Wang, C. T., Chen, T. S. & Xu, Z. M. (2004). "A Robust Watermarking System Based on the Properties of Low Frequency in perceptual Audio Coding", *The Institute of Electronics, Information and Communication Engineers (IEICE) Trans. Fundamental*, Vol. E87-A, No. 6, 2152-2159.
- [14] Takao, M., Sato, M. & Taketomi, S. "A General Intelligent Text to Speech System 'Voice Sommelier'", *Hitachi Business Solution Journal*, Vol. 3, 5-10 (Japanese).
- [15] Yamada, T., Takahashi, Y., Ebisawa, R., Sato, Y. & Susaki, S. (2009). "Evaluation of audio watermark system using short detection time frame", in Proc. of *IEEE Int'l Conf. on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP2009)*, 44-47.
- [16] Fujii, Y., Echizen, I., Yamada, T., Tezuka, S. & Yoshiura, H. (2004). "An Improvement of Error Correction Coding for Digital Watermarking", *Journal of Information Processing Society in Japan*, Vol.45, No.8, 1980-1997, (Japanese).
- [17] Lancini, R., Mapelli, F. & Tubaro, S. (2002) "A robust video watermarking technique for compression and transcoding processing", in Proc. of *IEEE Int'l Conf. on Multimedia and Expo (ICME2002)*, Vol. 1, 549-552.
- [18] The Institute (1999). of Image Information and Television Engineers, "Evaluation video samples".
- [19] Rec. (2002). ITU-R BT.500-11, "Methodology for the subjective assessment of the quality of television pictures".
- [20] Linnartz, J. P., Kalker, T. & Depovere, G. (1998). "Modelling the false alarm and missed detection rate for electronic watermarks", *Lecture Notes in Computer Science*, Vol. 1525, 329-343.
- [21] van Oorschot, P. C. (2003). "Revisiting Software Protection". in Proc. of 6th Int'l Information Security Conf. (ISC 2003), pp.1-13, Springer LNCS 2851.

Chapter 3

A CASE OF MIDDLEWARE FOR INFORMATION SYSTEMS OF PUBLIC TRANSPORT BY ROAD

Carmelo R. García, Francisco Alayón and Ricardo Pérez

Department of Computer Science and Systems
University of Las Palmas de Gran Canaria

Abstract

A practical case of use of the middleware is explained in this chapter. The contents are related with the fields of the ubiquitous computing, and more specifically with the theory and operation principles of the middleware, and the intelligent transport systems. The chapter describes how the theory and operation principles of the middleware can be applied in order to resolve a traditional problem of the information systems of the public transport corporations. Specifically, we explain the design and operation of a middleware, this middleware permits a proper integration of mobile information system of the vehicles of public transport corporation of passengers. In this context, the concept of proper integration means that the all the information related with the vehicles operation is available at time, real time, and the required amount by all the processes of the information system of the transport company. The description will be based on the theoretical and operation principles of the middleware, these principles are: context modeling, spontaneous interaction, context-triggered task management and development system for ubiquitous applications. Other relevant aspect of the middleware consists of that its design rely principles of network administration systems in order to control and administrating on automatic and unsupervised way the mobile information systems. The first point of the chapter is the introduction and it is dedicated to describe the main aspects of the mobile information system in the public transport context. The second point of the chapter is dedicated to describe the technological bases of the middleware, specifically the mobile communication infrastructures and the ubiquitous computing paradigm. The third point is the kernel of the chapter, in this point the formal description of the middleware will be presented. Finally, we will explain how to achieve several functionalities, commonly required in the operation of the vehicles of a public transport company, using the middleware.

1. Introduction

The advances in the information and communications technologies have allowed to automate processes in the information systems of the organizations and to put in practice models that integrate the different stages of the operations of the information system. Because of use of these advances, information systems are capable of collecting of the data from the activity of the organization and generate useful information for its operation.

In the case of the companies of public transport by road, where productive activity is performed in geographically dispersed areas using mobile systems, those advances have been introduced in a slow way, this slowness has occasioned that this mobile systems be weakly integrated in their corporative information systems. The accomplishment of its productive activity requires data that inform about where, when and how this productive activity is carried out. The result of this activity provides critical data that allow calibrating the performance of the performed activity. This data flow needs a transport media and later, once has been analyzed and integrated into the data of the organization, to be transformed in useful information that contribute to know and to improve the organization's activity.

New computation devices has been placed on board the vehicles that increase the level of automation of task performed into them and the capacity of recording data obtained from its operation. On the other hand the advances in wireless communications make possible the automatic transport of the data generated in the movable systems, avoiding so these data be carried by the personnel of the company, and as a result of it, making possible the immediate access to the data and in the proper amount in order to dispose of a better knowledge of the state of the production. Therefore, in the field of public transport systems, arises the challenge to develop telematic tools that take advantage of information and communications technologies to improve the operation of these systems, solving traditional problems and adding new functionalities. The major goal is to improve this crucial public service and increase its consideration by the citizens. In this chapter is exposed the architectural model that would allow to develop these tools.

The initial hypothesis of this model is the affirmation that it is possible to dispose of an information system that integrate a group of movable devices based on standard technologies which communication costs be acceptable. The chapter describes the architecture of the information system, based on ubiquitous computation model where all sections of the organization are taken into account. In the system has special relevance the integration of the movable systems (for example computers on board of the vehicles and hand held devices) in order to be considered at same level of relevance as any other information device of the organization network. Other aspects to emphasize are two characteristics of this architecture, that distinguish it from the traditional information systems operating in actual organizations in the field of public transport by road, first, it is intended to be open to the use of standard technologies, and second, it has an scalable structure, all of which make easy the adding of new elements and adapting to the changes of the future.

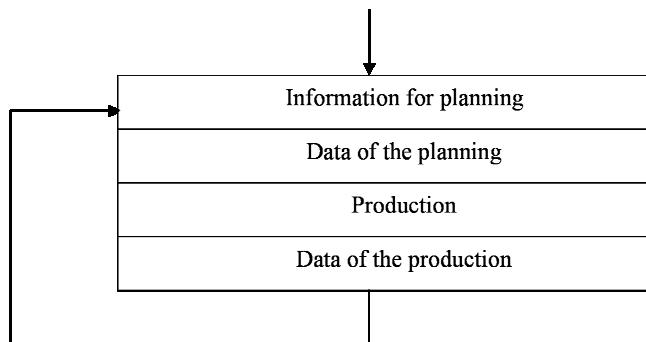


Figure 1. Information's life cycle.

This chapter has by purpose to explain the architectural model that allows the production systems be gracefully integrated into the information system of the organization. This architecture allows the information flows by the different sections of the organization at the convenient time, with security and in the amount the best adapted to the requirements, improving this way the quality of the management. Specifically, this architecture fulfills the following goals:

1. Easy integration of current standard tools into the architecture, allowing the company to be more independent from dealers at same time is updated in the use of technological advances.
2. It is based on open systems specifications, in order to facilitate the easy expansion to new components and services at effective cost.
3. To facilitate the development of specific solutions economically suitable, taking advantage from standard technology and making a reasonable use of the communications.

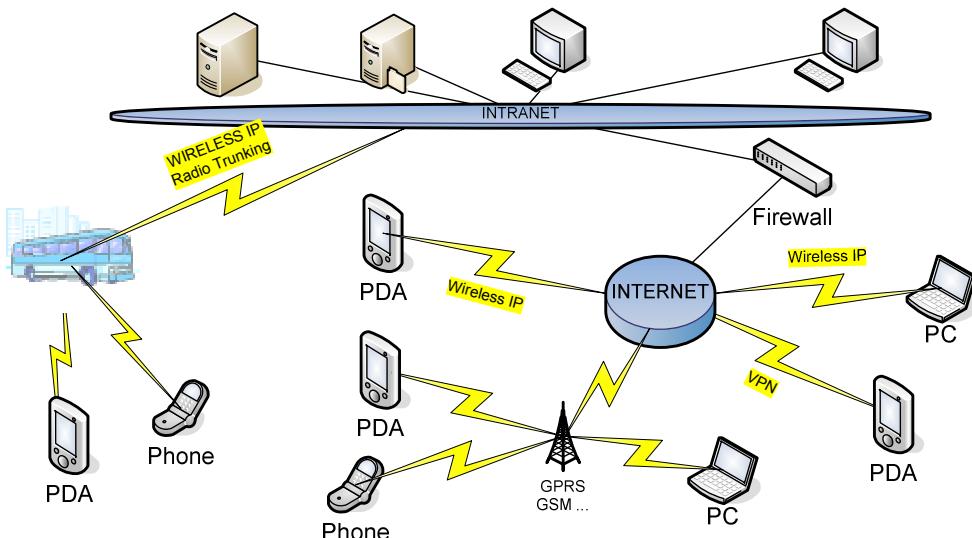


Figure 2. General vision of the system.

2. General Description of the System

In 1991 Mark Weiser introduced the ubiquitous computing paradigm as the computing paradigm of the 21st century [1]. But it is nowadays, with the generalized use of the mobile computing and communications technologies, when this computing model is been applied to solve real problems. The key goal of this paradigm is to develop computer systems adapted to the persons in order to achieve natural interaction between persons and systems.

In this chapter a successful case of use of this paradigm is described; specifically this description consists of an explanation of a middleware for information systems of the public transport corporations by road. With this middleware, the mobile onboard information systems, installed in the buses of the transport company, are integrated in a proper way in the information system of the corporation. This integration permits to make important tasks of automatic way, to facilitate the incorporation of new functionalities and to use quality data. Quality data means data processed in real time and data obtained in the proper amount. The figure two shows a general vision of the system.

The middleware executes in a ubiquitous system, figure three shows the functional structure of this system. In the top level the ubiquitous programs use mobile computing services, these programs are executing in mobile systems located in a disperse geographic area where the company works .These mobile systems have all the hardware and software resources required to work in autonomous way; it means that the permanent connection with the central services is not required. The mobile systems must to control the production operations and the technical status of the vehicles and if an exception occurs, then the communication with the control center of the company is made. In this context, an exception is any event that can affect to the operations planning of the vehicles or to the proper work of the onboard resources (hardware and software). Similarly, the control center can act in the vehicle mobile information system using the communication infrastructure. The role of the middleware is to provide transparent communication channels to the ubiquitous programs not dependent on the kind of communication infrastructure used, for example, long distance mobile communication infrastructure versus local mobile communication infrastructure. Another role of the middleware is to enable the spontaneous interaction between ubiquitous programs using the available communication infrastructures. The localization of the mobile systems is very important for the middleware, because this information can use to choose the specific of communication infrastructure to use in order to fulfill requirements related with the operational cost of the system, for example, if the positioning of the vehicle is inside of a area covered by the local wireless network, then the data transfers can be accomplished using this infrastructure because the use of it is free from the economic cost point of view. The positioning information of the mobile system is provided by GPS and this resource provides a common timer for all the equipments integrated in the information system. Of this way, distributing functionalities is performed in mobile and no mobile systems of the company.

In general, a ubiquitous computing system is formed by the following elements: the ubiquitous devices, the ubiquitous network, the middleware and the ubiquitous programs, Saha [2]. The communications channels required by the different ubiquitous programs are provided by the middleware, the goal of this chapter is to explain this middleware. In our context, public transport of passengers by road, the ubiquitous programs develop all the

production, control and information tasks related with the vehicles activities, so these programs can be classified as:

- Production programs, these perform all the tasks required to provide the services offered to the transport clients.
- Control programs, these perform the supervision and control of the operations of the vehicles in order to guarantee the fulfillment of the planning.
- Maintenance programs they supervise the hardware and software resource in order to detect technical failures and to update, in an unsupervised and automatic way, the data and programs of the different systems.
- User information programs, they are responsible to inform to the clients about the services of the company, for example timetables, incidents of the services, payment systems, etc.
- Transactional programs, they have to perform the automatic transfer production data of the vehicles to the control center.

2.1. Ubiquitous Devices

By these elements the users can interact with the system. In the system, figure one, PDAs, cellular phones, contactless card, mobile personal computer, etc. can be used to access to the services provided by the information system of the transport company. For example, the transport clients can use the contactless cards or the cellular phones to pay the trips or the maintenance staff uses PDAs or personal mobile computers to interact with the onboard systems. These onboard systems, and more specifically the computer installed in the vehicles, has the more important role between the ubiquitous devices in the system architecture because of it plays a main role with respect to other system elements. It acts as storage element of the all the production data generated in the vehicle until the data can be transferred to the control center. To achieve this transfer, the services of the middleware run-time are used. The onboard computer integrates the onboard system with the corporation network and, additionally, it controls to other onboard devices (contactless cards readers, information panels, driver console, etc) permitting an intelligent management of the rest of onboard elements (hardware and software). From the point of view of the connection structure, the onboard system has a star topology, where the central point is the onboard computer; it is connected to the following elements: payment system (driver console and card readers), positioning system (GPS), communication system and physical monitoring system (for example: power and temperature sensors).

2.2. Ubiquitous Network

In a ubiquitous context, the network is responsible of the data transfers between ubiquitous programs which are executing in mobile and fixed devices. In the system, to control the connections of the mobile devices, a fixed device plays the role of communications node connecting the transport company network to the mobile devices. The communications node routes the user requests, it is responsible of the reception of the data

from the mobile devices and sending to their destinations. The ubiquitous network is formed by different communication infrastructures: the long distance mobile communications is achieved by a radio trunking systems, the local mobile communication is achieved by IEEE 802.11 wireless networks operating in a set of sites located in the geographic covered by the transport company (station, garages, public center, etc.) and very short distance is achieved by Bluetooth and RFID networks operating in the vehicles.

The services of a public provider are used for long distance mobile communication, specifically a public radio trunking system, which covers the entire geographic area where the transport company operates. The reason to use a public infrastructure and not a proprietary infrastructure is that the use of a proprietary solution implies high costs from a point of view economic and installation. In this kind of infrastructure, the availability and the quality of service are related directly with the numbers of relay stations of the infrastructure. The option based on private communications infrastructure has two drawbacks: the difficulty to adopt the advances in the communications technology and the return on investment. The advances of the communications technologies are made quickly, for this reason the public communications providers offer frequently new products and service of added value with attractive costs. Additionally, the costs of the maintenance and renovation of the private infrastructure are high. In the opposite side, the election of a public provider of communications implies low costs of infrastructure; normally only the related with the communications users terminals. When the communications services of a public provider are used, the clients share the use of the infrastructure, normally it is of high capacity, paying only the use of it. The popularity of the public mobile communications services and the competition between providers has produced a wide products spectrum adapted to the clients' requirements with costs based on use thresholds.

There are some solutions to achieve the long distance mobile communications. The simplest is the PMR infrastructure, in this infrastructure the sender uses a fixed frequency and the messages are received by all the receptors that use the same frequency. The PMR is proper for short geographic areas and the main advantage is the permanent availability of an emergency channel.

Other solutions for long distance mobile communications are based on the competition for the radio-electric space; a case of this type is the radio trunking system. It follows the classic telephony model, where a set of channels is available for the users and only one of them is assigned when it is necessary. The goodness of this solution depends on the available channels number and the probability of channel availability. Following this scheme we find the radio cellular systems that have developed from the initial GSM until the currently UMTS. These systems provide wide coverage and improved speed. This speed increase has permitted to provide parquets commutation services, making possible to use popular protocols such as TCP/IP protocols, of this way, the applications are independent of the communication infrastructure and standard applications and services can be used. The costs of the use of the long distance communications infrastructure motivate us to achieve the goal of minimizing the communications costs by a rational use of this resource. In the transport context, an analysis of the information requirements show us that there are process which require real time communications, these are the process related with control planning and technical alarms.

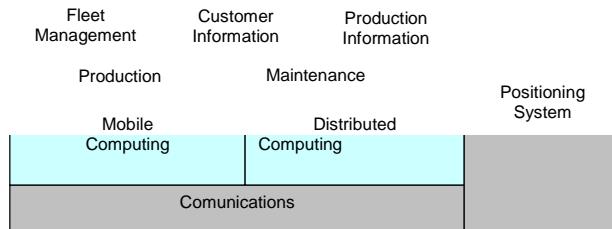


Figure 3. Functional structure of the ubiquitous system for transport.

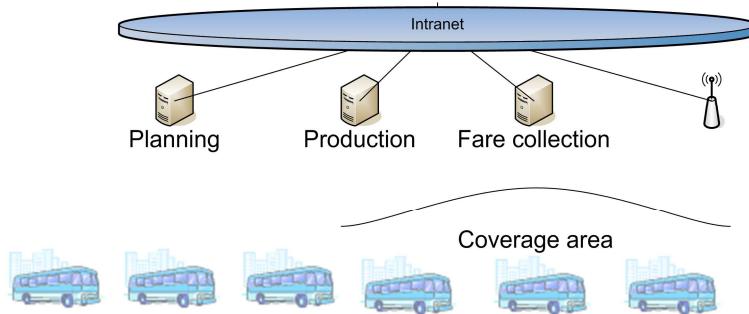


Figure 4. Integration of the mobile onboard system using Wifi infrastructure..

There are other tasks that don't require real time communications, the data are stored in the onboard computer and later they are transferred to the control center, for example the production data transferences, the updates of data and programs of the onboard systems, etc. For this type of process the ubiquitous network provides the local wireless network infrastructures, IEEE 802.11, installed in relevant places of the transport network of the company, normally these places are geographic points where there are a high concentration of buses frequently, for example: stations, garages, bus stops, etc. This network infrastructure provide a higher communications rate than the offered by public long distance communications infrastructures, permitting the massive data transferences between mobile systems and not mobile systems.

Therefore, the ubiquitous network integrates a mixed infrastructure managed by the middleware; it decides what type of infrastructure to use considering different criteria such as: infrastructures availability, costs, relevant of the data to be transferred, etc.

2.2.1. Data Transference

The middleware uses two communication modes: direct communication and communication by mailboxes. The election about mode to transfer the data depends on the priority of the process involved. When the process has a high priority, the direct communications channel is chosen using the proper network infrastructure. Therefore, when the communications have not priority, then the middleware uses a communication scheme based on mailboxes where first the data are stored in mailbox and latter they are sent. It is a distributed communication system specially designed for mobile agents. Nowadays is frequently to find systems that use mobile agents in fields such us ecommerce, e-government, network administration, etc. In general, mobile agents play an important role in distributed

systems and networks. A mobile agent is an element which works moving in the information system on an autonomous way, using the resources provided by the system, specially the required to work in mobility in an integrated manner. To guarantee the communications of the agents its necessary mobile communications protocols. Several mobile protocols have been proposed based on different principles and requirements, for this reason there is not standard protocol to communications mobile agents existing different methodologies to analyze and comparing between them.

The system uses a structured protocol based on mailboxes where the mobile agents can be executed in different stations to the stations where its mailboxes is placed, that is the migration of agents and mailboxes is independent. The system permits the automatic data transfers between agents in order to admin on an automatic and unsupervised way all the elements integrated in the information system. The general vision is a system integrated by elements that are used of a uniform way independently of the mobility degree of the element. The mobile systems, specially the onboard systems, move around a large geographic area and data transfers are spontaneous and automatic.

The principles of design of the system protocol are the following:

- Local transparency: the mobile agents are integrated on spontaneous way in the network using mobile stations, the receipt and transmission of data is made regardless of the location of the station.
- Asynchrony: in order to guarantee the data receipt, the system must to synchronize the transmission and receipt data between agents and this synchronization must not to limit the mobility of the agents.
- Reliability: the system must guarantee that the data arrive to the destination agents, regardless of the migration level of the agents.
- Efficiency: In order to optimize the communication costs for migrations and delivery operations, parameters such as distance, messages number and messages size must be used.

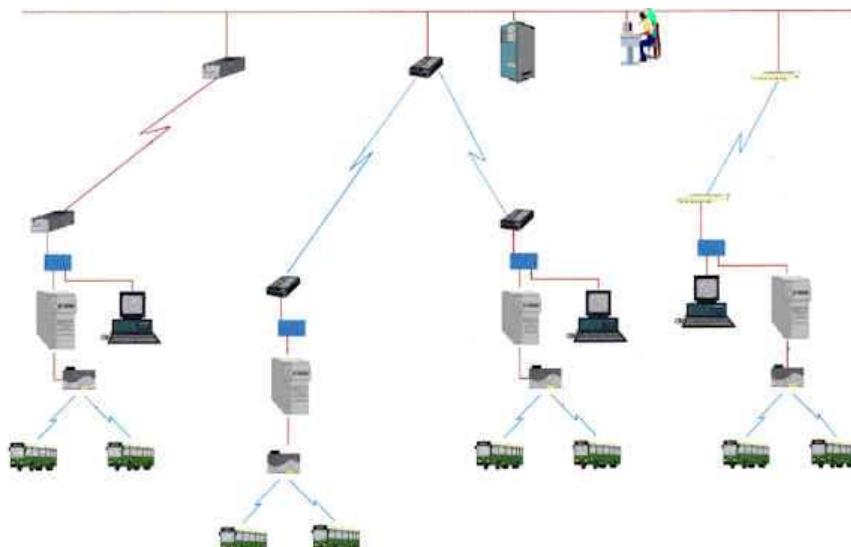


Figure 5. General vision of the transport company data network.

2.2.2. Scheme Based on Mailboxes

The model based on mailboxes used by the system protocol is explained in this section. The scheme defines three components: the mobile agents, the mailboxes and the mobile systems. Each mobile agent has a mailbox where messages are stored. The agents are autonomous and mobile and the mailboxes are mobile but non autonomous because they cannot determine where to migrate. A mobile agent can send messages to the mailbox of another agent. In general, to deliver the messages, two operations must be made; a pull operation or a push operation. The pull operation consists of sending of messages to the mailboxes of the destination agents and the push operation, for destination agents, in order to request its messages. So, to achieve the communication between two agents, two operations are executed, the first is to send a message to the mailbox of the receipt agent and, the second, is to send the messages from the mailbox to the destination agents. To apply these principles in the system, the following configuration has been defined: an agent can be any program that can send or receive messages. When the system, where the agent is executing, is mobile then the agent is a mobile agent too. To avoid the loss of messages during the transmissions and forward of messages, a high level and reliable network service is used. In the figure five a general vision of the system is presented, the mobile agents, which are executing on the vehicles systems, are integrated spontaneously in the network and the agents mailboxes are in non-mobile systems of the infrastructure.

To explain of systematic way the system protocol, we use the model introduced by Cao [3]. Basically, this model establishes three basic aspects in order to study and compare the mobile protocols based on mailboxes. These aspects are: frequency of the mail box migration, messages delivery and synchronization for migration and delivery messages.

- **Mailbox migration.** This aspect refers to the possibility of migration of the mailbox; three types of mailbox migration are defined:

No migration (NM): this case the mailbox stays in the same system permanently; commonly this system is called home system. It means that the mobile agent can migrate to different agent platforms. All the messages must be sent to the same home but these messages must be forwarded from the home to the mobile agent.

Full migration (FM): in this case the mailbox is considered part of the mobile agent and so, it migrates with the mobile agent.

Jump Migration (JP): The agent and mailbox can migrate separately. The mobile agent determines where to place its mailbox considering factors such as distance, number of messages, etc.

The system protocol uses the NM option because this option has the lowest cost and it is the less complex to implement. Additionally, NM works properly in system where there are a small or medium number of mobile agents.

- **Messages delivering.** This aspect refers to the method used to deliver the messages from the mailboxes to the mobile agents. Two operations are available:

Push method (PS): in this method the messages are sent to the mobile agents from the mailboxes, and so in this method the locations of the mobile agents must be known every time.

Pull method (PL): In this method the mobile agents take the messages of the mailboxes, and so the mobile agents must know the locations of the mailboxes.

The system protocol uses push method, because of two reasons: first, some time it is necessary that the deliveries messages are done as soon as possible, it is real time requirements and second, the PL method produces a highest delivery messages costs than the PS method.

- **Synchronization for migration and delivery messages.** This aspect refers to the reliability of the messages delivery, avoiding the loss messages. The key of this aspect is the policy used for synchronizing the migration of the mail boxes and agents. There are three policies:

No synchronization actions (NS).

Synchronization between host's message forwarding and mailbox's migration (SMH).

Synchronization between mailbox's messages forwarding and agent's migration (SMA).

Full synchronization (FS) that covers the two previous synchronization actions

The system uses SMA police; when a mobile agent connects with a no mobile agent, it sends register message and waits for an acknowledge message, Then the messages can be forwarded. For each message it sends an acknowledge message indicating that the message has been received. When a mobile agent migrates it sends a deregister message to the agent server.

3. Middleware

In the ubiquitous computing paradigm, the middleware is the interface between ubiquitous network and the ubiquitous programs, providing a transparent environment and uniform managements of the ubiquitous communications channels and context information for developing of programs. In the explained system, the main goal of the middleware is the proper integration of the mobile onboard system in the information system of the transport company, this implies to permit the spontaneous interactions between ubiquitous programs.

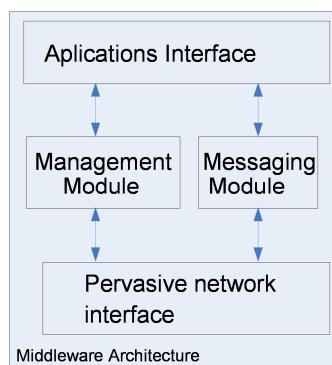


Figure 6. Middleware structure.

The middleware works in a transport company that has fleet of 300 buses, transporting about 25 millions of passengers yearly. Daily, the company executes about 3000 expeditions generating this activity about 2 millions of data transactions. To realize these transactions the middleware uses the different communication infrastructures of the ubiquitous network considering technical, strategic and economic criteria. This amount of data transactions are processed by a set of ubiquitous programs executing in different devices of the company infrastructure. The run-time middleware, which is executed in the mobile devices, decides the type of communications to use and the type of network to use depending of the location of the mobile device. Basically, the algorithm used by the middleware, for transaction from the mobile system, can be expressed of the following way:

```

If(the location of the mobile system is covered by IEEE 802.11 network) Then
    Use this infrastructure
Else
    If(the data to send belong to priority process) Then
        Use the long distance communication infrastructure
    Else
        Wait until the system is in IEEE 802.11 network coverage to send the data
    End if
End if

```

If the origin of the data transaction is not a mobile system and the data do not belong to a priority process, it is sent to the mailbox of the destination mobile system. Alternatively, if the data has priority, then the long distance communications infrastructure is used.

The context information plays an important role. Basically, the technical criteria are related to infrastructure availability and capacity, the strategic criteria are related to the policies and priorities of the task made by the ubiquitous programs and the economic criteria are related to the cost of the communications infrastructure use. The context information not only is used by the middleware, it is accessed by the ubiquitous programs too, for example data related to positioning, velocity of the vehicles, timer, etc, are frequently required to the programs.

The middleware has a hierarchical structure, figure six, formed by an interface layer with the ubiquitous programs, an interface layer with the ubiquitous network and a middleware kernel.

3.1. Ubiquitous Applications Interface

This layer provides the access to the context information. From the system point of view, a relevant context data is a context entity and each context entity has the following attributes:

- Name: it consists of a characters string that identifies to the entity in all the system, so if a ubiquitous program need to access to an entity, the program must to know the entity name.
- Data type: it defines the set of value that the entity can adopt.

- Entity location: it specifies the location in the system where the value of the entity is stored.

All the entities of the system are organized forming a hierarchical structure, specifically an entities tree. Each entity is located in a leaf node of the tree and each entity is addressed by the path to the leaf node associated. Conceptually, all the context information of the transport information system is structured in different functional contexts, these are:

- The system context formed by all the entities needed to manage the communications and data flows. The process related to middleware access to this context.
- The client context; formed by all the entities needed by the process of the traveller information subsystem.
- The production context; formed by all the entities needed by the tasks related with the activities of production and payment.
- The control context; formed by all the entities required for planning and operations controlling.
- The maintenance context; formed by all the entities required by the maintenance and administration of devices, data and programs of the mobile system.

3.2. Actions Triggered by Context

The information system supported by the ubiquitous computing context explained in this chapter can be classified as a transactional information system. Where a large amount of data is exchanged with the mobile system installed on vehicles. The control of the data flow is made using rules that are executed on automatic and unsupervised way in the mobile system. These rules are defined with the context data, specifying actions to execute and the values of the context data that triggers the actions. The ubiquitous programs are defined by a set of context sensitive rules that triggers data flow on automatic way between programs throw generic communications channels that are transparent to the network infrastructure used. In the system architecture, the middleware run-time installed on mobile systems provides the generic communications and setting the communications scheme, direct communications channels versus communication using mailboxes. A rule is structured in two components, the specification of the condition which produces the shot of an action and the action to execute. The specification of the condition is based on context data that processed using logical or relational operations and compared with constant values or the value of another context entity. The actions can be of two types: programs execution or the data transfer with using the generic communications channel provided by the middleware. Each ubiquitous application is structured of the following way:

- Context data declaration used by the program.
- Set of event to control by rules.
- Set of configuration values such as names of the context entities, sampling period of the context data, etc.
- Middleware run-time.

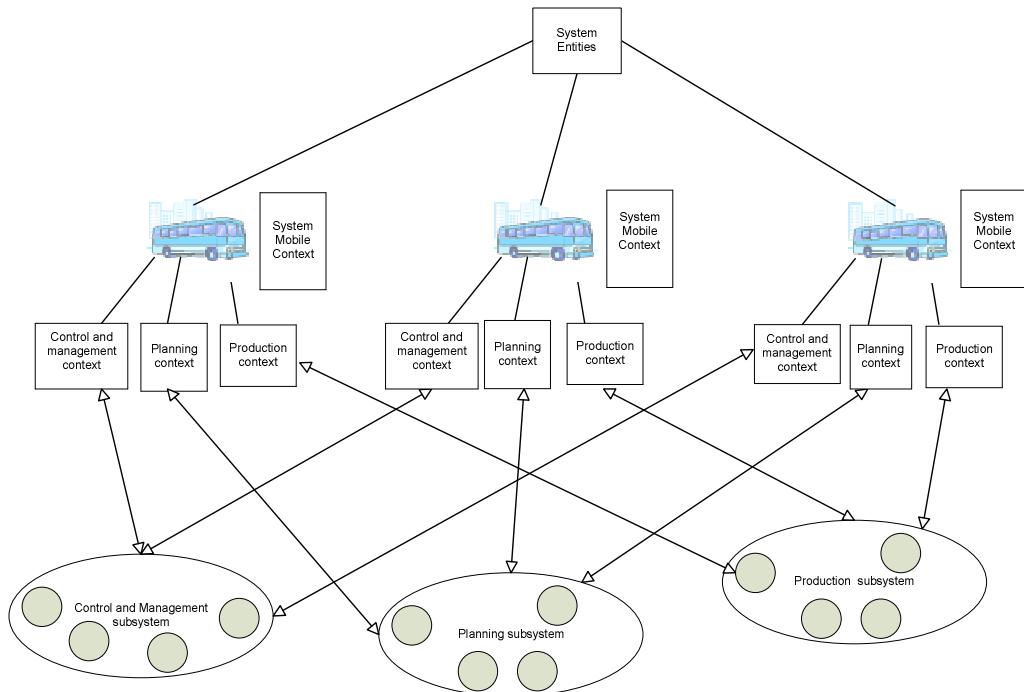


Figure 7. Hierarchical structure of the context information.

3.3. Ubiquitous Network Interface

The layer of ubiquitous network interface is responsible of providing the generic communications channels. These channels permit the communications regardless of the network infrastructure used. This layer plays a role of adaptor to the communication infrastructure, achieving the access to ubiquitous network and sending and receiving the data using the primitives of this layer. There are two types of primitives in this layer; the first type permits the communication between ubiquitous programs and the second to type permits the communications with the ubiquitous network. To send data, two basic specifications must be provided to this layer: the address of the devices where the ubiquitous programs are executing and the infrastructure to use. Also, this layer is responsible of the encapsulation of the data using the specific format of the communications infrastructure used. Similarly this level receives the data packet of the ubiquitous network, extracts the data field from the packet format and sends to the receiver application.

4. Ubiquitous Programs

The ubiquitous software is characterized by its high capacity of integration of the physical environment. As a consequence of this integration, the ubiquitous software is able to work in an autonomous and spontaneous way at different environments. For this reason, the ubiquitous software development must fulfil the following principles [4].

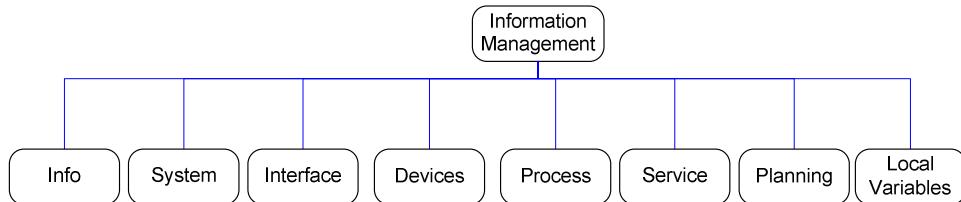


Figure 8. Context data space.

- **Border principle:** The environments must be defined by information and not physical or technical limitations. Thus, the environment detection depends on information context borders instead of physical borders. These borders must not limit the interoperability between systems.
- **Volatility principle:** In ubiquitous systems we must assume that the number of users, devices and applications in a ubiquitous environment is unforeseeable. Therefore, it is necessary to specify a set of invariant functional principles which manages system execution.

The goal of the system is to make on automatic way all the tasks belong to the subsystem of production, planning, traveller information and maintenance. All of these tasks are achieved by the ubiquitous programs. Like was previously explained, the data used by the middleware and the ubiquitous programs are structured in a context data space with a structure of a tree formed by eight groups of context data subspaces.

Each ubiquitous program can use one or more groups of context data subspaces. For example:

- **Transactions of the vehicle data production:** When a production service is made in a vehicle, then the onboard ubiquitous programs of production subsystem prepares the production data, delivering to the run-time middleware and, finally, the data are sent to control center when the vehicle enters coverage of the IEEE 802.11 network. The operation of sending of production data to control center has an expired time, if this expired time is reached, then the run-time send the data using the long distance communications infrastructure. To realize these actions the following context data subspaces are used:

From the context system data subspace the basic Identifiers are obtained, for example the vehicle identifier.

From the context of the devices data the Identifiers of the hardware devices of the onboard payment system are obtained, for example: driver console and card readers).

From the context of the planning data the identifiers of the production data are obtained, for example: identifier of the service, driver and expeditions of the service.

- **Production control:** when a exception occurs in the vehicle, the subsystem of production control builds a exception data packet, determining the kind of exception and depending of the relevance of the exception, the data packet can be sent immediately to the control center or it can be sent to the mailbox of the control center. In this context, the exception is a mechanism used to signal a relevant event

that can affect to the normal operation of the vehicles, for example: risk of delayed expedition, delay in the time of pass in an expedition control point, vehicle full of passengers, onboard device fault, etc. To realize these actions the following data subspaces of context are used:

From the data subspace of system context: the time, location and the identifier of the vehicle where the exception has occurred are obtained.

In case in technical alarm, from the data subspace of devices context the Identifiers of the hardware device that has faulted.

From the data subspace of the planning context; the identifiers of production data are obtained, for example: identifier of the service and expeditions of the service where the exception has occurred.

- **Data and software updating:** When a change in the onboard data or programs is made, the new version of software must be sent to all the fleet mobile systems. To achieve this updating, the maintenance subsystem of the control center sends a updating packet to the mailbox of the mobile systems affected, later each time that one of them will be covered by the IEEE 802.11 local network the updating packet is sent to the mobile system from its mailbox. The subsystem of maintenance receives the updating packet from the mailbox and considering the date and time of the updating, that is a field of the packet, performs the updating task on automatic and unsupervised way. In this example of task developed by ubiquitous agents, all the data required are obtained from the data subspace of system context; specifically the required entities are the identifiers of vehicle and onboard system and the numbers of the onboard programs version.

4.1. Payment System

Payment systems play an important role in the productive activity of the passengers transport companies. The companies, using automatic payment systems, improve security and commercial speed, providing to their clients reductions in the trip cost. We describe how to apply the ubiquitous paradigm in order to improve the payment systems and present a payment system model that allows clients the use of different mobile devices such as contactless smart cards, cellular phones or PDAs as means of payment. The system has the necessary mechanisms to be used in a combined way in different services contexts (public transport, parking, museums, etc). Relevant aspects of the system are the flexibility to implement different payment schemes, and the scalability to incorporate new mobile devices as means of payment. These characteristics are consequences of the system architecture, based on two main generic elements: the mobile device model and the user application model.

Payment automation provides service-oriented enterprises with many benefits: simplifies the interaction of users with payment systems, allowing to the enterprise the customization the types of payment media and adapting them for the different client characteristics, improves fare system by optimizing information flows, increases security and fraud control, and decreases exploitation costs, by using resources provided by the clients themselves. Also, the veracity of the mobile devices permits a greater services offer flexibility and their combination, for example: transport with tickets for cultural and leisure activities, parking.

The development of this automation has some requirements that we can summarize as: availability of versatile payment devices, supporting a variety of modes and customization parameters, development of processing and communication mechanisms permitting us to reduce interaction to the minimum, and development of mechanisms guaranteeing the security of data and transactions. Nowadays, contactless smart cards fulfil rather well those requirements. They are being incorporated in the latest years into enterprises payment systems. The next qualitative step is being carried out using mobile communication and data support systems, as they make possible the progress towards more evolved payment systems. We believe that a payment system based on those devices can offer the following properties.

- Scalability; it is the ability to incorporate new payment functionalities for the means of payment, communication systems, and clients.
- Security; it is the ability to detect wrong transactions, both fraudulent and caused by technical or accidental failures.
- Maintenance easiness; it is the ability to detect and respond to operation failures of its elements. Also, it is the ability of its physical and logical elements to be easily updated.
- Robustness; it is the ability to work in adverse physical conditions, both due to environmental reasons and to a massive and continued use.
- Speed; ability to carry out every transaction that is required in order to provide services access to the users at speeds that do not interfere with productive organization activity.
- Plain interactivity; it is the ability to permit the users to easily employ the means of access to the services.

In the system, a client application is defined as a set of structured data and a set of commands for its manipulation. Basically, these commands permit us to perform: communications with other remote applications, simple arithmetic and logical operations and storage operations. The sets of structured data are composed of data units characterized by two properties: identification and type of access permitted. By means of these data unit we represent the different and relevant aspect of the client applications such as: client identifications, device identifications, general parameters of the payment system, specific parameters of the payment system oriented to the clients and payments accounting. A context application is associated to a specific service such as transport, parking, etc. A client application that permits the user to access to a specific service is associated to the corresponding context. In a multi-service payment system we can find different contexts. In each context, the associated application permits the client to pay the service in different ways depending on the parameters stored in the application data. The extreme case will be a payment based on personal characteristics of the client. The client application commands are executed following a common scheme:

- 1 Context identification: The local infrastructure sends context identification packets.
- 2 Context information request: When a client application identifies a context, it sends an information context request, producing an information interchange between the client application and the local infrastructure.

- 3 Data transformation: A subset of the client application data is modified and generated by the client application or by an application running on the local infrastructure.
- 4 Transaction confirmation: The client application confirms to the infrastructure the complete execution of the client application commands.

The data transformation step must be emphasize, because it can run both on the client payment device, and on a specific station of the local infrastructure. Thus, the client application commands can be executed in a distributed way. This capability is motivated by several reasons: First, with this execution scheme the client applications are independent of the client devices capabilities. Second, in order to improve the security, we can ban the execution of certain operations on client devices. Third it allows us to support the dynamic nature of certain aspects of the payment systems.

A virtual payment device has been defined in order to establish a common client applications execution support that isolates the specific characteristics of real client devices. It integrates the variety of technologies and functionalities used nowadays as means of payment such as contactless smart cards, cellular telephones, PDAs, etc. To define this model of generic payment device we have taken into account aspects and working principles of devices and products of the market, specially JavaCard and Symbian operating system. The virtual payment device is a virtual machine composed of four layers:

- Physical device layer: It is composed of the physical elements of the device. Its minimum requirements are: about 16 Kb of storage capacity and radiofrequency communications capability.
- Logical device layer: It consists of the software support provided by device's manufacturers. At this layer, some common services are provided: communications, cryptography, file system, etc.
- Common services layer: It is the first level supplied by our system. Basically it provides the client applications execution support and the specific mechanisms for security.
- Client application layer: It is composed of the different applications installed on the payment device.

In general, the security is a critical aspect of any payment system. Thus, there are international specifications on the payment systems; especially we have considered the ICAO recommendations about passwords security in contactless smart cards [5]. Considering that in our system the means of payment are supported by non-proprietary technology devices such as PDAs and cellular telephones, security is a key issue. We have distinguished three main security matters. First, authentication of client application; a client application cannot be replicated, for this reason any client application has to be associated to a unique user and a unique payment device. In our system, every client application verifies that it is running on the device where it was installed. Thus, a basic requirement of the payment device consists of providing a service to get the unique identification key associated to the device at run-time. Second, access control to the applications; any access to data structures and application commands have to be done by authorized users. If an unauthorized access is detected, the transaction has to be rejected. And finally, transactions control; unauthorized client

applications must be detected and rejected by the infrastructure. Another transactions control aspect consists of the detection of incomplete transactions. When it happens, the system has to complete the transaction by itself. Finally, the system has to permit auditory processes that allow the company to verify the fulfilment of the security principles. To achieve this requirement it is necessary the traceability of clients operations.

How the system has been applied to the public transport context is illustrated following. Specifically, how to implement two kind of electronic tickets are explained. Ticket of fixed travels: It consists of finite number of trips with departure point and arrival point that cannot vary. Every time the client uses this ticket, one trip is decremented. Transport electronic money: It consists of an amount of money that can be spent to pay an arbitrary trip. The use of the electronic tickets follows these steps:

1. Context detection: The application is waiting for the context detection state. Before it is reached, the application can request some information. For example to use the transport electronic money, the application must request the destination point.
2. Authentication of the application: This step consists of sending a context authentication packed in an element of the infrastructure. This permits the client application to leave the waiting state.
3. Context information request: In this step the client application requests some context information in case it is needed to run any command.
4. Validation: The client application sends a data packet and commands packet to an element of the infrastructure to validate the application.
5. Confirmation: The validation element of the infrastructure sends to the client application the updated data.
6. Registered: The client application registers the updated data in its data structures stored in the payment device memory.
7. Transaction notification: The client application sends a packet of complete transaction notification to the infrastructure.

The data used by these client application examples for public transport are the following. Client personal data; this set of data consists of at least a unique identification key associated to the clients and a unique identification key associated to the application. Electronic ticket configuration data; they specify the type of electronic ticket and other specific parameters needed by the electronic ticket. And finally, history of movements registering all use of the ticket.

5. Conclusions

This chapter has explained how technological advances can help a road public transport company to improve its service and undertake new challenges. Specifically, given the current state of information technologies, the chapter has described how the pervasive computation model can be applied to build information systems with functionalities not proposed up to the moment in the context of the public transport of passengers by road. Basically, the system described in this chapter permits us to appropriately integrate mobile on-board information systems into the enterprise network vehicles and to offer high quality services to the public

transport user, playing a main role the middleware of the system. Besides, this system will allow the developing of flexible and scalable ubiquitous computing services at an attractive cost.

Those statements explained in the chapter have been supported by the results obtained by the system working at the *Global Salcaí-Utinsa, S.A.* Company. This enterprise has a fleet of 300 buses, transports 25 million passengers every year and carries out more than 4000 expeditions every day.

Acknowledgement

We want to express our gratitude to all the organizations which have supported us in the execution of the projects that have given rise to the ideas that we have expounded in this chapter: the transport company *Global Salcaí-Utinsa, S.A.*, the *Consejería de Economía y Hacienda del Gobierno Autónomo de Canarias* and the *Consejería de Turismo y Transportes del Gobierno Autónomo de Canarias*.

References

- [1] Weiser, M. (1991). The computer for the 21st century. *Scientific America*. Vol. 265, nº.3, 94-104.
- [2] Saha, D. & Mukherjee , A. (2003). *Pervasive computing: A paradigm for the 21st century*. Vol. 36, nº3, 25-31.
- [3] Cao, J., Feng, X., Lu, J. & Das, S. K. (2002). *Mailbox-based scheme for mobile agent communications*. Vol 35, nº 9, 54-60.
- [4] Kindberg, T. & Fox,, A. (2002). System Software for Ubiquitous Computing, IEEE Pervasive Computing. *Mobile and Ubiquitous Systems*, Vol. 1, nº 1, IEEE Computer Society, 70-78.
- [5] ICAO Recommendations (2004). Report of the 33rd meeting of ISO/IEC JTC1/SC17/WG8.

Chapter 4

INSTANT MESSAGING IN PRIMARY SCHOOLS

Damian Maher
University of Technology, Sydney

Abstract

The use of instant messaging (IM) in primary schools is a recent phenomenon having been around for less than 10 years in most schools with internet access. There is an expectation by Educational authorities, parents, teachers and students that interactive technologies such as IM be included as part of learning experiences. To date there has been very little research examining the use of IM with primary school students.

The use of IM has the potential to change the nature of education by expanding the range of participants with whom students can interact, both while at school and in their homes. Students now have access to experts online and other community members which vastly increases their access to different ideas and opinions. In addition, students can interact with other students who are geographically distant which enables increased cultural awareness. Students are also able to interact with each other, family members and their teachers while at home, which is dissolving the boundaries between school and home.

Access to other participants via IM has brought with it new challenges. In particular, the safety of students online has been a main focus of schools, parents, educational authorities and Governments and is examined in this chapter.

Introduction

Until quite recently, many primary schools tended to be very self contained. Occasionally students would set out on an excursion or a visitor would come to the school to speak to them. Now with the advent of Instant Messaging (IM), students have access to a vastly increased number of participants with whom they can interact with both at school and at home. Education using IM is very different to that of pre-IM use. This use of IM in primary schools provides access to an expanded audience, which has implications for the way primary education is constructed, implemented and assessed.

The use of Interactive technologies in schools such as IM is being encouraged and even mandated by both Governments and Education Authorities in many developed and developing countries. The Melbourne Declaration for example, sets out national goals for

schools in Australia and here it is stated: "Rapid and continuing advances in information and communication technologies (ICT) are changing the ways people, share use, develop and process information and technology. In this digital age, young people need to be highly skilled in the use of ICT." (MCEETYA, 2008, p.4). The New South Wales Department of Education and Training (DET) states the following in its Online Communication Services: Acceptable Usage for School Students Policy: "Online communication links students to provide a collaborative learning environment and is intended to assist with learning outcomes. Today's students are exposed to online communication tools and the internet in their community." (DET, n.d.).

Much of the data that is used in this chapter were collected as part of a research study whose focus was in examining the use of communication technologies in an Australian primary school classroom. A number of projects were implemented whereby students interacted with other students from different schools. The students also interacted with each other, other family members and friends and myself in the evenings. The results of the projects clearly demonstrated that IM has an important place in education.

Before proceeding further it is important to provide a definition of IM. Originally IM allowed real-time (synchronous) communication between two or more people based on typed text. The text was conveyed via devices connected over a network such as the Internet. Today IM also provides for text as well as encompassing features such as sound and video. Popular programs that use IM features include Windows live Messenger, Skype, Google talk and iChat to name only a few. IM is also embedded in many social networking sites such as Myspace and Facebook. The terms Chat rooms and Instant Messaging are used interchangeably by some authors. In this article IM does encompass chat rooms, but not with random strangers as is often the case with some public chat rooms.

Here in this chapter, some of the advantages of using IM in the primary school setting are reported on. While IM offers many advantages, there are also challenges in using it, particularly with young people. Safety is an important consideration in facilitating online access. In focusing on this issue, contact with strangers and cyberbullying are examined in this chapter.

Using IM in the School Setting

There are a number of new learning opportunities that are possible by incorporating IM into the classroom learning environment. Links between schools can be established which provide students with the opportunity to share expertise with each other. The use of smartphones now allows students to contact each other and other schools while in the field, which adds a new dimension to excursions. The use of IM also allows young people to contact experts for educational, health or other reasons.

Facilitating Access between Schools

IM can be utilised in the classroom to facilitate access to a wide range of participants. One way that IM can be used to facilitate access between schools is to link students to each other to support their academic and social needs.

One example of this linking is a project carried out for the research study where students between a primary school and the local high school interacted with each other over a two month period using IM. The purpose of this aspect of the research was to explore how the use of IM could help facilitate the transition to high school for the primary school students. Students interacted once a week for 40 minutes each week and then the primary school students visited the high school in the eighth week. The educational focus of the project was based on Drama. Students also had time each week to discuss issues related to high school during their online interactions. The students were asked to comment on their interactions at the end of the project and the majority of students' responses were very favourable as is indicated in this student's response:

I chatted about high school is like, what happens if you get into trouble. I learnt that high school is really different to primary.

The teachers involved in the project also found that using IM extended the possibilities for student learning. One example of this is that without using IM, the students would have only had contact with the students during the school visit. This would have meant the information they would have received about the school would have been much less. Providing the opportunity for students to interact online each week gave them some freedom to ask the questions they were interested in which were then followed up in later online sessions.

Another possibility for using IM is to link students in different parts of the country, or in different countries. This allows students to share cultural experiences with each other and learn more about what it means to live in other parts of the world. In one example from the research study, students in a city school were linked with students in a rural school and were invited to share ideas about their homes and hobbies. The city students were working on a unit on cultural identity. The aim of the unit was for students to understand how identity, which is influenced by events, people, family, peers and the media is shaped. Using IM allowed students to understand a very different way of Australian living of which they had limited experience.

Factors in ensuring successful projects using IM:

Ensure that you carefully plan with other teachers the scope of the project.

Interact often with other teachers to ensure that the project is on track.

Set allocated times for students to interact with each other.

Take into account differences in time zones and differences in the timing of school holidays.

Treat the project as you would any other project where there are clear curriculum outcomes.

Other possibilities for projects that include the use of IM are where students live in the same geographical region and there is a common concern or problem that students from different schools are working on together. Examples of this are where a bush regeneration project or a river regeneration project is taking place. Many primary schools do participate in projects similar to these that have tangible benefits for the community. Providing access for students during such projects via IM can help facilitate for improved outcomes and also

allows students to liaise more easily with different Governmental authorities throughout the project.

Contacting Experts

Another way that IM can facilitate learning in the classroom is that it allows students to chat directly with experts on a whole range of different topics. One of the topics about which students can chat online with experts includes discussion about animals. Students can chat to an expert and ask them about their pets, for example at the American Society for the Prevention of Cruelty to Animals (ASPCA), an online community website. Students can interact with experts about different animals that could support their work on a project focusing on pets. Students can also interact with students from different schools on various topics where the students themselves provide the expertise. One example here is the British Broadcasting Corporation's (BBC's) web site for students where there is a variety of different forums on issues such as sport that are available to support students' school work.

As well as interacting with experts to support class-based projects another important area that students can chat online is about what they may feel too embarrassed to talk about in a face-to-face situation or by telephone, ie mental and/or physical health. This was a factor identified by Livingston and Bober (2003) who found that "especially for girls, seeking advice online is less embarrassing as it can be done anonymously" (p. 19). Part of Health Education in schools provides students with information on people they can contact should they need help and these forums offer much needed support for young people who can sometimes feel alone and isolated.

Parents, relatives, community groups and local businesses are also able to provide expertise online. An example of this could be where a project is being completed in class on the different types of professions in the local community and what is involved with these professions. Students could interact with different individuals and ask them about their job and what it entails. Having synchronous chat means that students are able to interact more freely with participants compared to using email.

Ubiquitous Learning

The use of IM in schools is not limited to classroom use. As indicated earlier in the chapter, IM is evolving. One of the more recent innovations is the use of IM coupled with a smart phone. These devices allow for mobile learning to take place in new and innovative ways. These technologies allow students to learn in a variety of situations, which transcend the school setting. The type of learning coupled with the use of these wireless technologies is known as ubiquitous learning (Rogers, et al., 2004; Syvänen, Beale, Sharples, Ahonen, & Lonsdale, 2005).

Researchers have begun to explore how smart phones equipped with IM can facilitate learning in primary schools. One example of this type of innovative learning is the trial undertaken in Sweden where ubiquitous learning featured in the Amulets project (Kurti, Spikol and Milrad, 2008). In this trial 29 grade five students were divided into two groups. One group went to a museum and one group went to a city square. In this trial students collaborated using instant text messaging with the smartphones. This allowed communication between the smartphones in the city square and the stationary computers in the museum.

As a result of this trial, it was found that there are many benefits of ubiquitous learning for students which include:

"learn to explore a topic in authentic settings
collaborate in order to construct common knowledge
reason and to argument in order to come to the solution of a problem
reflect upon things and to support abstract thinking" (Kurti et al, p. 183)

Other advantages in adopting ubiquitous learning practices include the possibility of inter-group collaboration where students are mobile and can interact with a range of participants, rather than just the person on a computer next to them (Danesh, Inkpen, Lau, Shu, & Booth, 2001). Ubiquitous learning also support social interactivity, enables individualised scaffolding, and facilitate cognition distributed among people, tools, and contexts (Klopfer, Squire, & Jenkins, 2003).

Using IM to Facilitate Links to the Home

Communication technologies like IM are embedded in young people's lives. They have grown up with this technology and many are comfortable using it. Tools such as IM facilitate communication between young people, their friends and family as well as other participants. Young people's identities are constructed where face-to-face and online interactions are intertwined. This use of this technology allows a seamless off and online experience.

A study by Pew Internet and American Life project demonstrates just how embedded the use of IM is in American teenagers' lives. Of the teenagers who go online, 75% use instant messaging (Lenhart, Hitlin, & Madden 2005). Typically, the cell phone is the tool of choice when using IM. These teenagers are not only using IM to chat according to this report, they are also using IM to share links, photos, music and video. Not only do these young people chat with each other, about one third of them use IM to keep in contact with their parents.

This degree of access is repeated in many countries around the world. In Australia for example, it was found in a study published in 2007 by the Australian Communications and Media Authority (ACMA) that nine out of 10 Australian families have a networked computer and 95% of households with children have a mobile phone (ACMA, 2007).

As young people use the Internet and in particular IM as a way of interacting, it is important that teachers have an idea of what they are interacting about and who they are interacting with socially. This helps to inform teachers of the interests and skills of students, which is very important in an environment where links between the home and school are seen as educationally beneficial, as is increasingly the case in many classrooms. Interacting online with students also opens a window into young adolescents' social networks, which can include family as well as friends.

As part of the research study students invited me to join their online group where we interacted in the evening (with parents' permission and knowledge) using *Messenger* about matters that both related directly to school life but also about matters that could potentially contribute to school life.

The use of IM at home facilitated learning projects that then carried over into the classroom. One of the purposes for which students logged into Messenger in the evening was to discuss contents of a class web page. As a result of the discussion students were able to

contribute towards the creation of the class page with content that they felt was important to them during class time. The ability to interact in groups like this means that students and teachers have the potential to collaborate online as a group on a whole host of different subjects which can then be incorporated into the learning occurring in the classroom.

By chatting socially with the students I was able to get to know more about their interests and hobbies. I was able to learn more about what type of music the students enjoyed and what sites they visited as the following extract indicates:

<Jack> Korn is on [channel] v again
<JG> [Jack] R U WATCHIN KORN!!!!

Students often discussed music and bands when they interacted online with me. It was through such interactions I learnt that bands like *Korn* were popular and that one of the students' favourite web site was called *Channel V*, which contained music related material. Knowing about the music interests of the students means that teachers could use songs that students know as a starting point when teaching music.

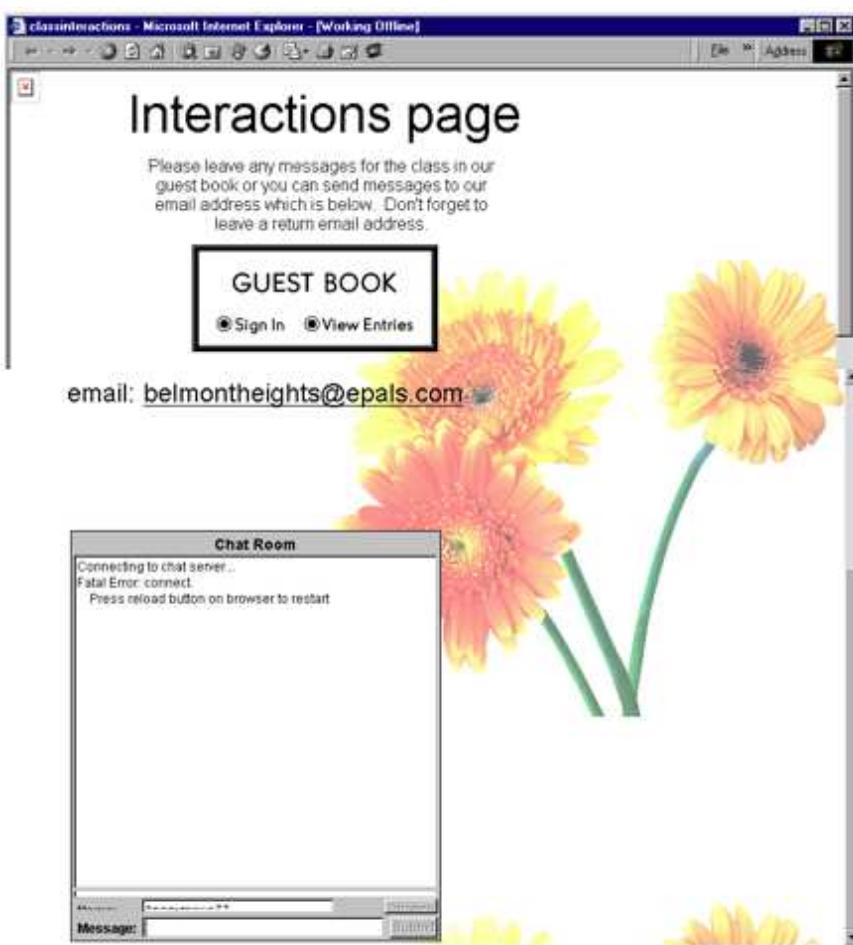
Having knowledge of students' interest outside of the classroom can assist in facilitating learning in the classroom. This knowledge is important from a constructivist viewpoint of learning where the premise is that "human learning is *constructed*, that learners build new knowledge upon the foundation of previous learning" (Hoover, 1996). By knowing what interests students, the teacher is more easily able to facilitate successful learning that is authentic for the students. Authentic learning is a pedagogical approach that allows students to explore, discuss, and meaningfully construct concepts and relationships in contexts that involve real-world problems and projects that are relevant to the learner (Donovan, Bransford, & Pellegrino, 1999).

Through the use of IM the links between school and home can be much stronger allowing this type of learning to occur. This link between school and home via the use of IM is being more widely recognised as an important aspect of contemporary education (Maher & Lee, forthcoming).

There are a number of related issues that need to be considered if teachers are to interact online with their students in an informal manner. First, there is the time required to interact online in the evening. Many primary school teachers (which mirror many occupations) feel constrained by time and so if there is an expectation that they might interact out of school hours then this may be resisted by the teachers. Secondly, there are the skills needed by teachers in being able to interact online. There is a large range of skills required by teachers to be able to interact online using IM. As well as having an understanding of how to use IM, teachers also need to learn how to implement learning activities in the classroom that incorporate the use of IM. It would be fair to say that many teachers do not get training opportunities to learn the required skills which are constantly changing as technology changes. Generally if teachers have expertise in this area it is because they have up-dated the skills in their own time. Thirdly, there is the aspect of safety and approval. Parents may feel that they do not want their child interacting online in the evening and may object to their child's teacher interacting.

Facilitating access between schools, experts and the home using IM means that the tools required need to be easily found and accessed by all participants. One of the outcomes of the

research project was that simply having an IM link did not provide the type of access required to enable purposeful online interactions to occur. A suggestion to ensure this can happen is for classrooms to construct their own web site and host all the interactive tools on one page. This was what was done in the research study referred to previously. Here is a screen shot of the interactions page:



As can be seen, the chat room is part of the interaction page. The link to this page was clearly signposted on the main page of the class web site. This meant that any time the class wanted to invite someone to contact us we gave them the link to our class web site with instructions to click on the interaction page. This made it very simple for the students and parents to use. It was found that having this facility allowed students to interact socially with each other although they did then use *Messenger* as they had existing accounts and this was their preferred medium.

IM and Literacy

The literacy practices of students using IM is very different compared to traditional paper and pen practices. Writing is becoming more visually based. The use of emotions which allow

participants to express their emotions quickly and easily is a common literacy feature employed by many young people when using IM. The use of colour as well as different size and types of fonts also allow students to use different literacy features as a way of helping to construct their identity online.

Another change to literacy practices online the use of abbreviations such as U for you. These type of abbreviations were constant throughout the interactions as part of the research project where students used IM both at school and while interacting informally in their homes. Other examples of abbreviations students used in the study include:

G2G- [I've got to go].

LOL- [laugh out loud, or lots of laughs].

Cul8er- [see you later].

GR8-[great]

This type of literacy using IM comes about partly as a result where speed is of more importance than accuracy. This type of spelling is as Merchant (2001) suggests, phonetic in nature and illustrates how online chat combines both elements of written and spoken conventions. In using such as abbreviations and emoticons the students were using "linguistic features to manipulate the written tone, voice, word choice, subject matter, and structure of messages" (Lewis & Fabos, 2006, p. 482).

In several interactions I had with students from the research project they assumed the role of the teacher and I that of the student. This is evident in the following interaction where Paul teaches me how to access the different colors where I am 'The D':

The D: What are the symbols?

Paul: EMOTICONS



The D:

Paul: YEA

Paul: **U CAN CHANG NAME AS WELL**

The D: I see

This role reversal demonstrates the assertion of Leu Jr., Kinzer, Coiro, and Cammack (2004), which is that many young people have a greater understanding of technology use than many adults. The model of the teacher holding all the knowledge and being in charge of imparting it to students is changing as a result of the rapid rate at which technologies like IM are being adopted in both schools and homes. It is important therefore that students' knowledge be something that is allowed to shape the school curriculum.

The use of this type of language shown here demonstrates that literacy is changing (Lankshear Snyder, & Green, 2000; Livingstone, 2003). While this has always been the case, the introduction of the Internet has changed the literacy landscape in a very short period of time. Literacy practices amongst young people have gone from being primarily paper-based in the written form, to one of both paper-based and electronic form. Writing has also gone from being a static process to a more interactive one online. Students have new literacy requirements and opportunities, in particular establishing their identity online using new

literacy devices such as emoticons. This is particularly important in online spaces where many participants interact at the one time. Young people need a way of establishing their identity due to the lack of visual resources they normally employ in face-to-face situations such as the clothing or jewellery they wear. Online identity markers serve several purposes; to mark young people as slightly different from others and to indicate to other participants their interests and personality so that they are one of the group.

The use of emoticons and computer-based literacy is certainly not a new phenomenon. Wilkins (1991) noted the use of emoticons and textual changes in a study examining group conversations on a conferencing network at a tertiary level. What is different is that these practices are becoming more commonplace among young people. It is likely that as students are encouraged to go online they will adopt more of the conventions of visual practices, which are constantly changing and becoming more accessible. As suggested earlier, IM is changing and now allows participants to see and hear each other while interacting online. Online literacy practices will continue to change as the new technologies are adapted in the classroom.

It is important that the teaching of literacy practices that includes electronic aspects be included as part of the curriculum in primary schools and that new literacy and literate behaviours be adopted (Labbø, 2006). In order for this to occur, there needs to be an understanding of some of the literacy practices of young people.

Safety Issues

Much has been written about children and Internet safety. There are many web sites advising teachers, parents and students how to interact safely when using IM. Two examples are Safekids.com and Connectsafely.org. As well, much has been published in the popular press on the perils of online interactions for students. There was a much publicised case of a 12-year old British school girl who ran away with a U.S. Marine she met over the Internet, which was reported in the television, radio, newspapers and online (The Scotsman). These types of reports focus on the shocking, presenting singular events as typical online interactions, and are therefore misleading. Nonetheless, there is limited rigorous research on safety issues associated with IM and young children. In this section in considering safety issues, stranger danger and cyberbullying are examined

Stranger Danger

The use of IM in schools can encourage young people to use this for interacting socially in their homes. This can lead to the possibility of contact with strangers. There have been a number of studies conducted which focus on young people's contact with strangers online. These studies have mixed findings.

A number of large-scale studies indicate that students are often harassed, or they themselves engage in dangerous behaviour while interacting online. A large scale study conducted in 2003, (which followed on from and included a UK study by O'Connell, Sange, & Barrow, 2002) examined online use by stage two students (eight to 11 years old), (O'Connell, Price, & Barrow, 2004). It looked at their use of the World Wide Web, chat rooms, email, instant messaging, peer to peer applications and mobile phones through data

gathered by questionnaire from 1661 students. Almost one-fifth of the students reported being harassed online while in chat rooms. Students also reported that they engaged in dangerous behaviour online. Some of the students contacted participants online in a chat room then went on to meet this person in a face-to-face meeting. At this meeting a friend, rather than a parent, usually accompanied the student. A number of students reported giving out their telephone numbers and addresses online while using email.

In contrast to these findings, another study conducted on online behaviour found that there was minimal harassment by strangers while young people interact online (Finkelhor, Mitchell & Wolak, 2000). In this study, it was found that approximately 6% of 10 to 17 year olds reported being harassed or threatened by strangers while online.

Whilst there are concerns about contact with strangers online for young people, according to a report published by the Australian Communications and Media Authority (ACMA, 2009) “Those of primary school age (8 to 12 years) are often content with immediate friendship circles and to not actively seek or desire contact with people they do not know” (p. 8). This implies students in this age bracket are less likely to come into contact with strangers.

There are some basic rules that can be given to young people to help ensure their safety online in relation to strangers which are:

- Never give out personal information, such as your name, address and telephone number.
- Do not communicate with strangers online.
- Report abusive or offensive behaviour to parents and teachers.
- Protect your user name and password. Do not share this information.

Cyberbullying

One of the major concerns with young people’s use of interactive technologies such as IM that has been a focus by both the media and researchers is cyberbullying. Bullying is defined here as aggressive behaviour where a dominant individual or group abuses their greater power by threatening a less dominant individual (Farrington 1993; Rigby 1996). Here there is the desire by individuals or groups to exert power over others.

Cyberbullying involves the internet or other digital communication devices (Willard 2004) for bullying. The various forms of cyberbullying include flaming (making hostile statements), flooding (holding down the send key to prevent anyone else interacting), harassment (making threatening comments), cyberstalking (stalking online), denigration (putdowns), masquerade (pretending to be someone else) and exclusion (organising for someone to be excluded from interactions).

There were numerous examples of cyberbullying that occurred during the research project.

Cyberbullying also occurred in class while students interacted online during class time. One student posted a message using another student’s name, which is an example of masquerade which brought this immediate response from the student:

Helen

who wrote OKAY Helen in the name box

Helen

mary was it u

Mary

me I didn't put my name sorry Helen

The history of this episode is that there were two groups of girls in the classroom who were having difficulties working together due to personality clashes. The bullying in the classroom was also transposed to the internet. It would appear that one of the students felt that writing a statement using another person's identity would hurt that person's feelings, which is what happened. Of importance here also is that bullying occurs both at school and at home, both in a face-to-face situation and online. Students often bring practices with them to school and part of the teacher's role is to help students understand what is appropriate in different settings.

One of the features of IM that can minimise the negative effects of cyberbullying is that conversations are generally not permanent as they are with email or message boards. Permanency is one of the features that attract bullying online. This lack of permanency is also a downside for educators and adults who may wish to keep some track of the interactions of young people. Once the conversation has finished in an IM environment it is much harder to record, meaning any problems that might arise can be more difficult to follow up and resolve.

There seems to be a belief by some commentators that since cyberbullying occurs, the use of the internet should be limited for young people. In the context of this chapter which focuses on primary school students, typically from the age of five to 12, there is limited cyberbullying by this age group that has been by researchers. It appears to be older adolescents in secondary school that are victims and perpetrators of cyberbullying. For example, in a survey conducted with 1500 youths aged between 10 and 17, 89% of youth harassed were ages 13 to 17 (Wolak, Mitchell & Finkelhor, 2006). Not all cyberbullying is as extreme as often reported by the press either (Maher, 2008).

Whilst cyberbullying is not a major issue for primary age students, it does not mean there should not be attempts made to minimise it. There are a number of ways that the safety of students using IM can be ensured. Two main ways that are addressed here are through education and through participation.

Education

One of the most important ways that young people can be taught to deal with cyberbullying is through education.

In the research study from which the data draws, it was found that participants who interacted informally online tended to belong to the same face-to-face community. This has been found in other studies. For example, when Wolak, Mitchell and Finkelhor (2007) interviewed 1500 participants aged between 10 to 17 by phone, they found that "45% of known peer harassers had offline contact with targets, suggesting online incidents may have been an aspect of offline bullying" (p.57). It was possible then, that students who were prone to be bullied, or were bullies in the class were likely to interact in similar ways online. The teacher has an important role in managing and minimising bullying online as he/she is in a

strong position to know who is being bullied in the classroom. Teaching students to be able to recognise and deal with bullying is another important aspect of minimising the effects of bullying, along with management strategies both in the classroom and online. I support the view expressed by Amis on this issue.

The problem is not the invasiveness of new communications technology - that is something that most children are perfectly able to learn how to control. The problem is not even the supposed increase in bullying. The problem is the level of adult concern and intervention that encourages children to see themselves as victims rather than making them learn how to cope with the disputes they inevitably have with each other. It is this level of intervention that is denying them the agency to sort out their own problems.

(Amis, 2002).

Children therefore need to be encouraged to manage their relationships in an online environment without excessive intervention from adults (Jenkins, 1997) as the quote above suggests. The problem of bullying exists in many primary school classrooms, and many students have lessons in how to interact in a positive way. Through the learning of skills, students might be better able to deal with situations when they arise, rather than always relying on a teacher or parent, and in this way they can become empowered to recognise and deal with bullying.

There are some basic rules that can be given to young people to help ensure their safety online in relation to cyberbullying which are:

- Don't respond. If someone bullies you, remember that your reaction is usually exactly what the bully wants. It gives him or her power over you.
- Don't retaliate. Getting back at the bully turns you into one and reinforces the bully's behavior.
- Talk to a trusted adult. It's always good to involve a parent or a school counselor.
- Save the evidence. Messages can usually be captured, saved, and shown to someone who can help.
- Block the bully. If the harassment's coming in the form of instant messages, use preferences or privacy tools to block the person. If it's in chat, leave the "room."

*Adapted from material from ConnectSafely.org

Participation

Whilst students need to learn to deal with cyberbullying without excessive intervention from adults, they should not be expected to interact online in an adult free environment. When I interacted with students online in the evening as part of the research project, other family members such as brothers or sisters would sometimes log in and sometimes friends of the students also joined in on the conversation, but there was no instance of parents logging in to chat.

In the case of supporting young people while they interact online using IM, the saying 'it takes a village to raise a child' holds true. One of the main ways therefore that cyberbullying can be minimised is through participation by adults online. This issue of who is responsible

for ensuring that the use of IM is monitored needs careful consideration but this appears not to be happening in many primary school communities. Parents and teachers might assume a cooperative approach to managing online bullying. This cooperative approach means that links between school and home need to be stronger than they are presently. Ensuring the safety of students online will require greater communication between the school and home so that appropriate supervision occurs. I fully support the view of Livingstone (2001) on this matter where she states: "To establish the home-school link effectively will require a considerable investment of resources, particularly in terms of staff time. It will also require a transformation in the formal definition of appropriate use of educational resources."

In considering safety of students in using IM it is noteworthy to consider the different Internet use policies in various parts of the world. In focusing on policies it is clear that Australia, the United Kingdom and the United States have adopted a very different approach compared to countries such as Denmark, Sweden and the Netherlands. In the Scandinavian countries, the policy is to keep ports unblocked unless at a local level there is an agreement to close them. In Australia, the US and the UK the policy position is to "close ports and use filtering software unless teachers and others at the local level request a specific URL to be unblocked" (Moyle, 2009, p.3). This policy of keeping ports closed means that it is very difficult for IM to be facilitated because the sites it is linked to are often blocked. The approach adopted by the Scandinavian countries is one where rather than controlling specific technologies, "an emphasis is placed upon building social responsibility, harnessing positive social behaviours and controlling anti-social behaviours" (Moyle, p. 3) which as has been discussed earlier, through education.

Advantages of Using IM in Primary Schools

- Provides contact between schools.
- Provides contact to experts.
- Provides contact between the school and home.
- Allows for multimodal interactions to take place.
- Breaks down barriers of distance.
- Allows for groupings between schools to be easily formed for projects.

Challenges of Using IM in Primary Schools

- Can take considerable effort to set up projects using IM.
- Can increase the workload of teachers.
- There are issues of safety that need to be addressed.
- IT systems that are needed to use IM cost a great deal of money to maintain.
- Needs close communication between parents and teachers.

Conclusion

The introduction of communication technologies into primary school classrooms has fundamentally changed the way that learning is conceptualized, implemented and assessed.

The use of IM allows primary school students access to an increased number of participants outside of the classroom. Students are able to make contact with experts and with other students, who have the opportunity to offer their expertise such as in the high school project. Parents, who have a wealth of expertise can potentially participate in the education of their children through interacting in an IM environment. As new technologies are emerging, the use of IM incorporating video and audio allows students access to places such as factories and workplaces that would have once been out-of-bounds for safety or logistical purposes.

The use of IM has also opened the way to link between the school and home. This has a number of benefits for students where they can interact with each other to organize projects as well as interact with their teachers. This access can assist teachers to have a greater understanding of the students allowing the learning in the classroom to be more authentic and meaningful.

The use of IM as well as providing new opportunities provides new challenges. In particular, allowing and encouraging students to interact in an IM environment can potentially increase safety concerns. With education and participation by adults, these concerns can be minimised. Encouraging the use of IM for educational purposes also blurs the boundaries between school and home which can create uncertainty as to whom is responsible for ensuring the safety of students. Greater dialogue and participation between the school and home can help ensure a shared understanding and also a shared responsibility.

References

- ACMA, (2007). *Media and Communications in Australian families*, Retrieved February 26, 2009 from http://www.acma.gov.au/webwr/_assets/main/lib101058/media_and_society_report_2007.pdf.
- ACMA, (2009). *Click and Connect: Young Australians' Use of Online Social Media*. Melbourne, Vic. Australian Communications and Media Authority.
- Amis, D. (2002). *Kids will be kids*. Retrieved October, 22, 2003 from <http://www.netfreedom.org/news.asp?item=186>.
- Danesh, A., Inkpen, K., Lau, F., Shu, K. & Booth, K. (2001). *Geney™: Designing a collaborative activity for the Palm™ handheld computer*. Paper presented at the Conference on Human Factors in Computing Systems, Seattle, WA.
- DET (n.d.). *Internet and Email Services: Acceptable Usage for Schools*. Retrieved April 13, 2008 from https://www.det.nsw.edu.au/policies/general_man/general/accep_use/PD20020046.shtml?level=Schools&categories=Schools%7CFacilities+%26+assets%7CInternet+%26+email
- Donovan, M. S., Bransford, J. D. & Pellegrino, J. W. (Eds.). (1999). *How people learn: Bridging research and practice*. Washington, DC: National Academy Press.
- Farrington, D. P. (1993). Understanding and preventing bullying. In M. Tonny & N. Morris (Eds.), *Crime and Justice*, v.17. Chicago: University of Chicago Press.
- Hoover, A. (1996). *The practice implications of constructivism*. Retrieved July 19, 2009 from <http://www.sedl.org/pubs/sedletter/v09n03/practice.html>.
- Jenkins, T. (1997). *Children Squabble: It's a fact of life, not always a case for the helpline*. Paper presented at the conference Childhood and Friendship in a Fearful World.

- Retrieved October 22, 2003 from <http://www.generationyouthissues.fsnet.co.uk/bullying/Children%20Squabble.htm>.
- Kurti, A., Spikol, D. & Milrad, M. (2008). Bridging outdoors and indoors educational activities in schools with the support of mobile and positioning technologies. *International Journal of Mobile Learning and Organisation*, 2(2), 166-186
- Klopfer, E., Squire, K. & Jenkins, H. (2003). *Augmented reality simulations on handheld computers*. Paper presented at the American Educational Research Association Conference, Chicago, IL.
- Livingstone, S. (2001). *Online Freedom & Safety for Children*. Retrieved July, 12, 2003 from Livingstone, S. & Bober, M. (2003). *UK children go online : listening to young people's experiences* [online]. London: LSE Research Online. Retrieved July 20, 2009 from Available at: <http://eprints.lse.ac.uk/388/1/UKChildrenGoOnlineReport1.pdf>.
- Lenhart, A., Hitlin, P. & Madden, M. (2005). *Teens and technology*. Pew Internet & American Life Project. Washington, D. C. Pew Research Center.
- MCEETYA (2008). *The Melbourne Declaration on national Goals for Schooling in the Twenty-First Century*. Retrieved April 10, from http://www.mceecdya.edu.au/verve/_resources/National_Declaration_on_the_Educational_Goals_for_Young_Australians.pdf.
- Maher, D. (2008). Cyberbullying: An ethnographic case study of one Australian upper primary school class. *Youth Studies Australia*, 27(4), 32-39
- Maher, D. & Lee, M. (2010). Student 'net usage in a networked school community- the challenge, In: M., Lee, M & G. Finger, G. (Eds) *Developing a Networked School Community*. Melbourne: ACER Press
- Moyle, K. (2009). Varying Approaches to Internet Safety: The Role of Filters in Schools. Consortium of School Networking (CoSN). Retrieved September, 19, from <http://www.cosn.org/Portals/7/docs/Varying%20Approaches%20to%20Internet%20Safety.pdf>.
- Rigby, K (1996). *Bullying in schools - and what to do about it*. ACER, Melbourne.
- Rogers, Y., Price, S., Fitzpatrick, G., Fleck, R., Harris, E., Smith H., Randell, C., Muller, H., O'Malley, C., Stanton, D., Thompson, M & Weal, M. (2004). Ambient wood: designing new forms of digital argumentation for learning outdoors. *Proceedings of the 2004 Conference on Interaction Design and Children: Building the Community*, ACM, Maryland 1-3 June.
- Syvänen, A., Beale, R., Sharples, M., Ahonen, M. & Lonsdale, P. (2005). Supporting pervasive learning environments: adaptability and context awareness in mobile learning. *Proceeds from the third IEEE International Workshop on Wireless and Mobile Technologies in Education*, Tokushima, Japan.
- Wilkins, J. (1991). Long distance conversation by computer. *Written Communication*, 8, 56-98.
- Willard, N. (2004). *An educator's guide to cyberbullying and cyberthreats*. Retrieved April 9, (2008). from <http://www.cyberbully.org/cyberbully/docs/cbctedicator.pdf>
- Wolak J., Mitchell K. & Finkelhor, D. (2006). *Victimization: 5 years later*. Alexandria, VA: National Center for Missing & Exploited Children.
- Wolak, J., Mitchell, K. & Finkelhor, D. (2007). Does online harassment constitute bullying? An exploration of online harassment by known peers and online-only contacts. *Journal of Adolescent Health*, 41, 51-58.

Chapter 5

SEARCH ENGINE INTERFACES

Gondy Leroy^{*}

School of Information Systems and Technology, Claremont Graduate University,
Claremont, CA, USA

Abstract

More often than not, we turn to the Internet when we need information about products to buy, places to visit, or even doctors to consult. We use a search engines to locate information and expect to find answers immediately and without effort. A search engine's interface is a critically important component in this process. This chapter reviews the user query options in general-purpose search engines and the underlying technology used to match that query to web pages. It also describes different query options provided by special-purpose search engines.

General-purpose search engines use a simple interface, a text box, and require only a few keywords to search. Most people use only 2 or 3 keywords, which is very little information, to search among billions of documents. Increasing the number of keywords increases the information available to search and improves the results. There are two approaches to increasing this information: establishing a user profile or using query expansion techniques. User profiles are predominantly used to filter results from a search. Static user profiles are built on information supplied by users about themselves. This information is then used of toward the selection a subset of results. On the whole, it is not very popular with users and too stringent for prolonged use. Dynamically built profiles, requiring no user effort, are continuously updated. However, many users do not like tracking of their behavior. In contrast to filtering results based on a profile, query expansion aims to add additional terms to the original query to make it more precise so that fewer but more precise results are found. A few extra, relevant keywords increase the available information leading to better results. Query expansion, whether it is automated or manual and interactive, generally improves the results and many search engines provide query expansion options as an effortless and dynamic augmentation to their basic search.

Special-purpose and newer search engines provide a different interface. For example, music or image search engines benefit from techniques that use sounds or images in the query. Natural language search engines allow users to type a query in their own words. In our own work, we evaluated query diagrams as an input method and found that they are easy to

* E-mail address: Gondy.leroy@cgu.edu

understand and the query itself contains much more information resulting in more precise queries.

Introduction

While the Internet is increasingly used for gaming, e-commerce, or gambling, information retrieval still is the most important activity on the Internet. Starting with the earliest search engines such as Archie, to today's Google, search engines have become our main and often sole source of information. When we look for information about products to buy, places to visit, or doctors to consult, we turn to the Internet. Even when browsing, a search engine is only one click away and any browser's toolbar readily includes a text box for searching with access to a variety of search engines. Some search engines cover all information, such as Google or Bing, while others are specialized for the type of content, such as images, or for a specific domain, such as medicine.

Over the years, our interaction with and expectation of search engines has not changed: we type keywords and expect a list of pages. It seems that every question can be asked by typing in the right keywords. However, increasingly more and different document collections are made available online, ranging from full-text journal articles and text books to medical information pamphlets and legal information. And increasingly, our important life decisions are influenced or even completely based on information we find online. Even though the task has become more difficult, search engines seem able to provide satisfactory results for every day users. The search engines provide ten or twenty pages - most users do not look further - selected among billion in response to a user query expressed by two or three words. Few people question the results or wonder about alternative or missing information.

Although searchers are satisfied, there is room for improvement. Current search technologies come up short in at least two ways. First, they are not equipped to deal with long texts. Keywords may appear several hundred times in a long document and highlighting will not help the user find the relevant information. Second, with more reliance on independent learning online, more attention is needed to ensure that results are complete and address all aspects of a question. Interactions with search engines are limited to a few keywords and a set of pages, largely selected by the search engine algorithms, in answer to that query. Although the effectiveness and efficiency of search engines is admirable, search engines need a better interface, a difficult task since it has to be a zero-learning interface that requires no effort. In the following, the search engine interface is reviewed together with the types of queries it allows and the consequences of these choices for searching and finding. Both general purpose and special purpose search engines are discussed.

Common Search Engine Interfaces

Search engines have become an essential part of our online experience. Millions of people search online for texts, music, photos, and videos. Regardless of the type of information sought, the Internet is the place to go. We answer our information needs with navigational, transactional, or informational queries [1]. While search engines provide information, their activities are not regulated and not guaranteed to lead to correct and

trustworthy information. Although there are attempts to rate the web information [2], it is the searcher's responsibility to formulate alternative queries and integrate the information found.

User Queries

User search queries can be divided into three groups. They are navigational, transactional, or informational [1]. Navigational queries are intended to lead to a particular site, e.g., the BBC News website. Transactional queries are intended to lead to further online activities, e.g., download music. Informational queries, the most common kind, are intended to lead to particular information. For example, "Who won the Tour de France in 2008?" is a simple, fact-finding informational question, while "Do men with sleep apnea tend to be overweight or have diabetes?" requires more in-depth reading to answer. A 2001 survey indicated that almost half of the queries are informational, while later work indicated that 80% of queries are informational [3]. In fact, for anyone of college age or younger, the Internet is the normal source of all information.

Most research evaluating user queries focuses on informational queries. A common finding is that users' online search behavior has not changed over the years. Users provide a 'bag of words' and expect a list of documents with those words highlighted. Searcher expertise does not lead to very different search strategies but is noticeable in small differences in the queries and query strategies. For example, it has been shown that novices tend to start out with very general, imprecise queries [4]. However, regardless of expertise and topic of the search, most people's queries contain very few keywords, usually two or three words [5-10]. With so few keywords, it is difficult for a search engine to deduce the user intention and find the matching documents. Moreover, those keywords are often vague which worsens the problem. The users' mental models of the use of Boolean terms, stop word removal, and term order in a search engine are also often incorrect [11]. This lack of precision in keywords combined with limited search engine understanding requires many query formulations [9] before a satisfying, complete answer can be found.

While any topic can be searched for, health information is especially popular. In 2003, Baker et al. [12] reported that 40% of a 60,000 household sample looked online for health information. Although it may lead to incorrect information, in many cases there are advantages to searching for health information online. Foremost is that people become more knowledgeable and empowered to ask more informed questions when seeing a caregiver, while their fear of the unknown is lessened [13]. Then, the online information often supplements the information given by caregivers who spend a very limited amount of time with patients [14] and often instruct patients to go to a specific website.

With the increase in chronic diseases, many patients need in-depth information to manage their health and even make life-altering decisions. The information found online affects decisions about health and healthcare as well as frequency of visits to healthcare providers [12, 15]. Because of the serious and costly consequences of misinformation and misunderstandings, the medical domain has paid close attention to how medical information is used on the Internet. It turns out that users searching for medical information are not more careful with medical than with general information. Early on, McCray and Tse [16] found that when their system suggested a correct alternative for a misspelled term, users accepted it in only 45% of cases. Later studies showed that with both medical and general purpose search

engines, user relevancy ratings of web pages found in response to questions were not related to the correctness of their answers to those medical questions [17]. Users may believe they have found relevant information, but it is often imprecise, incomplete, or too difficult to understand.

Query Processing

Search engines have become very efficient at retrieving documents when given only a few keywords to work with. Decades of research have seen to this. Enormous progress has been in the ability to store and index large collections, retrieve items in a very short time, refine user searches with keyword suggestions, and present relevant results. In contrast to these backend processes, the input text box has not changed at all with the exception of a few popular improvements such as spelling correction, phrases versus single words distinction, and term suggestion. Searching remains limited to forming a sequential string of words in a search text box.

Conventional Components

To retrieve documents, it is important to recognize the important words in a document. These important words can then be matched to the user keywords. Since there are thousands of documents that will contain the user keywords, matching documents need to be ranked according to their relevance relative to the user keywords. The basic approach to ranking documents that match keywords is based on term frequencies and the relation between the frequency of a term in a particular document to its frequency in a collection: $tf\backslash idf$ [18, 19]. Words that appear frequently in one but not in every document are considered better descriptive words for that document. Figure 1 shows how Term A is a relatively frequent term in all four documents. A term provides an indication of the content of a document if it appears frequently in the document. However, terms that appear frequently in all or many documents are not very good descriptive terms. In contrast, terms that are frequent in some but not all document, such as Term B in Figure 1, are much better descriptive terms of the unique content of a document. If such terms match the user keywords, the document gets a high ranking score.

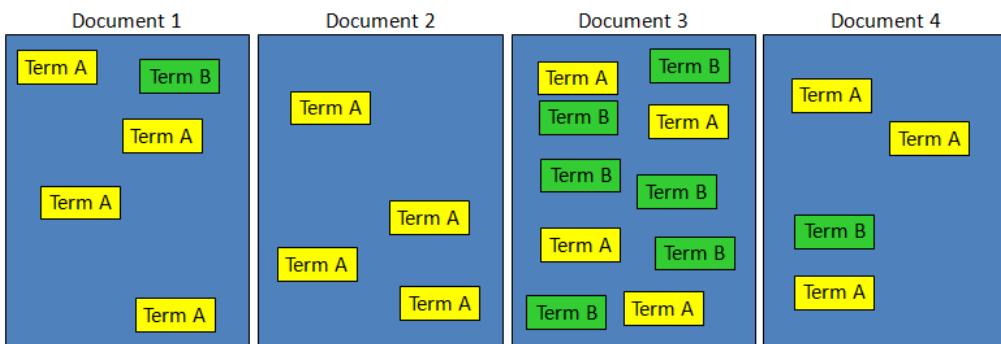


Figure 1. Term Frequencies.

With millions of documents and an average of two keywords in a user query, the simple tf\idf model is insufficient and today's commercial search engines use more sophisticated models that have become increasingly sensitive, fine-tuned, and augmented with additional information to better rank documents. For example, many optimize results by taking network or web page characteristics into account. PageRank uses the structure of the websites and the link between them as an indication of page importance [20]. It is worth noting that the solid results demonstrated by the PageRank algorithm leverage structural relationships but don't infer relatedness based on the richer semantics and predicates in the content itself. In addition to network characteristics, page characteristics can be used. Machine learning algorithms can recognize high quality pages using metrics such as word counts to improve the rankings [21]. More sophisticated ranking optimization approaches exist, e.g., using genetic programming to optimize ranking functions [22, 23]. Many details are not published today as they belong to the search engines intellectual property and are part of the business model's foundation.

Augmentations

However sophisticated a ranking function may be, any algorithm would benefit from having more information from users. Longer and more precise queries lead to better, customized results. With very few keywords to search millions of documents, much research has focused on obtaining additional information to improve search results. Personalization is one such approach. The goal of personalization is to take user characteristics and personal interests into account when searching. The personal information is often used as a filter and collected in a user profile. However, most users are reluctant to spend the time or divulge the information. Furthermore, user-constructed profiles are sometimes inferior to system-constructed profiles [24]. To avoid such problems and by using increasingly more sophisticated and fine-tuned algorithms for information gathering and tracking, user profiles can be constructed automatically. User profiles can be built automatically based on a person's search and browse history, which is assumed to represent the user interest, or by analyzing a person's short term click behavior, which is called click-based personalization. A comparison of click-based and personalized profiles that any type of personalization benefited queries with large click entropy, but hindered queries with small click entropy [25]. Click entropy is the variety in links followed when different users look at results from the same query. In addition to gathering information from one person, it is also possible to use the search history of an entire user community [26] or by the use of information that can stand in for profiles, such as a personal health record [27].

User profiles are intended to be used to filter results after a search engine has provided them in response to a query. In contrast, query expansion techniques are intended to improve the query itself by adding additional keywords, thus increasing the query's precision. To select which words to add, some type of user feedback is needed. The feedback from the user can be explicit or implicit. Explicit user feedback requires users to indicate which results are relevant. For example, users can rate results after reviewing them. With implicit feedback, such an evaluation is deduced from the user behavior. No extra effort or actions are required from the users. Additional terms are selected from the 'relevant' results for future queries. With both explicit and implicit feedback, the additional terms are selected from the available text, e.g., the text snippets shown by the search engines, and the best terms are determined with term ranking functions.

Many query expansion algorithms and term selection methods exist. In the early 90s, several different approaches were systematically compared. For example, adding all terms from relevant results was compared and considered slightly better than expansion with only the most common terms [19]. Different numbers of additional keywords were also tested, ranging from twenty terms [28, 29] selected using a tf\idf approach to lower numbers, e.g., six additional terms [30]. Many of these expansion techniques were compared at the Text REtrieval Conferences (<http://trec.nist.gov/>). For example, the use of text passages was compared to using entire documents for term selection [31, 32]. The culmination of research led to the conclusion that any form of query expansion improves results [33] and automated query expansion became even more sophisticated, e.g., automatically selecting terms by taking their term distribution in the corpus into account [34] or by using genetic algorithms for expansion and filtering of Google queries [35].

Even though beneficial, many people using a search engine do not like automated query expansion. Furthermore, it is not effective for all types of users. Query expansion was shown to be beneficial for novice users but could hinder experts [35]. As a result, manual or interactive expansion was developed [3], requiring some user input. In TREC-7's Interactive Track, it was shown that explicit relevance feedback was critical in boosting performance [36]. Results improved when users highlighted the "context" in a document which could be used to augment the query [37]. However, until recently, users seldom requested query expansion: 5% of the time or less according to a study by Jansen et al. [38]. Today, query expansion is offered by Yahoo! and Google and can be expected to be much more popular. When typing in the search box, a user can choose from suggested queries that complete the personal one. These keyword options are based on overall popularity of queries submitted by others [39, 40].

Advanced Search Engine Interfaces

Search engine interfaces can benefit from different kinds of improvements. The first type of improvement would be a better interface for the most common searches on a variety of devices. Today's text-based search engines provide only a single search box. This leads users to submit a linear string of text. These searches are increasingly conducted from mobile devices and so the interfaces will also need to be adjusted for the small screens. This adds limitations, e.g., lists of results are difficult to view on mobile devices, but also offers new opportunities not available with standard computer access, e.g., position and orientation awareness. The second type of improvements is to facilitate searching for other types of information, such as images or audio. Although today's popular search engines also provides images, sound and even video, their retrieval mechanism are largely based on the text associated with this media, such as file names or the surrounding text. The image pixels or sound waves are not commonly used in the search process. Last but not least, improvements in interfaces are needed for users with different abilities. Different cognitive abilities, e.g., limited memory in older users, and different physical abilities, e.g., lacking fine motor skills or limited eyesight, need to be taken into account.

Searching for Text

Two types of improvements for text searches are discussed below. The first does not change the look and feel of the common search engine interface, but allows a different type of query. These are the natural language search engines. The second type changes the search engine interface while retaining a more structured input format. This is a new type of interface that uses diagrams.

Natural Language Interfaces

Natural language processing (NLP) is a set of techniques and algorithms used to interpret, understand, or generate natural language. Natural language processing is already used by search engines when they process the texts they provide access to. For example, recognizing noun phrases and being able to ignore prepositions requires natural language processing. Most of these techniques have been applied at the backend of search engines. However, newer applications are experimenting with interacting with users using natural language or improving query processing using more advanced NLP. Quite often, the developers hope to get more information from users by encouraging them to use their own language. Interestingly, these search engines continue using the single search text box in their interfaces.

Several new search engines are available. Powerset (www.powerset.com) uses natural language processing of queries and texts to search for Wikipedia articles. True Knowledge (www.trueknowledge.com) relies heavily on NLP to automatically answer questions. Their knowledge base consists of Wikipedia and manually contributed knowledge. Their results provide an explicit answer to factoid questions and a list of websites to back up the answer. Hakia (www.hakia.com) is a semantic search engine with a SemanticRank ranking algorithm that uses concept-based instead of word-based matching. Hakia employs natural language processing and is taking on Google directly, by allowing users to compare results on their website.

Naturally, in addition to adding advanced NLP techniques to the front end, the newer search engines also aim to improve their underlying algorithms with more NLP. A health information search engine, Healia (www.healia.com), allows keyword-based searching of the web, ClinicalTrials.gov and PubMed. With the results, filters are provided that can be applied, such as limiting results to females or teenagers.

Query Diagram Interfaces

While searching for text, the input method does not have to be limited to phrases or sentences in a text box. Instead of one line of text, a figure or diagram could be used. By changing the input method, more options or affordances become available to work with. An affordance, first coined by Gibson [41], is a property of an entity or object that allows interaction with that object in a specific way. Manipulation of affordances has been used for decades to guide our interaction with the physical environment. For example, the type of door handle will influence whether you try to push or pull. Doors with a flat metal plate as door handle are meant to be pushed. On the Internet, such physical affordances are mimicked to help users behave in similar ways, e.g., buttons are *pushed* and checkboxes are *checked*,

although the physical behavior is a *mouse click* for both. A change in affordances can have large scale consequences. Baron [42] provides several examples of behavioral and social change, sometimes unintended, as a result of changed affordances. Phones becoming mobile allows us to roam (an affordance) and this make us available everywhere all the time. Another one of her examples is instant messaging, which facilitates multi-tasking during conversations. One can compose an essay while carrying on several IM conversations simultaneously.

In contrast to linear text boxes, a 2-dimensional interface consisting of multiple search boxes with connections between those searches boxes can be used to form a query. Figure 1 shows an example query for the question “Can sports be used to treat depression in teenagers?” Each box represents a search term, e.g., “depression.” As with current search engines the keywords can consist of a single word or a phrase. The labels on the arrows and the directionality of these arrows show how the search terms need to be related to each other: “treats.” In addition to encouraging more keywords, this interface provides two additional, easy-to-use affordances. The first is the use of relationships between keywords and the second is the option to add meta-information to search terms, for example “medication.” With this type of query, a user can specify very precisely what the question is. Although these precise queries may be unnecessary for simple factoid questions, they will allow more precise and useful searches in long texts, a search type not very efficiently executed today. To enable this type of searching, different pre-processing techniques and data structures are needed based on predicates instead of phrases. Users can draw search diagrams but current document representations and retrieval mechanisms cannot generally employ the extra information. To leverage this type of user query, search engines need to improve four components: 1) the user interface so that diagrams can be drawn, 2) the document representation so that predicates are used to represent contents, 3) the search mechanism to match diagrams to the predicates, and 4) the results presentation to make better use of the information available.

To evaluate if everyday Internet users, who customarily use existing search engines with a single input text box, would be willing and able to use a different type of query, a user study was completed with 22 users [43]. The Google interface was compared to two versions of a diagram search interface: Template and Blank Diagram queries. In general, users adopted the query diagrams easily, even with limited training. They had no difficulty grasping the idea of diagram-based searching. This different query method led to more search terms being used in comparison to Google-type queries. Moreover, the search terms were used in a structured format and relationships were added between them. This, in turn, provides even more information useful to a search engine. Even so, users preferred using Google, which was not unexpected given the years of training and comfort levels that have been achieved using Google. They found it easier to use Google and also expected better results. When asked to compare the use of template diagrams versus drawing their own, users preferred to form their own diagrams and also expected these to lead to better results.

Searching for Different Media

Search engines that provide access to text are the most common and the best developed and most advanced. With different media becoming increasingly popular, text search engines have added those collections, such as photos, music, video, and made them available for

searching. However, they often do this by matching user keywords to indexed text. For example, Google, Yahoo! and Kazaa (www.kazaa.com) use the text found with the images to enable search. The text is provided by authors, taken from the surrounding image text, or from the filenames themselves. Others, e.g., Bing (www.Bing.com), use an alternative interface and require users to combine meta-information about the images, such as ‘square’ and ‘people’. This meta-information reflects the image itself, not the text surrounding it.

With improvements and dropping prices in both hardware and software, new search alternatives become available. With non-text media, the relevant features need to be extracted as the basis for search and matching results, e.g., colors or sound waves. In addition, an intuitive user interface is needed since few users are willing to spend time learning how to use a search engine. Most of the advances are made in the algorithms underlying these search engines. Similar to a text search engine, the images need to be indexed before they become accessible for search. This can be done, for example, by using wavelets instead of words and phrases. For each image, multi-resolution wavelets decompositions are constructed. The coefficients of these form a signature of images which can then be compared against signature of other images [44], e.g., an images drawn by users or an image for which similar images are needed. Current research focuses on developing alternatives and improving the algorithms to match images to other images, to match images to drawings and to video, and to effectively index large collections and making them available. Additional applications of such searches are face recognition applications for images or video or the recognition of objects from sky images. Once the underlying algorithms are developed and validated, applications can go a step further and e.g., group similar music by clustering based on acoustic features [45], perform analysis of artwork authorship, etc.

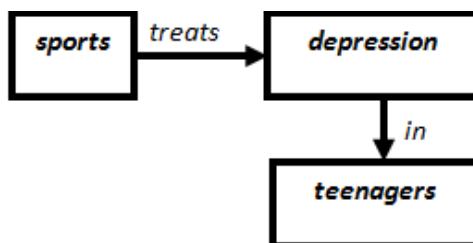


Figure 2. Example Query Diagram.

Search engines for music, objects, or images are becoming available for use by the general public. One of the earliest such search engines is Retrievr (<http://labs.systemone.at/retrievr/>). It indexes images from Flickr (www.flickr.com) which can then be searched by drawing sketches. Another is Gazopa (www.gazopa.com), where users can upload an image which is then compared, based on extracted feature images, to available images on the web. There are many more prototypes available online and it can be expected that several will turn into full-fledged search engines. Similar to image search, now also voice, sound, and music search have become possible. A user can sing, hum, or whistle a song at Midomi (www.midomi.com) to find the matching music. Video search engines are also becoming available. For example, Blinkx (www.blinkx.com) extracts information from the video, sound, voice, and images. Interestingly, searching in this case is still based on a single search text box.

Searching by Different Users

Today's search engines are optimized for the average person who does not have physical or mental disabilities. However, with increasing use of search engines by all people and for a variety of devices, improvements in the interface are becoming a necessity. Many ergonomic guidelines apply to the search interface design for users with special needs. Users with reduced cognitive abilities, memory problems, or lack of experience may require help. Elderly or people with disabilities who have impaired fine motor skills may need a different interface. Buttons and scroll bars need to be larger and the text size needs to be adjustable.

Normal visual abilities are necessary for our common search engines. However, some improvements have recently been made for users without good eyesight. Blind users use screen readers and voice browsers. These suffice for very simple search pages, i.e., with only a text box and search button, but they are insufficient for search engines integrated in other websites, e.g., e-commerce sites. Most shortcomings are the result of web pages that are designed to be scanned while blind users necessarily have to rely on sequential, line-by-line information and navigation. For example, excessive sequencing of information, lack of context, missing expressive power, but also information overload and difficulty navigating are especially cumbersome [46, 47]. Interfaces, such as Google's interface [48], can be simplified for screen readers with increased satisfaction for users. More advanced solutions include inclusion of haptic feedback and sounds to be associated with specific navigation and html objects. For example, a force feedback mouse can be used to provide haptic feedback and enable blind users to distinguish between different elements in Google's interface [49].

Searching with Different Devices

With increasingly more powerful but smaller processors and better, cheaper displays, we can expect many alternatives for searching for information. Intelligent watches, eyeglasses, see-through computer displays, and smart clothing may become search enabled. Alternatively, with established devices becoming more intelligent, such as display boards, ATMS, and gaming consoles such as Wii for health, we can expect also different types of search at different places. Naturally, with mobile devices becoming affordable to large groups in the community, even where computers are not, e.g., many parts of Africa, their increased use makes them the next most common device to search and browse online. The obvious difference between website and mobile device that should be taken into account by search engines is their small screen to form queries and display results. Additional features, such as location and position awareness, may bring new search strategies.

Mobile search engines can be expected to be the next most common type of search engine. Comparable to early research on web site search engines, there currently are enormous differences in coverage of content and large differences in precision and recall between the different alternatives. For example, 4Info (www.4info.com) is a mobile search engine based on text messaging. A message is texted and the answer is received by text message. The content covered is limited (list available online) and focuses on entertainment, such as sport scores, movies, horoscope. Church et al. [50] compared 7 mobile search engines, Google, Moooble, Click4WAP, Seek4WAP, WAPAll, WAPly, Ithaki, with their own work and discuss the differences in coverage and information provided by each of the

alternatives. Although these search engines still rely significantly on typed queries, the use of natural language processing to enable spoken queries seem a natural fit for mobile devices. Fabbrizio et al. [51] describe a mashup architecture to deliver speech-to-text services for mobile devices. Paek et al. [52] combine voice with text hints for higher reliability in their mobile search interface: Search Vox.

Conclusion

A search engine is composed of at least four layered components: a user interface for querying, a user interface for result presentation, a representation layer to represent and store the information in the documents or other media, and a search mechanism to match the user query and underlying media. In addition, many search engines use a fifth component: algorithms to collect and refresh their document collections. To significantly improve a search engine, advanced in each of the components are often required, making it a non-trivial exercise.

With advances in displays, e.g., flexible displays, total immersion and other 3-dimensional displays, and additional modalities, such as smell and touch, the future is bright for search engines and search engine research. When these hardware improvements will be combined with software improvements, e.g., texture interfaces for blind users, we will see an entire new generation of interfaces and search engines.

References

- Broder, A. (2002). "A taxonomy of web search," *ACM SIGIR Forum*, Vol. 36, 3-10, September.
- Martin, M. J. (2004). "Reliability and Verification of Natural Language Text on the World Wide Web," in *ACM-SIGIR Doctoral Consortium*, Sheffield, England.
- Jansen, B. J., Booth, D. L. & Spink, A. (2007). "Determining the user intent of web search engine queries," in 16th international conference on World Wide Web, Banff, Alberta, Canada, 1149-1150
- Navarro-Prieto, R., Scaife, M. & Rogers, Y. (1999). "Cognitive strategies in web searching," in 5th Conference on Human Factors & the Web, Gaithersburg, Maryland 20899.
- de Lima, E. F. & Pedersen, J. O. (1999). "Phrase Recognition and Expansion for Short, Precision-biased Queries Based on a Query Log," in 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Berkeley, CA USA, 145-152.
- Ross, N. C. M. & Wolfram, D. (2000). "End User Searching on the Internet: An Analysis of Term Pair Topics Submitted to the Excite Search Engine," *Journal of the American Society for Information Science*, Vol. 51, 949-958.
- Spink, A., Wolfram, D., Jansen, M. B. J. & Saracevic, T. (2001). "Searching the Web: The Public and Their Queries," *Journal of the American Society for Information Science and Technology*, Vol. 52, 226-234.

- Toms, E. G., Kopak, R. W., Bartlett, J. & Freund, L. (2001). "Selecting versus Describing: A Preliminary Analysis of the Efficacy of Categories in Exploring the Web," in Tenth Text REtrieval Conference (TREC 2001), Maryland.
- Hölscher, C. & Strube, G. (2000). "Web search behavior of Internet experts and newbies," in Ninth International World Wide Web Conference, Amsterdam, The Netherlands, 337-346.
- Lau, T. & Horvitz, E. (1998). "Patterns of Search: Analyzing and Modeling Web Query Refinement," in Seventh International Conference on User Modeling.
- Muramatsu, J. & Pratt, W. (2001). "Transparent Queries: investigation users' mental models of search engines," in 24th annual international ACM SIGIR conference on Research and development in information retrieval, New Orleans, Louisiana, United States, 217-224.
- Baker, L., Wagner, T. H., Signer, S. & Bundorf, M. K. (2003). "Use of the Internet and E-mail for Health Care Information: Results from a National Survey," *Journal of the American Medical Association*, Vol. 289, 2400-2406, May 14.
- Fox, S. & Fallows, D. (2003). "Internet Health Resources - Health searches and email have become more commonplace, but there is room for improvement in searches and overall Internet access.," Pew Internet & American Life Project, Washington D.C. July 16.
- Peterson, E. B. (2005). in Library Trends, *Health information literacy: a library case study*.
- Warner, D. & Procaccino, J. D. (2004). "Toward Wellness: Women Seeking Health Information," *Journal of the American Society for Information Science and Technology*, Vol. 55, 709-730, June.
- McCray, A. T. & Tse, T. (2003). "Understanding Search Failures in Consumer Health Information Systems," in *AMIA Symposium*, Washington, DC, 430-434.
- Coiera, E. W. & Vickland, V. (2008). "Is Relevance Relevant? User Relevance Ratings May Not Predict the Impact of Internet Search on Decision Outcomes," *Journal of the American Medical Informatics Association*, Vol. 15.
- Salton, G. (1971). "The SMART Retrieval System: Experiments in Automatic Document Processing," in *Automatic Computation*, G. Forsythe, Ed.: Prentice-Hall.
- Salton, G. & Buckley, C. (1990). "Improving Retrieval Performance by Relevance Feedback," *Journal of the American Society for Information Science*, Vol. 41, 288-297.
- Page, L., Brin, S., Motwani, R. & Winograd, T. (1998). "The PageRank Citation Ranking: Bringing Order to the Web," Stanford Digital Library Technologies Project.
- Mandl, T. (2006). "Implementation and evaluation of a quality-based search engine," in Proceedings of the seventeenth conference on Hypertext and hypermedia, Odense, Denmark, 73-84.
- Fan, W., Gordon, M. D. & Pathak, P. (2000). "Personalization of Search Engine Services for Effective Retrieval and Knowledge Management," in International Conference on Information Systems (ICIS), Brisbane, Australia, 20-34.
- Fan, W., Gordon, M. D. & Pathak, P. (2004). "Discovery of context-specific ranking functions for effective information retrieval using genetic programming," *IEEE Transactions on Knowledge and Data Engineering*, Vol. 16, 523- 527.
- Fan, W., Pathak, R. & Wallace, L. (2006). "Nonlinear ranking function representations in genetic programming-based ranking discovery for personalized search," *Decision Support Systems*, Vol. 42, 1338-1349.

- Dou, Z., Song, R. & Wen, J. R. (2007). "A large-scale evaluation and analysis of personalized search strategies," in 16th international conference on World Wide Web, Banff, Alberta, Canada, 581-590
- Coyle, M. & Smyth, B. (2007). "On the community-based explanation of search results," in 12th international conference on Intelligent user interfaces, Honolulu, Hawaii, USA, 282-285.
- Silva, J. M. & Favela, J. (2006). "Context Aware Retrieval of Health Information on the Web," in Fourth Latin American Web Congress (LA-WEB'06), 135-146.
- Harman, D. (1988). "Towards Interactive Query Expansion," in Eleventh International Conference on Research & Development in Information Retrieval, New York, 321-331.
- Harman, D. (1992). "Relevance Feedback Revisited," in 15th International ACM/SIGIR Conference on Research and Development in Information Retrieval.
- Magennis, M. & Rijsbergen, C. J. V. (1997). "The Potential and Actual Effectiveness of Interactive Query Expansion," in the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 342-332.
- Yang, K. & Maglaughlin, K. (1999). "IRIS at TREC-8," in Eighth Text REtrieval Conference (TREC 8), Maryland, 645.
- Yang, K., Maglaughlin, K., Meho, L., Robert, J. & Sumner, G. (1998). "IRIS at TREC-7," in Seventh Text REtrieval Conference (TREC 7), Maryland, 555.
- Hawking, D. & Craswell, N. (2001). "Overview of the TREC-2001 Web Track (TREC 2001)," in Tenth Text REtrieval Conference, 61-68.
- Amati, G., Carpineto, C. & Romano, G. (2001). "FUB at TREC-10 Web Track: A Probabilistic Framework for Topic Relevance Term Weighting," in Tenth Text REtrieval Conference (TREC 2001), Gaithersburg, Maryland, 182-192.
- Leroy, G., Lally, A. M. & Chen, H. (2003). "The Use of Dynamic Contexts to Improve Casual Internet Searching," *ACM Transactions on Information Systems*, Vol. 21, 229-253, July.
- Bodner, R. C. & Chignell, M. H. (1998). "ClickIR: Text Retrieval using a Dynamic Hypertext Interface," in Seventh Text REtrieval Conference (TREC 7), Maryland, 573.
- Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G. & Ruppin, E. (2002). "Placing Search in Context: The Concept Revisited," *ACM Transactions on Information Systems*, Vol. 20, 116-131, January.
- Jansen, B., Spink, A. & Saracevic, T. (2000). "Real Life, Real Users, and Real Needs: A Study and Analysis of User Queries on the Web," *Information Processing and Management*, Vol. 36, 207-227.
- Google Labs, "Google Suggest [<http://www.google.com/webhp?complete=1&hl=en>]."
- Yahoo!, "Search Assist [<http://tools.search.yahoo.com/newsearch/searchassist>]."
- Gibson, J. J. (1986). *The ecological approach to visual perception*. New Jersey: Lawrence Erlbaum Associates.
- Baron, N. S. (2008). *Always On: Language in an Online and Mobile World*. New York: Oxford University Press.
- Leroy, G. (2009). "Persuading Consumers to Form Precise Search Engine Queries," in *American Medical Informatics (AMIA) Fall Symposium*, San Francisco.
- Jacobs, C. E., Finkelstein, A. & Salesin, D. H. (1995). "Fast Multiresolution Image Querying," in 22nd Annual Conference on Computer Graphics and Interactive Techniques, 277-286.

- Frank, J., Lidy, T., Peiszer, E., Genswaider, R. & Rauber, A. (2008). "*Ambient Music Experience in Real and Virtual Worlds Using Audio Similarity*," in 1st ACM International Workshop on Semantic Ambient Media Experiences, Vancouver, British Columbia, Canada.
- Leporini, B., Andronico, P. & Buzzi, M. (2004). "*Designing Search Engine User Interfaces for the Visually Impaired*," in Proceedings of the 2004 International Cross-disciplinary Workshop on Web Accessibility (W4A), 57-66.
- Takagi, H., Saito, S., Fukuda, K. & Asakawa, C. (2007). "Analysis of navigability of Web applications for improving blind usability," *Transactions on Computer-Human Interaction (TOCHI)*, Vol. 14, Article 13.
- Andronico, P., Buzzi, M., Castillo, C. & Leporini, B. (2006). "*A prototype of google interfaces modified for simplifying interaction for blind users*," in 8th International ACM SIGACCESS Conference on Computers and Accessibility, Portland, Oregon, USA 267-268.
- Kuber, R., Yu, W. & McAllister, G. (2007). "*Towards Developing Assistive Haptic Feedback for Visually Impaired Internet Users*," in SIGCHI Conference on Human Factors in Computing Systems, San Jose, CA, USA, 1525-1534.
- Church, K., Smyth, B. & Keane, M. T. (2006). "*Evaluating Interfaces for Intelligent Mobile Search*" in 2006 international cross-disciplinary workshop on Web accessibility (W4A), Edinburgh, U.K. 69-78
- Fabbrizio, G. D., Okken, T. & Wilpon, J. G. (2009). "*A Speech Mashup Framework for Multimodal Mobile Services*" in 2009 International Conference on Multimodal interfaces, Cambridge, Massachusetts, USA 71-78
- Paek, T., Thiesson, B., Ju, Y. C. & Lee, B. (2008). "*Search Vox: Leveraging Multimodal Refinement and Partial Knowledge for Mobile Voice Search*" in 21st Annual ACM Symposium on User Interface Software and Technology, Monterey, CA, USA, 141-150.

Chapter 6

SOCIAL MOTIVES FOR WINDOWS LIVE MESSENGER USE: RELATIONS TO SOCIAL CAPITAL THEORY AND DEPRESSIVE PERSONALITY STYLES

Craig Ross and Emily S. Orr

University of Windsor, Ontario, Canada

Abstract

The purpose of this chapter is to explore the motivations associated with Windows Live Messenger (Messenger) use within the context of social capital. Furthermore, the present chapter will also explore these motivations in relation to depressive personality styles. Motivations for the use of online technology can be understood within the framework of social capital (Williams, 2006) which proposes that the appropriate use of communication tools can lead to interpersonal and coping benefits for the individual user. Our results identified three motivations for Messenger use: Emotion Regulation and Coping, Positive Practicality, and Passing Time. These motivations were consistent with our understanding social capital theory. Further to our understanding of Messenger use from a social capital perspective, the present data supports that these motivations can also be understood within the context of clinically-related personality styles. Blatt (1974) argued that depression stems from one of two personality dimensions: self-criticism or interpersonal dependency. Individuals who exemplify either style are particularly prone to repeated episodes of depression. The relationship between those motives identified above and the personality styles identified by Blatt (1974) will be explored. Taken together, this chapter will explore how the motivations for Messenger use relate to social capital theory as well as within the context of depressive personality.

Introduction

Regardless of the specific function of the Internet, people utilize this technology to communicate with others. Previous research (e.g., Heinrich & Gullone, 2006) indicates that “satisfying social relationships are vital for good mental and physical health” (p. 696). As such, it is not surprising that individuals are motivated to maintain social connections with

others (Baumeister & Leary, 1995). When beneficial social relationships are not possible, individuals are at higher risk for both physical and emotional disorders (House, Landis, & Umberson, 1988).

The purpose of this chapter is to explore the motivations associated with Windows Live Messenger (herein referred to as Messenger) usage within the context of social capital. Furthermore, the present chapter will also explore these motivations in relation to individual differences. For some, the use of instant messaging (IM) technologies appears to be eroding traditional social structures by eliminating face-to-face interactions or more traditional forms of communication (Putnam, 2000). However, a closer examination indicates that there may actually be considerable social capital realized through the use of IM technologies such as Messenger (e.g., Wellman, Quan Haase, Witte, & Hampton, 2001).

Social Capital

In its current form, social capital is the result of theoretical contributions by numerous researchers. Common to all theories, however, is that the concept of social capital has emerged as a way of describing the sorts of interpersonal and institutional benefits that emerge through relationships with others (Bassani, 2007). Social capital, therefore, is understood as describing the processes by which people interact, the benefits that can result from these interactions and the various outcomes that can ensue. While some disciplines such as political science and economics understand social capital from a *group* perspective (Bassani, 2007), research in psychology has demonstrated that social capital can emerge from the types of social support (e.g., personal advice and emotional reinforcement) that result from interactions with individuals (Beaudoin & Tao, 2007).

Until recently, the concept of social capital has not been widely embraced by psychology researchers. One of the largest problems with traditional notions of social capital is that the term does not immediately express the underlying theory (Williams, 2006). Williams (2006) noted that social capital engenders an idea of financial capital in that individuals can choose to generate, save or redeem social resources. Thus, much as a person can have money saved in a bank, an individual can stockpile a series of personal favours that can be used in future interactions. However, social and financial resources are different in that social resources require interpersonal interactions to be realized. In this way, Williams observes that social capital can be viewed as both an outcome (e.g., positive social bonds) and a process (e.g., seeking social support or helping a friend move).

Portes (2000) argues that social capital refers to the benefits that exist or are created when individuals form relationships with others. These benefits can be felt directly by the individual (e.g., being able to arrange help when moving), or can help to build a stronger society (e.g., creating safer neighbourhoods). According to Wellman et al., (2001), use of tools such as Messenger most closely fits with the concept of network capital. Network capital concerns relationships with friends, neighbours and family that significantly provide companionship, emotional aid, information and a sense of belonging.

Within the increasingly connected world, there also appears to be changes in the sources of social capital. For example, Green and Brock (2005) observed that in today's society, social capital is more likely to be generated through informal means such as talking to neighbours than through more formal means such as churches and social organizations. In

part, this may be due to the fact that modern life is too hectic to engage in regularly scheduled social activities. Thus, it is encouraging to note that the levels of social capital generated in online relationships are similar to the levels observed in face-to-face interactions such as club and church participation (Best & Krueger, 2006).

But there is more to understanding social capital than to simply observe the outcomes of friendship and social support when a collection of individuals begin to associate. For example, Francescato, Mebane, Porcelli, Attanasio, and Pulino (2007) noted that there are two distinct ways in which social capital can be created. The first, bridging capital, occurs when characteristics such as trust, norms and networks improve the collective of society. In this way, bridging capital is understood as a resource that is part of the social network and which ties one individual to a number of other individuals (Adler & Kwon, 2002). The idea of bridging capital is most consistent with the ideas of Putnam (2000) which will be discussed below. The second approach, bonding capital, refers to the types of capital that an individual can achieve through personal networks. Bonding capital tends to focus more on the internal characteristics of the individuals in the relationship and moves away from the reciprocal, instrumental quality that defines bridging social capital (Adler & Kwon, 2002). With these two categories defined, it can be observed that bridging capital is dependent on the particular social network or communication medium in question, whereas bonding capital is more influenced by the efforts of the individual. From a technological viewpoint, tools such as Messenger automatically provide the possibility for bridging capital because they are designed to connect people. However, in order for bonding capital to develop, an individual must be willing to participate in discussions and be willing to invest effort in using the tool.

One of the most common venues for the existence of bridging capital is in voluntary organizations where trust and cooperation are required (Stolle, 1998). Because of the connections that are created in these types of organizations, bridging social capital is viewed as inclusive (Putnam, 2000). In other words, bridging social capital allows individuals from different backgrounds to connect and experience alternative views of the world. Not surprisingly, research with job-seekers has found that the most successful job-seekers are the ones with the greatest bridging capital (Williams, 2006).

Despite the relatively large number of contacts that can be generated through bridging capital, little emotional support is created because relationships tend to be surface-level (Williams, 2006). In fact, Green and Brock (2005) observed that individuals who were more organized in their affiliations tended to demonstrate higher levels of social capital and civic competence, whereas those who were more informal in their interactions were found to build greater feelings of connectedness and support. In part, this may indicate that social capital as a resource has to be mobilized in order for an individual to experience the benefit of their generated social capital (Bassani, 2007). Thus, although individuals with high levels of bridging capital may have stockpiled large amounts of social capital as a resource (e.g., they are owed favours), they may have difficulty in mobilizing those resources because of the relatively weak connections they have with their social contacts (i.e., low bonding capital). This mobilization can be even more difficult for contacts which exist solely on Messenger, if such mobilization requires face-to-face contact.

Bonding capital, on the other hand, is primarily intended to provide emotional benefits to the individual. By definition, groups dominated by bonding capital are composed of strongly-tied individuals such as family and close friends (Williams, 2006). Thus, when the process of utilizing social capital is understood in terms of reciprocity, prosocial norms, the sharing of

information and interpersonal trust (Best & Krueger, 2006), it is not surprising that groups defined by close relations are the ones which are most likely to benefit from high levels of social capital. In other words, it is much easier to make substantive requests of those with whom you have developed a strong relationship than those who are only casual acquaintances.

Given the appropriate conditions, situations designed to elicit bridging capital can encourage bonding capital as well. In a study of a neighbourhood community provided with electronic communication tools including IM, Hampton and Wellman (2003) observed a trend by which community members significantly increased their connections in the neighbourhood through the available technology. Most notably, those with access were recognized by name three-times more often, talked with twice as many neighbours and visited 50% more of the community than those without online community access. Moreover, the provided tools were often used for the types of information exchange and social support that define bonding capital.

Despite the opportunities for emotional support, there are some risks to bonding capital when Internet tools are considered. Most notably, individuals have a tendency to reinforce pre-existing ties rather than to create new ones (Matei & Ball-Rokeach, 2001). This process seems to underscore criticisms by Best and Krueger (2006) that increasing time spent with previously known individuals can actually decrease or inhibit social capital because it prevents the development of new relationships.

With its focus on the importance of social interaction, the framework of social capital, including bridging and bonding capital, can be used to understand the effects of technology on interpersonal relationships.

Computer Mediated Communication

From the perspective of social capital, it now appears that the Internet is not necessarily understood as an inherently dangerous entity in terms of eroding our social structures, social communities and relationships. While it may be true that some early Internet adopters suffered negative social consequences by spending large amounts of time online (Kraut et al., 1998), it appears that today's users are more savvy and can actually benefit from online activities (Kraut et al., 2002). Whether it is the result of exposure to different perspectives (Bryant, Sanders-Jackson, & Smallwood, 2006), the ability to find others of the same ethnic group (Tynes, Reynolds, & Greenfield, 2004), the opportunity to practice social skills in a less threatening environment (Mazur, Burns, & Emmers-Sommer, 2000), or some other phenomenon, the Internet allows individuals of all ages to broaden horizons and experience situations that might not be possible with more traditional forms of communication.

In order to understand how some of these benefits can result from Internet use, it is necessary to draw some distinctions between common Internet practices. For example, while many people may equate the Internet with the World Wide Web, this would only reflect one component of Internet use. Although important, the World Wide Web is more about information gathering than it is about communication. Therefore, it is only when other components such as e-mail, IM and chat rooms are considered that the Internet as a communication tool can be properly understood. The use of these tools falls into the category

of Computer Mediated Communication (CMC) which reflects the fact that computer technology is the medium through which individuals interact.

Perhaps one of the most compelling aspects of CMC is that it can take so many different forms. When one thinks of the telephone, there is a relatively narrow spectrum of expected uses. For example, the phone can be used to communicate with family or friends, to make appointments, and in some limited capacity can also be a source of information when used to contact specialists and professionals. When dealing with CMC, however, there are a number of considerations that must be made. For example, while almost everyone has a telephone and there are standards to ensure that telephones from one service provider can be used with telephones from a different service provider, the same cannot be said with CMC technologies. Not only are there different formats available depending on the type of communication desired (e.g., real-time chat versus letter-writing), there are also different formats for different service providers. For example, if a user prefers Messenger, that person may be cut off from potential online contacts that choose to use AOL Instant Messenger or ICQ. Therefore, even with this simple example, it becomes clear that CMC can be significantly different than some of the more traditional forms of interpersonal communication. Thus, it cannot be assumed that correlates of these traditional methods hold true for contemporary CMC tools. One such correlate is that of individual motivations for CMC use.

Individual Motivations

The theory of social capital is most effective in describing the broad social reasons that individuals might choose to use CMC tools like Messenger. However, there are also a number of intermediate motives which can also be significant. A motive is a construct that encompasses specific goals, desires, and needs, which are relevant to an individual's well-being (Horowitz, 2004). There are a number of different motives (e.g., an intimacy motive), each of which is made up of many goals (e.g., wanting to spend time with your romantic partner). In other words, we conceptualize motives as hierarchies wherein motives represent a higher level of abstraction, and specific goals, desires and needs represent the specific components that comprise the motives. In the context of social capital, we might understand motives as serving one function. In other words, while social capital can be understood as describing the overall processes and benefits that can result in interpersonal relationships, motives describe the drives that encourage individuals to act in a certain ways to facilitate the generation and maintenance of social capital. Thus, within the context of social capital, there exists a hierarchy of broad, intermediate motivations and specific goals that comprise these motives.

Researchers such as Leung (2001) have investigated some of the personal motivations that can lead to technology use. Leung (2001) conducted a preliminary investigation of motives associated with usage of ICQ (i.e., 'I seek you'), an IM program which predated Messenger. He identified seven motives for ICQ use. These motives included affection (when a user expressed affection toward others), entertainment, relaxation, fashion (when an individual used ICQ to appear stylish), inclusion (when an individual desired to feel deeply involved in a relationship), sociability (when an individual wanted to meet new people), and escape (when an individual used ICQ to avoid other responsibilities). The results revealed that affection, entertainment, inclusion, escape, and sociability were all significantly positively

correlated with frequency of ICQ use. Moreover, Leung found that entertainment, inclusion, sociability, and affection were also significantly related to the time spent on ICQ in a typical session (i.e., intensity of use).

Of course, many of the motivations described by Leung (2001) can be understood within the social capital framework. For example motivations related to inclusion, sociability and affection likely fall under the category of bonding capital (Adler & Kwon, 2002). Likewise, most motivations such as affection, inclusion and sociability are representative of what Wellman and colleagues (2001) termed network capital. It can also be observed that, should an individual not demonstrate the motives associated with the usage of CMC tools that facilitate social capital benefits, they may be at risk for the development or continuation of psychological pathology.

Poor Relationships and Depression

Historically, there have been multiple models used to explain the construct of depression (Beck, 1973). Within the conceptualization of Blatt (1974), depression is manifested in one of two forms: anaclitic depression or introjective depression. Those individuals who are more sensitive to interpersonal loss develop an “anaclitic depression” (Blatt, 1974). Anaclitic (herein referred to as dependent) depression is characterized by feeling unloved and helpless. Individuals who are depressed in this manner desire protection and feeling cared for and demonstrate abandonment as their greatest fear. Due to their fear of abandonment, these individuals often suppress feelings of anger in fear of losing an individual that can satisfy their need for love and support (Blatt, 1974). These individuals are also susceptible to developing depression when they perceive rejection or a loss of support (Blatt & Zuroff, 1992). This fear of loss is not only present when these individuals are depressed, but is a pre-existing personality profile, known as a dependent personality, that is susceptible to depression (Blatt and Zuroff, 1992).

Personal failure, as opposed to issues of abandonment, leads to introjective depression (Blatt, 1974). Introjective (herein referred to as self-critical) depression is characterized by feelings of guilt and unworthiness (Blatt, 1974). These self-critical individuals are most vulnerable to developing depression when they perceive failure on their part or when they feel they cannot control their environment (Blatt & Zuroff, 1992). Like the dependent personality configuration, those suffering from self-critical depression also exhibit a pre-existing personality profile that leaves them vulnerable to developing depression (Blatt & Zuroff, 1992). This is known as the self-critical personality.

Blatt's (1974) model, which stresses that depression is manifested in persons who are highly dependent on others or who are very self-critical, has been replicated by researchers from a number of different theoretical perspectives (Blatt & Maroudas, 1992). Although the two distinct experiences can lead to depression or depressive feelings, these two processes (self-criticism or dependency) are not exclusive of one another, as an individual can be susceptible to both vulnerabilities. Although individuals can experience both dimensions, there is usually more of a vulnerability to one dimension over the other (Blatt & Zuroff, 1992).

In terms of Messenger use, individuals with dependent personalities are most likely to be sensitive to the unavailability of online contacts or be negatively impacted when someone

declines a contact request. Conversely, individuals with self-critical personalities are likely to enjoy the kind of control that is possible in an IM environment. For example, it is relatively easy to control who is able to contact you through block lists and the intentional posting of status (e.g., "Available" or "Busy") can help to determine when others are allowed to communicate with you.

The Present Research

To date, very little work has been done to try and understand individual motivations from the perspective of social capital. This seems to be a significant omission, however, as many of the motivations which have been found to influence the use of tools like Messenger often have a relation to social capital. In order to examine the potential influence of social capital and individual motivation, a study was conducted which focussed solely on Messenger. This particular tool was chosen as it has the greatest market share, being selected by 61% of worldwide IM users (Lipsman, 2006). Given that tools like Messenger are so popular, they are a valuable means to investigate motivations which match the traditional understanding of social capital.

Method

Participants

Participants were recruited through three different venues to ensure a wide-ranging sample. These venues included: the psychology participant pool at a university in Southwestern Ontario, Canada; the Facebook profile of the secondary author (posted via an open Event function); and through the online free-advertisement site, Kijiji (posted on the sites for all Canadian listings). Participants were recruited as part of a larger parent study investigating the motives associated with usage of CMC tools and other forms of contemporary communication.

One hundred fifty-five Messenger users (72.3% women) participated in the present study. The sample had a mean age of 22.92 years ($SD = 4.49$). A total of 75.5% of the participants came from the participant pool. Distinctions between participants recruited through Facebook and Kijiji could not be made. All participants indicated that they had utilized Messenger within the week prior to their participation in the present study.

Materials

A questionnaire was developed to assess the motives of individuals for using Messenger. The questionnaire consisted of 99 specific items (e.g., goals, desires, and needs) that were drawn from previous research in the domain of CMC motives (e.g., Amiel & Sargent, 2004; Leung, 2001; Shepherd & Edelmann, 2005) and that were developed by the authors for the purpose of the larger parent study.

In order to assess participants' personality styles, the Depressive Experiences Questionnaire (DEQ; Blatt, D'Affliti, & Quinlan, 1976) was administered. The DEQ is a 66-

item Likert style self-report measure, originally developed to categorize everyday depressive-related experiences. The measure specifically focusses on items that are correlated with depression (e.g., "I urgently need things that only other people can provide") but that are not symptoms of the disorder (Zuroff, Moskowitz, Wielgus, Powers, & Franki, 1983). Participants indicate their agreement to each statement on a scale from "*1 – Strongly Disagree*" to "*7 – Strongly Agree*." Repeated data reduction analysis of this scale has suggested that the measure is comprised of three subscales: self-criticism, dependency, and self-efficacy (Zuroff, Quinlan, & Blatt, 1990). Participants' responses are scored via computer to generate standard scores, multiplied by a factor weight, and then summed to produce scores for each of the three sub-scales (Nietzel & Harris, 1990).

The DEQ has demonstrated adequate test-retest reliability and construct validity (Zuroff et al., 1983). Moreover, the dependency and self-criticism scales of the DEQ have been found to correlate with scores on the Beck Depression Inventory-II and depressive affect more generally (Zuroff et al., 1990).

Procedure

An online questionnaire was developed by the first author to administer the materials. Participants who indicated that they had used Messenger within the week prior to survey completion were asked to complete the motives questionnaire for Messenger usage, the DEQ, and additional measures relevant to the parent study. Individuals who indicated that they had not used Messenger within the week prior to survey completion were dropped from the analyses for the present study.

Results

Data Reduction

A Principle Component Analysis (PCA) was conducted on the Messenger motives questionnaire. PCA was selected as the data reduction method as it allows for the maximum amount of variance within each component that is extracted (Tabachnick & Fidell, 2007). Moreover, PCA reduces large numbers of variables to a smaller set of components, thus summarizing the data into related constructs.

The preliminary, unrotated, PCA scree plot indicated that there were three components to be extracted from the data set. Thus, all further analyses forced three components. A second PCA was conducted forcing three factors. Subsequent to this, variables that did not load on any of the three components at acceptable levels (less than .55; Comrey and Lee, 1992) were removed from further analysis in order to facilitate interpretation. After these variables were removed, the Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) and Bartlett's Test of Sphericity indicated that the data was now suitable for interpretable analysis (Tabachnick & Fidell, 2007). Specifically, KMO = .793 (which is above the recommended cut-off of .6), while Bartlett's Test of Sphericity was $< .001$.

Another PCA was conducted with a direct oblimen rotation of $\delta = -1$. The direct oblimen rotation was selected as it is oblique, thus allowing the components to be correlated with one

another. This PCA revealed that the three factors accounted for 41.269% of the total variance in the solution. Specifically, the first component accounted for 24.110% of the variance; the second component accounted for 12.236% of the variance in the solution; while the third factor accounted for 5.283% of the variance. Within the first component, 22 items loaded at .55 or greater; six items loaded on the second component at this cut-off; and five items loaded on the third component at the .55 cut-off.

Those items that loaded on the first component were identified as being related to emotion enhancement and coping strategies. That is, participants endorsing this component indicated that they used Messenger to improve their mood, facilitate ongoing positive moods, or cope with negative stressors in their lives. The second component was associated with practicality and use of Messenger as a communication tool to connect with offline friends and family members. This component also reflected a general enjoyment of Messenger for the features it provided. The third component reflected the use of Messenger as a means of passing time or alleviating boredom. Cronbach's alpha of the first component was .937; .832 for the second component; and .793 for the third component. None of the factors were significantly correlated with one another.

Correlations

In order to investigate the relations between motives for Messenger usage and depressive personalities, Bartlett factor scores were generated for each of the three Messenger motives. Subsequently, a series of planned bivariate correlations were conducted between these factor scores and scores on the DEQ. Analysis of the Emotion Enhancement and Coping motive revealed that it was significantly correlated with the self-criticism scale of the DEQ ($r = .185, p < .05$). The Positive Practicality motive was positively correlated with the dependency scale ($r = .204, p < .05$) and self-efficacy scale ($r = .425, p < .001$) of the DEQ. Finally, the Apathy and Passing Time motive was positively correlated with the dependency scale of the DEQ ($r = .316, r < .01$).

Discussion

As a theory, social capital can explain the types of benefits and experiences that can occur when individuals interact with each other (Bassani, 2007). Current technology like Messenger, however, has the potential to substantially alter how people communicate and provide opportunities that were not possible with more traditional forms of interaction.

If social capital is considered to be the overarching force behind communication, then individual motivations are best thought of as the unique factors that drive an individual to act in a certain way. For example, a motive to regulate emotion is likely to draw a person to technologies and experiences that allow for meaningful communication with others in order to facilitate positive affect. Thus, low levels of certain motivations, or an inability to satisfy them, can have very real psychological consequences for the individual (House et al., 1988). This process can be observed in Blatt's (1974) conceptualization of depression, where depression often results from a failure to satisfy a certain motive (e.g., either personal success or connectivity with others).

An investigation of the components resulting from the Messenger motives questionnaire generally supports the social capital framework in terms of reasons for Messenger use. The Positive Practicality motive closely maps onto the ideas of bridging capital (Adler & Kwon, 2002). This motive for Messenger use involves more structural concerns such as cost and ease of communicating with others. In this way, Messenger seems to be clearly considered in terms of use as a tool, rather than as a transparent means of communication (Lewis & Fabos, 2005). In other words, there is a kind of conscious awareness that using Messenger provides interpersonal benefits, which means that Messenger is chosen for those benefits. In contrast, other “transparent” technologies like the telephone are simply a means to an end as it is rare for people to think about whether the phone might be the most beneficial way to communicate.

The Emotional Enhancement and Coping motive, however, more clearly maps onto the idea of bonding capital (Adler & Kwon, 2002). In this case, motivation is not about the “practical” matters of communication, but rather about emotional support and emotional regulation. In other words, individuals are attempting to make use of existing relationships in order to improve their mental state. This finding is reminiscent of Bassani (2007) who observed that it is possible to engage in meaningfully reciprocal relationships online. Thus, those who might be at risk for depression because of difficulties finding emotional reassurance (Blatt, 1974) might turn to Messenger to try and seek emotional support from a virtual support group.

The Passing Time motive appears to demonstrate the type of negative activities that Putnam (2000) described in his discussion of how technology was reducing levels of social capital. Most notably, this motive describes Messenger as a kind of time-waster. However, while it may be true that Messenger can impair productivity, the impacts on social relationships and social capital may not be detrimental. In thinking about the time-passing aspect of Messenger, it is likely that individuals will engage in shallow-level discussions with online contacts instead of engaging in a required task. This activity seems to be very similar to the “water cooler” conversations of traditional office settings. Thus, the Passing Time motive may be best understood as another example of how communication styles are shifting in contemporary society (Green & Brock, 2005), rather than as a negative outcome. The risk of this motive, however, is that the nature of the casual conversation pulls primarily for bridging capital-type relationships, which may not be able to provide meaningful levels of support when required (Williams, 2006).

Individual-level Implications

In addition to supporting the general concept of social capital, the three motivations derived from the Messenger-use motives questionnaire also map on to some of the individual-level processes that are correlated with clinical characteristics such as depression. Regarding Blatt’s (1974) self-critical depressive style, it was observed that there was a positive correlation between the self-criticism factor from the DEQ and the Emotional Enhancement and Coping motive. Since individuals who are high on self-criticism are more likely to judge themselves, it appears that they are using Messenger in an adaptive way by seeking opinions that might challenge their personal view (i.e., that they are a failure). Similarly, self-critical individuals are less likely to have many offline contacts, as they are more geared toward work

responsibilities than interpersonal relationships. Thus, they may be able to find social support through online means such as Messenger. This format may also be seen as less threatening than asking for help face-to-face, and thus represent an opportunity to avoid failure in a more social realm.

The use of Messenger to compensate for poor or insufficient offline relationships is an example of how individuals can build social capital with the effective use of technology. This process may help to explain the positive correlation observed between the self-efficacy subscale of the DEQ and the Positive Practicality motive. Because of their positive self-beliefs, individuals with high scores on self-efficacy are likely able to effectively control the stressors in their lives and are more likely to demonstrate adaptive coping skills. In addition, a positive correlation was observed between the Positive Practicality motive and the Dependency subscale of the DEQ. Given that individuals with highly dependent personalities attempt to surround themselves with others in an effort to avoid abandonment (Blatt, 1974), these individuals are likely to have larger offline social circles and thus use Messenger as a means to communicate with that offline social network. In terms of social capital, this may reflect the process of building, maintaining and utilizing bonding capital.

Individuals who endorse the Passing Time motive for Messenger use may indirectly derive positive social interactions (e.g., the “water cooler effect” noted above). Thus, it is not surprising that a positive correlation was observed between the Passing Time motive and dependency scores on the DEQ. That is, individuals who have highly-dependent personalities are likely to attempt maintain connections with others, even in potentially superficial ways. Unfortunately, when Messenger is used in this way, it is more facilitative of bridging capital which may not provide the kind of intimate emotional connection that those with dependency crave. Therefore, dependent individuals who use Messenger in order to pass time may be more vulnerable to developing depression as Messenger does not sufficiently satisfy their need to have deep, meaningful connections with others. Thus, although using Messenger to pass time may help a dependent person feel connected in the short-term, it is likely that the limitations of the online contact will become apparent when greater depth of interaction is desired. In order to generate a deeper connection, one would have to be motivated to use Messenger in a way that facilitates generation of bonding capital (e.g., the Positive Practicality motive).

Conclusion

An investigation of the outcomes of social relationships conducted by Baumeister and Leary (1995) found that the key to understanding the benefits of relationships can be understood in terms of a need for belongingness. Viewed in this way, relationships are more than just a way of reciprocating for financial gain or increasing chances of survival. As such, these authors observed that individuals prefer a number of close and reciprocal relationships rather than a large number of one-sided interactions. These types of needs are likely what generate the Emotion Enhancement and Coping, and Positive Practicality motives identified in the present study. Moreover, the dependent personality identified by Blatt (1974) is consistent with Baumeister’s (1995) understanding of the importance of social relationships.

The findings of the present study are consistent with the concept of social capital and the potential benefits that can occur through positive interactions with others. When used

effectively by individuals at risk for self-critical depression, Messenger can provide opportunities for emotion regulation, which is one of the psychological benefits of social capital. The results from the present study also indicate that individuals with highly dependent personality styles are likely to use Messenger for one of two motives. When used to satisfy the Positive Practicality motive, individuals who use Messenger for reasons of connecting with friends and family are likely to facilitate the development of bonding social capital. Conversely, bridging capital can result when dependent individuals use Messenger in order to pass time. In this case, dependent individuals might indirectly be accumulating social capital which is not of sufficient intensity to satisfy their need to connect with others and thus they may be at greater risk for developing a depressive episode.

Taken together, these results suggest that social capital can provide a beneficial framework for understanding the motives for Messenger use that were observed in the present study. Subsequently, these motives, within the context of social capital facilitate our understanding of why individuals with self-critical and dependent personality styles use Messenger. Thus, in order to fully appreciate the reasons why individuals use CMC tools such as Messenger, it is necessary to consider broad social motivations as predicted by social capital as well as individual-level characteristics such as the personality styles identified by Blatt (1974).

References

- Adler, P. S. & Kwon, S. W. (2002). Social capital: Prospects for a new concept. *The Academy of Management Review*, **27**(1), 17-40.
- Amiel, T. & Sargent, S. L. (2004). Individual differences in Internet usage motives. *Computers in Human Behavior*, **20**, 711-726.
- Bassani, C. (2007). Five dimensions of social capital theory as they pertain to youth studies. *Journal of Youth Studies*, **10**(1), 17-34.
- Baumeister, R. F. & Leary, M. R. (1995). The need to belong: Desire for interpersonal attachments as fundamental human motivation. *Psychological Bulletin*, **117**(3), 497-529.
- Beaudoin, C. E. & Tao, C. (2007). Benefiting from social capital in online support groups: An empirical study of cancer patients. *CyberPsychology & Behavior*, **10**(4), 587-590.
- Beck, A. T. (1973). *The Diagnosis and Management of Depression*. Philadelphia, PA: University of Pennsylvania Press.
- Best, S. J. & Krueger, B. S. (2006). Online interactions and social capital: Distinguishing between new and existing ties. *Social Science Computer Review*, **24**(4), 395-410.
- Blatt, S. J. (1974). Levels of object representation in anaclitic and introjective depression. *Psychoanalytic Study of the Child*, **29**, 107-157.
- Blatt, S. J., D'Afflitti, J. P. & Quinlan, D. M. (1976). Experiences of depression in normal young adults. *Journal of Abnormal Psychology*, **85**, 383-389.
- Blatt, S. J. & Maroudas, C. (1992). Convergences among psychoanalytic and cognitive-behavioral theories of depression. *Psychoanalytic Psychology*, **9**, 157-190.
- Blatt, S. J. & Zuroff, D. C. (1992). Interpersonal relatedness and self-definition: Two prototypes for depression. *Clinical Psychology Review*, **12**, 527-562.

- Bryant, J. A., Sanders-Jackson, A. & Smallwood, A. M. K. (2006). IMing, text messaging, and adolescent social networks. *Journal of Computer-Mediated Communication*, **11**(2), 577-592.
- Comrey, A. L. & Lee, H. B. (1992). *A first course in factor analysis* (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Francescato, D., Mebane, M., Porcelli, R., Attanasio, C. & Pulino, M. (2007). Developing professional skills and social capital through computer supported collaborative learning in university contexts. *International Journal of Human-Computer Studies*, **65**(2), 140-152.
- Gotlib, I. H. & Cane, D. B. (1989). Self-report assessment of depression and anxiety. In P.C. Kendall & D. Watson (Eds.), *Anxiety and depression: Distinctive and overlapping features* (131-169). San Diego, CA: Academic Press Inc.
- Green, M. C. (& Brock, T. C. (. (2005). Organizational membership versus informal interaction: Contributions to skills and perceptions that build social capital. *Political Psychology*, **26**(1), 1-25.
- Hampton, K. & Wellman, B. (2003). Neighbouring in netville: How the internet supports community and social capital in a wired suburb. *City & Community*, **2**(4), 277-311.
- Heinrich, L. A. & Gullone, E. (2006). The clinical significance of loneliness: A literature review. *Clinical Psychology Review*, **26**(6), 695-718.
- Horowitz, L. M. (2004). *Interpersonal foundations of psychopathology*. Washington, DC: American Psychological Association.
- House, J. S., Landis, K. R. & Umberson, D. (1988). Social relationships and health. *Science*, **241**, 540-545.
- Katz, R., Shaw, B. F., Vallis, T. M. & Kaiser, A. S. (1995). The assessment of severity and symptom patterns in depression. In E.E. Beckham & W.R. Leber (Eds.), *Handbook of Depression* (2nd ed., 61-85). New York, NY: Guilford Press.
- Kraut, R., Kiesler, S., Boneva, B., Cummings, J., Helgeson, V. & Crawford, A. (2002). Internet paradox revisited. *Journal of Social Issues*, **58**(1), 49-74.
- Kraut, R., Patterson, M., Lundmark, V., Kiesler, S., Mukophadhyay, T. & Scherlis, W. (1998). Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist*, **53**(9), 1017-1031.
- Knight, R. G., Waal-Manning, H. J. & Spears, G. F. (1983). Some norms and reliability for the State-Trait Anxiety Inventory and the Zung Self-Report Depression Scale. *British Journal of Clinical Psychology*, **22**(4), 245-249.
- Leung, L. (2001). College student motives for chatting on ICQ. *New Media & Society*, **3**(4), 483-500.
- Lewis, C. & Fabos, B. (2005). Instant messaging, literacies, and social identities. *Reading Research Quarterly*, **40**(4), 470-501.
- Lipsman, A. (2006). *Europe surpasses north america in instant messenger users, comScore study reveals*. Retrieved November/22, 2008, from <http://www.comscore.com/> press/release.asp?id=800.
- Matei, S. & Ball-Rokeach, S. J. (2001). Real and virtual social ties: Connections in the everyday lives of seven ethnic neighbourhoods. *American Behavioral Scientist*, **45**(4), 550-564.

- Mazur, M. A., Burns, R. J. & Emmers-Sommer, T. M. (2000). Perceptions of relational interdependence in online relationships: The effects of communication apprehension and introversion. *Communication Research Reports*, **17**(4), 397-406.
- Nietzel, M. T. & Harris, M. J. (1990). Relationship of dependency and achievement/autonomy to depression. *Clinical Psychology Review*, **10**, 279-297.
- Putnam, R. (2000). *Bowling alone: The collapse and revival of american community*. New York: Simon and Schuster.
- Shepherd, R. & Edelmann, R. J. (2005). Reasons for Internet use and social anxiety. *Personality and Individual Differences*, **39**, 949-958.
- Stolle, D. (1998). The development of generalized trust in voluntary organizations. *Political Psychology*, **19**(3), 497-525.
- Tabachnick, B. G. & Fidell, L. S. (2007). *Using multivariate statistics* (5th Ed.). Boston, MA: Allyn & Bacon.
- Tynes, B., Reynolds, L. & Greenfield, P. M. (2004). Adolescence, race, and ethnicity on the internet: A comparison of discourse in monitored vs. unmonitored chat rooms. *Journal of Applied Developmental Psychology. Special Issue: Developing Children, Developing Media: Research from Television to the Internet from the Children's Digital Media Center A Special Issue Dedicated to the Memory of Rodney R.Cocking*, **25**(6), 667-684.
- Wellman, B., Quan Haase, A., Witte, J. & Hampton, K. (2001). Does the internet increase, decrease, or supplement social capital? social networks, participation and community commitment. *American Behavioral Scientist*, **45**(3), 436-455.
- Williams, D. (2006). On and off the'Net: Scales for social capital in an online era. *Journal of Computer-Mediated Communication*, **11**, 593-628.
- Zung, W. W. K. (1965). A self-rating depression scale. *Archives of General Psychiatry*, **12**, 63-70.
- Zung, W. W. K. (1972). The Depression Status Inventory: An adjunct to the Self-Rating Depression Scale. *Journal of Clinical Psychology*, **28**(4), 539-543.
- Zuroff, D. C., Moskowitz, D. S., Wielgus, M. S., Powers, T. A. & Franko, D. L. (1983). Construct validation of the dependency and self-criticism scales of the Depressive Experiences Questionnaire. *Journal of Research in Personality*, **17**, 226-241.
- Zuroff, D. C., Quinlan, D. M. & Blatt, S. J. (1990). Psychometric properties of the Depressive Experiences Questionnaire in a college population. *Journal of Personality Assessment*, **55**, 65-72.

Chapter 7

FULL-TEXT SEARCH IN ELECTRONIC HEALTH RECORDS: CHALLENGES AND OPPORTUNITIES

***David A. Hanauer^{1,2,3}, Kai Zheng^{3,4,5}, Qiaozhu Mei⁵
and Sung W. Choi^{1,2}***

¹Department of Pediatrics

²Comprehensive Cancer Center

³Center for Computational Medicine and Bioinformatics

⁴School of Public Health

⁵School of Information, University of Michigan, Ann Arbor, Michigan, USA

Abstract

With the passage of the HITECH Act as part of the American Recovery and Reinvestment Act of 2009, the ubiquitous adoption of electronic health records (EHRs) in the United States is likely to occur in the next few years. However, transforming the storage media of patient data, in itself, does not guarantee desirable quality improvement and cost saving outcomes. The catalyzing effects of EHRs critically rely on the value-augmenting functionalities that could unleash the true power of electronically acquired data.

Unfortunately, while the data are electronic, most EHRs support only rudimentary search capabilities which limits the opportunities for making full use of the data. In this Commentary, we discuss some of the complex issues, and potential solutions, for designing an effective search engine for EHRs. The Commentary is based on the 4 years of experience we have had operating a search engine specifically designed for EHRs, referred to as the Electronic Medical Record Search Engine (EMERSE). We have found that concepts that work for general-purpose search engines do not necessarily apply to those used for an EHR system, namely [1] ambiguity exists with respect to how to define document ‘match’ when searching patient data; [2] documents should not be retrieved in isolation outside the context of all care episodes for the patient; and [3] ‘stop words’—minor words with little inherent meaning that are often automatically excluded by the search engine software—generally cannot be ignored since many are valid abbreviations (e.g., “AND” = axillary node dissection, “OR” = operating room). Further, medical data are subject to unique privacy restrictions and access to the data is limited based on specific user roles in the health care system.

In this Commentary, we discuss several use cases for searching data stored in EHRs as well as critical features which we have deployed that have helped to make our implementation

successful—more useful and more usable—at our medical center and beyond. Such features include a vast library of medical synonyms and abbreviations, functionalities accommodating the need to search for a phrase with only a subset of words in a case-sensitive manner (e.g. to distinguish “all” vs. “ALL” = acute lymphoblastic leukemia), and the ability to support basic negation and collaborative search.

Introduction

EHRs play a central role in transforming the U.S. healthcare system through their capability of enhancing clinician performance, streamlining wasteful processes, and improving the health of populations. [1] Further, clinicians’ day-to-day interactions with EHRs generate rich, individual-level patient data that can be utilized for secondary use purposes such as quality and safety measures, education, cost analyses, public health surveillance, and clinical and policy research. Such secondary use can greatly augment the value of EHRs through facilitating “rapid learning” in the healthcare system to fill major knowledge gaps about healthcare costs, the benefits and risks of drugs and procedures, geographic variations, environmental health influences, the health of special populations, and personalized medicine. [2, 3]

At our institution we have over 10 years of free text narrative clinical documents stored in electronic format with approximately 3 million new clinical narrative documents added every year. These clinical documents form the backbone of clinical care for patients and describe critical details including their medical and social histories, medications, diseases, treatments, etc. They often come in multiple forms including nursing, social work, and physician documentation. Among these categories there are also subcategories including admission history and physicals, progress notes, transfer notes, and discharge notes.

While the data are stored electronically, we were finding it increasingly difficult to retrieve relevant information for rapidly answering vital clinical care, research, billing, and other questions. This challenge is universal. For example, a recent study that monitored the implementation of an electronic patient record system in Norway reported that over a third of general practitioners gave up searching for patient information in the electronic system because it required too much time. [4]

Tools currently exist for extracting coded data such as laboratory and physiologic parameters (height, weight, pulse, blood pressure) [5] which are often stored in *queryable*, relational databases such as clinical data repositories or data warehouses. [6] However, providing database level access to clinicians or researchers, especially for complex free text data, is neither feasible nor advisable, because average users of medical data usually do not have the technical skills for constructing relational database queries. Hence, there has been a growing need to organize and find all of the data stored in EHRs in a timely manner that is accessible to average users, such as a free text search engine specifically designed to assist in free-text data retrieval.

Advantages of Free Text

In general, free text data are difficult to extract and use in computation due to the extreme variability in how any concept can be phrased. However, for exactly this reason free text

offers great value in supporting medical practice. The primary reasons clinicians create notes are to document care episodes and communicate the important elements to other clinicians who may be involved in a patient's care. In other words, the effort is directed at making data easy to understand for other *people* and not for other *computers*. [7]

Additionally, even though many in the medical informatics field have proposed more structured data entry, such documents would still require free text to capture the complexities of a patient along multiple dimensions. These can include not only basic medical information such as medications and diagnoses but also other factors which are vitally important to a patient including their home environment and the detailed timing of when events occurred in relation to each other. It would be risky to sacrifice all of the rich and detailed medical data often found in free text notes for the ease of use of much more limited data entered via structured data entry.

For example, a patient may come to the clinic with the new onset of a cough, fever, and night sweats. These elements may be reasonable to capture in a structured form. However, it can quickly get complicated trying to capture whether or not the patient had traveled to an area with endemic tuberculosis, the specific location and duration of travel, and the timing of the onset of symptoms in relation to the travel. All of these clues can help the practitioner include or exclude potential diagnoses. In essence, none of the contemporary medical taxonomy systems is comprehensive enough to meet all the documentation needs of clinicians.

Data Entry Methods

Among a wide spectrum of electronic health records systems available, many provide the option for structured data entry, storing clinical information as coded data rather than free text notes. Nonetheless, nearly all these systems allow free text data entry due to the necessity of describing abnormal findings, plans of care, and other reasons described above. A recent study of an EHR system that provided structured data entry found that while clinicians were using structured data entry, the concurrent use of free text actually has been increasing over time, demonstrating that there is a persistent need for the flexibility of documenting without constraints. [8]

Various approaches exist for entering clinical narratives. [9] Some EHR systems provide templates for users, allowing them to click on standardized elements that will form the text. Some have pre-populated fields that can import laboratory values into the document without having to transcribe them from another screen. At our institution, there are two primary methods for creating free text documents: dictation/transcription and typing. Among free text notes available in the system, about 43% were created via dictation/transcription with the remainder entered directly via a keyboard.

The advantage of dictation/transcription is that it is fast and requires less effort, as a note can be created as quickly as a clinician can record his or her voice, with typing and the majority of editing performed by professional transcriptionists. The disadvantages, other than costs, include various typographic errors that can end up in the medical record, likely a result of unclear speech by the dictating clinician and a resultant difficulty on the part of the transcriptionist in determining the precise spoken phrases. Such errors can lead to transcription blunders such as "nasopharynx" instead of "nasal flaring" or "albumin" instead

of “albuterol.” These errors can conceivably create problems for later document retrieval and reuse unless the source documents are corrected by the clinician; however, many clinicians do not spend the time to carefully identify and correct mistakes and these mistakes often persist in the medical record.

Data Mining

Interest in mining medical records for valuable information has been growing. Such initiatives primarily focus on the coded data, although the details of a patient’s medical history may only be partially (or minimally) captured in that form. For free text notes, natural language processing (NLP) is the favored choice among academic informaticians. [10] Such techniques involve complex algorithms to identify diagnoses or other treatments, taking into account complexities such as negation and ambiguity. It is a worthwhile technique when a very specific parameter of interest is desired such as smoking status in patients. [11] NLP represents a powerful approach for those who have the right skill set and the time to effectively develop the code to apply to a specific task. Additionally, NLP will certainly have a place for solving tasks at a very large scale such as screening all patients in a medical center (or nationwide) for a potential complication or identifying patients for a medical device recall.

NLP, however, is only useful when the information retrieval task can be clearly defined. In reality, users often want to answer poorly defined or very complex clinical questions which can be a challenging task even for human reviewers (e.g., “Was the pain control adequate?” or “Did the clinician provide smoking cessation counseling to the patient?”). Further, to achieve high sensitivity and specificity, considerable implementation and tuning efforts are needed which often requires sophisticated technical expertise. For average users, it is neither a scalable nor a generalizable solution, because they don’t have the time, resources, or expertise to implement an NLP solution for every question that needs answering.

Search Engines for the Medical Record

Given the amount of clinical research requiring medical chart reviews and care situations in which finding a specific element would lead to a rapid answer to a clinical question, it is surprising that little effort has been put forth into developing robust search engines to facilitate information retrieval in EHRs. Barriers include the need for specific medical domain knowledge, the difficulty for non-medical personnel to obtain access to the underlying database containing sensitive and protected health information, and the challenge of how to translate existing technology to an area where the normal search paradigm may be different from standard, general-purpose search engines.

Nevertheless, researchers and clinicians are searching the medical record, but they are doing so in a highly inefficient manner. Many who undertake what is known as “chart review” or “chart abstraction” simply go through each document manually, reading through each clinical note to find the information they are looking for. Sometimes the burdensome search yields no results, either because the information was not there at all or because it was overlooked. This may explain why there has been evidence suggesting that users will stop searching in medical records if it requires too much time. [4]

There has been very little published research regarding the use of search engines applied to medical record systems. [12-14] Part of this may be due to the privacy and security barriers surrounding EHRs, making it difficult to implement and evaluate systems that provide additional functionality. At the University of Michigan, we have developed a search engine for the medical record to address the varied needs of researchers, clinicians, and others for finding information quickly and accurately in the medical record. Our implementation is known as EMERSE – Electronic Medical Record Search Engine. [15] In developing this search engine, we have come across, and surmounted, various obstacles that illustrate the challenges of implementing such a tool in the medical domain.

Hipaa and Privacy Regulations

The Health Insurance Portability and Accountability Act (HIPAA) was instituted over a decade ago. It was designed primarily with health insurance protection for patients in mind but also included important legal mandates (and penalties if violated) for protecting the privacy of patient data. HIPAA, along with other privacy regulations, including locally instituted rules, limits the accessibility of protected health information (PHI) to only a small subset of users, and has important implications for those wishing to use the medical record for research. [16] Local institutional rules, for example, restrict access to sensitive psychiatry documents to only those who truly need to see them, so that even some physicians would not have access by default.

At our institution, access to the medical record, and to EMERSE, is strictly controlled. Both the EHR system and EMERSE must maintain an audit trail to determine which users have accessed which patient's records. Furthermore, if the work is being done for research, approval must be granted by the local institutional review board (IRB) and human subjects training must be completed by each user. This adds complexity to managing users and access and ensuring that patient privacy is maintained.

Significant challenges for protecting privacy exist with the way most search engines work. If an index of all patient documents were to be created, it is possible that a simple search could quickly turn up sensitive records for patients who did not want their identities to be disclosed. The documents themselves often contain identifying information as many physicians, nurses, and social workers document the social environment in which the patient lives. This could include history of family abuse, bitter divorces, or other sensitive information, and often is specific about names, places, and dates. All are considered PHI that could identify a patient and HIPAA mandates their removal under many circumstances in which patient data will be used for non-clinical care activities.

With EMERSE, we have bypassed one of the security challenges by requiring that users specify the patients they want to search. That may seem counterintuitive for how a search engine should work, but in most cases, the patients cohort has already been identified in some manner and the real task is to identify and abstract certain elements for each patient. Initial patient cohorts are sometimes identified by using international classification of disease codes (ICD-9) that most medical establishments use for billing purposes. Such codes can be inaccurate [17] so further examination in the search engine allows for better classification of patients.

Typical Users and Use Cases for Searching Electronic Patient Care Data

When considering who would use a search engine, one might initially think that physicians and nurses would be the primary users. However, we have found that to be only partly true. The use cases, and thus users, are wide and varied. We have conducted surveys of our users to determine the profiles of our user base. Nearly all medical disciplines are represented in the group that responded to the survey, ranging from primary care fields such as general pediatrics, family medicine, and general internal medicine to subspecialties including neurosurgery, urology, and gastroenterology. Among them, only 7% are full-time practicing clinicians. The majority of the users are cancer registrars, infection control specialists, quality assurance officers, pharmacists, data managers, and other research-oriented individuals who are sometimes medical fellows, residents, students and even undergraduates working on research projects within our health system.

When we asked users what they were using the search engine for, the answers were also quite variable. Two-thirds of the users responded that they used the search engine to determine medication use for patients and nearly as many reported using it for assisting with clinical trials. Other uses included detection of adverse events, determining eligibility for clinical studies, infection surveillance, internal quality assurance projects, study feasibility determination, billing/claims abstraction, and risk management review. For comparison, researchers at Columbia University recently implemented an EHR search engine. They found that searches for “Laboratory or Test Results” and “Disease or Syndromes” constituted the majority of the recorded usage. [13]

Distinctive Features of Emerse

No Ranking of Search Results

When searching the medical record, we argue that the concept of a search result, or ‘hit,’ is different than standard search engines. Often there is no ‘best document’ or ‘top hit’ to rank and display. Rather, the final ‘answer’ (which depends heavily on the initial question) is almost always determined by viewing a collection of documents for a patient and not a single document. This is largely due to the uncertainty and ambiguity that is inherent in medical encounters. An example of this includes a concept as simple as ‘asthma’. Diagnosing a young child with asthma can be difficult. On an initial clinic visit for a child with difficulty breathing, the clinical narrative may mention terms such as wheezing, coughing, reactive airway disease and, perhaps, asthma. However, this does not mean that the child actually has asthma, because many young children will develop wheezing during an upper respiratory viral infection. It is only through observing documented recurrent episodes that one might conclude that a child truly has asthma, also taking into account the medications that have been prescribed and the changes in the child’s condition based on use of those medications. Therefore, a single document mentioning asthma, regardless of how frequently mentioned in the document, would not truly provide a confident diagnosis. The same is also true for many other diagnoses.

Because of this common problem, we have chosen to show all hits for a patient throughout the medical record. Results are grouped by patient so that all hits for a single patient can be seen together, by default in chronological order. This grouping allows users to identify potentially discordant information to ultimately determine the correct way to extract the concept. A single patient may have some documents mentioning “history of smoking” and others mentioning “no history of smoking”. Careful review of these documents can help lead to the proper conclusions. Furthermore, documents for a specific patient that *don’t* have hits are also presented to the user, interleaved in chronological order with those that do have hits, as this helps the user review all of the potentially relevant clinical data.

Stop Words and Synonyms

Many search engines include the concept of stop words, which are common but minor words that often do not provide much value to search and are therefore excluded from the query. Such words are not ignored in our search engine since many small words are also medical abbreviations. A short list of examples include ‘AND’ (axillary node dissection), ‘OR’ (operating room), ‘IS’ (incentive spirometry), ‘ARE’ (active resistance exercise), ‘IT’ (intrathecal), and ‘ALL’ (acute lymphoblastic leukemia).

Furthermore, we have found it necessary to provide common synonyms and abbreviations used in the medical narratives since users often find it challenging to think of the appropriate terms (also see section on ‘Collaborative Search Features’). Our system contains a large list of these synonyms and abbreviations to help prompt the users with additional choices with which to search. If a user were to type in “operating room” it would suggest the abbreviation “OR” or if a user typed in the chemotherapy “CHOP” it would suggest a list of the 4 individual chemotherapeutic agents that CHOP represents (cyclophosphamide, hydroxydaunomycin, Oncovin, and prednisone) to allow the user to search on the names of the individual drugs. For medications, a large list of generic and trade names are included since both are commonly referred to interchangeably in the clinical documents, especially for commonly used drugs such as Motrin (trade name), Advil (trade name), and ibuprofen (generic name). These suggestions are merely provided to the user who can then choose to include them or disregard them—they are not automatically added to the list of search terms. There were multiple reasons for this, but a primary reason was due to the ambiguity of many abbreviations. For example, if a user were to type in “ARF” as an abbreviation to search for, the system would provide the following suggestions: “acute renal failure”, “acute respiratory failure”, and “acute rheumatic fever”. It would be up to the user to determine which of these, if any, were the appropriate terms.

Modifier Codes

Because stops words are included in the search, and due to a need for distinguishing common words from medical abbreviations, various modifier codes are also available to users of the search engine. By default, terms or phrases separated by a space are searched using the *or* modifier. Nested Boolean searches are not supported, but basic *and* conjunctions are permitted by adding the ‘+’ symbol in front of search terms.

We have found it important to enhance the likelihood of a search hit as much as possible by performing our searches by default in a case-insensitive manner and by highlighting any subset of text that matches. For example, when searching for “cervical ca” (“ca” is a common abbreviation for cancer) we would also want to highlight the phrases “cervical cancer” as well as “cervical carcinoma”. This inclusive approach can sometimes result in false positive results (e.g., “cervical cap” would also be a hit) and we therefore also allow the option for users to force certain terms to be distinct words using the ‘~’ symbol. Thus, searching for “hiv” would highlight the “hiv” in the term “archived” but searching for “~hiv” would only highlight the distinct word “HIV” (in a case-insensitive manner by default) which would represent “human immunodeficiency virus”.

Some of the terms that users seek are so common that we have also found the need to allow for searches to be performed in a case sensitive manner. Thus, “all” would highlight too many non-specific terms, but specifying a case-sensitive search using ‘^’ such as “^ALL” would allow the user to narrow down the hits to words that more likely represent “acute lymphoblastic leukemia” rather than the common English word. Other words and phrases must be ignored first in order to reduce the false positive rate. We support a simple form of negation by allowing users to include a “-“ in front of words or phrases. Rather than ignoring documents with those terms the minus sign simply mean to ignore those specific terms before searching. This provides the capability for users to search for “wound infection” but ignore phrases in the text that might state “no evidence of a wound infection”.

A few searches cannot be satisfied without the support of regular expressions, and our tool does support the use of these. Once, a physician was trying to identify cases of “myocardial infarction”, abbreviated “MI”. This was indistinguishable from the state abbreviation for Michigan, found in many of our clinical documents. We found a simple yet elegant solution by using regular expressions:

```
-$"MI\s*\d{5}"; -$"MI\s*,\s*\d{5}"; ~^MI
```

This set of terms above can be interpreted as “first ignore any MI followed by a 5 digit ZIP code and then search for any remaining upper-case distinct MI words”. It should be noted, however, that the typical users do not use nor understand regular expressions.

Presentation of Search Results

In order to make the presentation of the data as intuitive and straightforward as possible, three levels of granularity are available for users to review the ‘hits’. The top level view is what we call the clinical ‘heat map’ (Figure 1A), loosely modeled after the heat maps used in genomic analyses. At this level, every row in the map represents a patient and each column represents a type of document as they are organized in the medical record. These columns include diagnoses, procedures, common medical reports (including progress notes, discharge summarize, operative reports, etc), radiology reports, pathology reports, and others. The organization helps users focus in on the areas of greatest interest to their task. Cells in this grid view are color coded; darker colors represent an increasing number of documents that contain a hit.

A

Num	Name	CPI	PSL diagnoses	PSL procedures	Documents	Pathology	RadNuc	Other Results
1	[REDACTED]				[13/68]	[1/665]	[15/147]	
2	[REDACTED]		[3/28]		[20/146]		[1/23]	
3	[REDACTED]			[1/68]	[23/105]		[9/43]	
4	[REDACTED]				[51/233]		[10/56]	
5	[REDACTED]				[12/69]		[10/138]	
6	[REDACTED]			[1/48]	[1/2]	[1/10]	[4/4]	

B

Num	Date	Doc Type	Svc	Dept Summary
99		CONSULT- OUT	FPSY	PED
00		LETTER - RV	PHEM	PED
97		NUTRITION NOTE	FEDEHM	CCC
96		CONSULT- OUT	NPSYMI	PSY
05		LETTER - RV	FEDEHM	CCC
94		SOCWK/OP ASSESS	PHEM	PED
93		NUTRITION NOTE	FEDEHM	CCC
92		LETTER - RV	NEUS	NEUS
01		PREOP H&P	PAS	ANES
90		LETTER - RV	PHEM	PED
99		NOTE - RV	ORTS	ORTS
88		NOTE - NP	O&P	PMR
87		LETTER - NP	ORTS	ORTS
96		ED NOTE	PER	ER
85		LETTER - RV	PHEM	PED
84		LETTER - RV	PHEM	PED

C

I had the pleasure of seeing [REDACTED] today in the Vascular Surgery Clinic here at the University of Michigan. Ms. [REDACTED] is a very pleasant [REDACTED]-year-old female who is status post placement of an endovascular aortic cuff, AneurX cuff for an aortobronchial fistula. This was done through a right flank approach. She has no new complaints. She had a CT scan which showed no change in the appearance of the descending aortic stent graft.

She is on no present medications.

On physical exam, she is well appearing. Her right flank incision is well healed. She has easily palpable groin pulses.

Overall, [REDACTED] is doing extremely well over a year after endovascular treatment of her aortobronchial fistula. I would like her to get a chest x-ray today. I will see her back in approximately 1 year with a CT scan of her chest as well as a repeat PA and lateral chest x-ray.

Thank you allowing us to participate in her care.

Sincerely,

[REDACTED], M.D.

Figure 1. Three levels of granularity for viewing the search results. In the top view (A) results for all patients are shown in rows with all document types in columns. Clicking on a specific cell shows a timeline view (B) of all documents for a specific category for a single user. Hits are highlighted with snippets of surrounding words for context. Clicking on a row displays the original document (C) with all hits highlighted. Identifiable health data have been removed from these screen shots.

If a particular cell in this map is of interest to the users, he or she can click on the cell to bring up a more granular, patient-centric view of the data. At this level the user is no longer looking at all hits for all patients but instead gets a view of all documents of a particular category for that patient in chronological order (Figure 1B). The documents with a hit have snippets of text showing a summary of the hits and those documents without a hit have blanks in the summary section. This allows for rapid review of the documents so that the reviewer can determine if any specific documents are worth reading.

When a specific document is of interest, the user simply clicks on the appropriate row which will bring up the single document view with all search terms highlighted and color-coded (Figure 1C). Thus, in three simple clicks a user can go from a high-level view of all patients and all documents to a very narrow view of a single document for a specific patient.

Collaborative Search Features

The complexity and variability of medical terminology, and that of medicine in general, means that no individual can be completely well-versed in performing searches outside of their domain of medical expertise. Furthermore, since many of the users of the search engine do not have formal or standardized medical training (e.g. medical students still in training), it is important to ensure that they have the greatest probability of success when conducting a search. In order to facilitate and enhance the ability of users to complete a successful search we added into our search engine a collaborative search feature known as “Search Bundles”.

Search Bundles are collections of search terms and phrases that can be of any size and are built by the users. Each Bundle has a name and a description of what it was designed to find. To leverage the collective wisdom of the users and to foster transfer of knowledge across medical and health domains, we built into the Bundles the means with which users can share their Bundles with their colleagues, either with a specific group of users (privately shared Bundles) or with all users on the system (public Bundles).

We assessed the utility and acceptance of this feature by analyzing nearly 4 years of data in the system’s search logs which recorded detailed usage data (also discussed in the following section on System Usage). Network analysis techniques were used to examine the relationships among users, their home clinical departments, and the Bundles that each created, shared, or used.

We found that the Bundle sharing feature did support collaboration among disparate clinic departments, resulting in knowledge transfer between these departments, as shown in the two network graphs in Figure 2. While a few groups were not heavily involved in either sharing or using the bundles created by others, most departments embraced the feature and used it in their work.

Over 900 Search Bundles were created by users and about 40% of these were shared with specific users (private Bundles) and another 16% were made available to all users (public Bundles). Nearly half of all searches utilized Search Bundles and nearly 40% of these were Bundles that had been created and shared by other users. The largest Bundle contained 241 distinct terms or phrases and the average Bundle size was 20 terms or phrases.

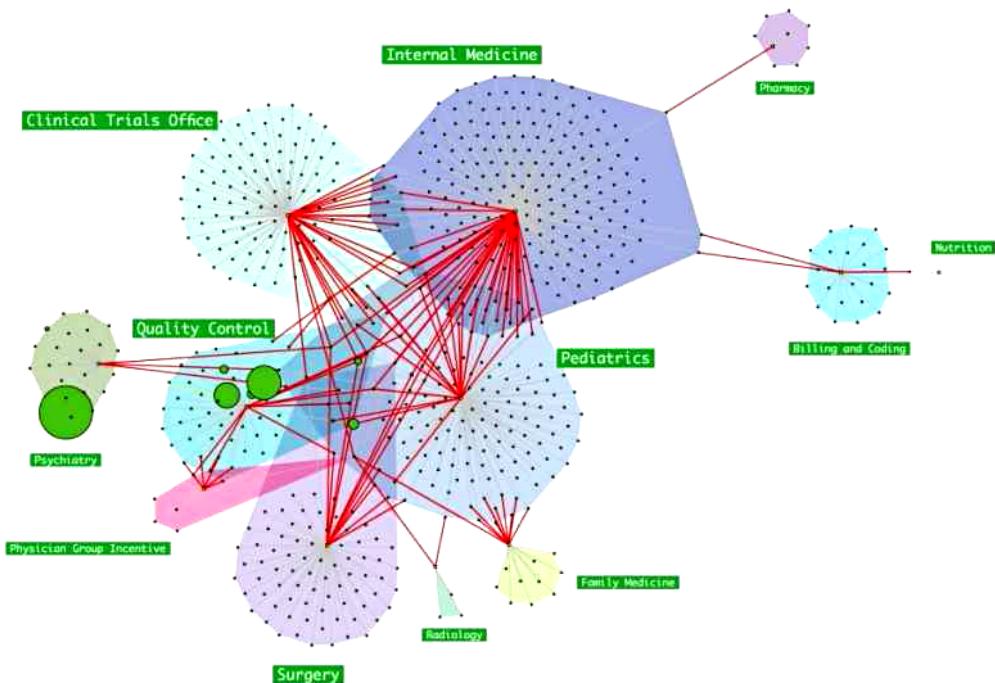


Figure 2. (A). Network plot showing collaboration among departments with shared Search Bundles. Each small circle is a Bundle. Colored convex hulls represent all bundles created by users in a single clinical department. Red edges connect shared Bundles to the department that created it and Gray edges connect a Bundle to its owner department.

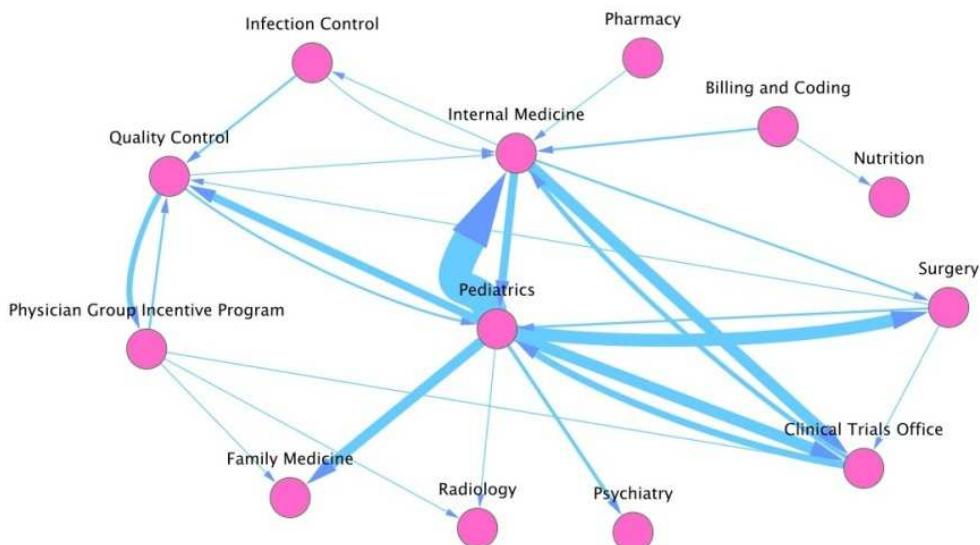


Figure 2 (B). A simplified network graph showing knowledge transfer between departments. Nodes represent clinical departments. Edges represent knowledge transferred between departments through shared Bundles. Edge thickness is proportional to the number of Bundles shared between two departments and arrows point to the direction of knowledge flow.

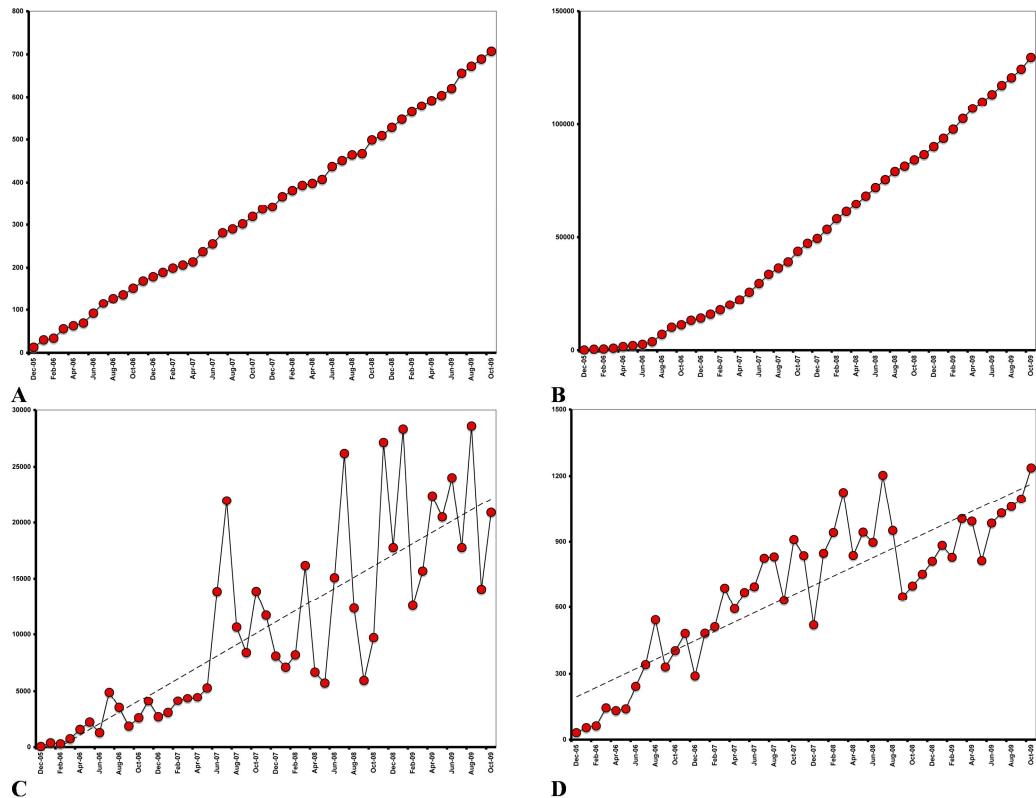


Figure 3. Various usage measurements derived from the audit logs maintained by EMERSE from December 2005 through October 2009 including (A) cumulative number of registered users; (B) cumulative size of the search term log, now with nearly 150,000 rows of search query data; (C) total number of system logins per month; and (D) total number of unique patients searched per month. A trend line has been added to the graphs in panels C and D for clarity.

System Usage

Our search engine system records detailed usage data in order to be compliant with HIPAA and other compliance regulations. From these logs we are able to report various usage statistics about the system since its inception in December 2005. The user base, defined as the cumulative number of registered users, has grown every month since that time (Figure 3A). The cumulative number of rows in the search log is also growing substantially (Figure 3B). Additionally, the trend has been increasing for total number of monthly logins (Figure 3C) and for the number of patients searched per month (Figure 3D).

Evaluation

Due to the novelty of applying search engine principles to the medical record, we have been very interested in evaluating how well the system has performed for users both through surveys and through empirical evaluations of the tool.

We have conducted two separate surveys of users. The first survey conducted in January 2008 had 78 responses. Among the respondents, users gave an overall rating to the system of 4.8 out of 5.0 (1 = “terrible”, 5 = “great”). They estimated an overall time savings of 143 minutes per day when using the search engine compared to prior methods for finding data. A second survey of prior and current EMERSE users conducted in March 2009 resulted in 305 responses. Among this group of respondents, when asked how important EMERSE was for their work compared to other software research tools, the average response was 8.1 (0 = “trivially important”, 10 = “critically important”). Average estimated time savings per day was 149 minutes, remarkably similar to the results from the prior survey.

Several groups have also conducted studies of the tool to determine how well the tool worked for real tasks. One recent psychiatry analysis compared using EMERSE to identify patients eligible for depression research compared to manual chart abstraction, considered the gold standard by which any tool should be compared. Conclusions from the study were that using EMERSE resulted in significant time savings (ranging from a 38 to 82% reduction in time depending on the abstractor) with no significant decrease in accuracy. [18]

Another study was conducted by our hospital’s infection surveillance team. This group’s charge is to monitor infections in the hospital to detect trends and promote practices that result in decreased infections with resulting decreases in morbidity and mortality. One of their tasks involves reviewing microbiology reports for infections in specific groups including post-surgical patients. They recently studied using EMERSE for identifying surgical site infections in ambulatory surgery centers. They reported a 91% reduction in workload with no loss of accuracy in identifying the appropriate infections.[19)

Conclusion

Electronic health records are continuously being adopted to improve care, but little attention has been paid to enhancing its information retrieval capability particularly for finding important information from free text patient care documents. The search engine we developed for medical and health records has been widely embraced and used at our institution with high satisfaction among users. Security is essential for such a system and so are other features that we have provided including synonym suggestions. The complexity of the medical lexicon has forced us to avoid ignoring simple stop words because they can easily be confused for true clinical terms. We have also found it necessary to provide collaborative features to enhance the sharing and transfer of complex search knowledge to other domains and to those with less expertise, and we have found that users have embraced the concept. While much more work needs to be done to enhance the technology, we believe that we have taken steps forward to open the vast stores of data to those needing access for a wide variety of clinical, research, and operational purposes. Building upon our prior success, we will continue to enhance our tool based on the unique features required for effective and efficient information retrieval within EHR systems.

References

- [1] Blumenthal, D. Stimulating the adoption of health information technology. *N Engl J Med.*, 2009, Apr 9, 360(15), 1477-9.
- [2] Etheredge, LM. A rapid-learning health system. *Health Aff.* (Millwood). 2007, Mar-Apr; 26(2), w107-18.
- [3] Safran, C; Bloomrosen, M; Hammond, WE; Labkoff, S; Markel-Fox, S; Tang, PC; et al. Toward a national framework for the secondary use of health data: an American Medical Informatics Association White Paper. *J Am Med Inform Assoc.*, 2007, Jan-Feb, 14(1), 1-9.
- [4] Christensen, T; Grimsmo, A. Instant availability of patient records, but diminished availability of patient information: a multi-method study of GP's use of electronic patient records. *BMC Med Inform Decis Mak.*, 2008, 8, 12.
- [5] Deshmukh, VG; Meystre, SM; Mitchell, JA. Evaluating the informatics for integrating biology and the bedside system for clinical research. *BMC Med Res Methodol.*, 2009, 9, 70.
- [6] Dewitt, JG; Hampton, PM. Development of a data warehouse at an academic health system: knowing a place for the first time. *Acad Med.*, 2005, Nov, 80(11), 1019-25.
- [7] van Ginneken, AM. The physician's flexible narrative. *Methods Inf Med.*, 1996, Jun;35(2), 98-100.
- [8] Madani, S SE. An Analysis of Free Text Entry within a Structured Data Entry System. *AMIA Annu Symp Proc.*, 2009, 947.
- [9] Johnson, SB; Bakken, S; Dine, D; Hyun, S; Mendonca, E; Morrison, F; et al. An electronic health record based on structured narrative. *J Am Med Inform Assoc.*, 2008, Jan-Feb, 15(1), 54-64.
- [10] Meystre, SM; Savova, GK; Kipper-Schuler, KC; Hurdle, JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform.*, 2008, 128-44.
- [11] Uzuner, O; Goldstein, I; Luo, Y; Kohane, I. Identifying patient smoking status from medical discharge records. *J Am Med Inform Assoc.*, 2008, Jan-Feb;15(1), 14-24.
- [12] Gregg, W; Jirjis, J; Lorenzi, NM; Giuse, D. StarTracker: an integrated, web-based clinical search engine. *AMIA Annu Symp Proc.*, 2003, 855.
- [13] Natarajan, KSD; Jain, S; Elhadad, N. CISearch: What Do Clinicians Search for within the EHR? *AMIA Annu Symp Proc.*, 2009, 473.
- [14] Schulz, S; Daumke, P; Fischer, P; Muller, M. Evaluation of a document search engine in a clinical department system. *AMIA Annu Symp Proc.*, 2008, 647-51.
- [15] Hanauer, DA. EMERSE: The Electronic Medical Record Search Engine. *AMIA Annu Symp Proc.*, 2006, 941.
- [16] NIH. National Institutes of Health. Clinical Research and the HIPAA Privacy Rule. NIH Publication Number 04-5495. 2004, [updated 2004; cited 2009 November 27, 2009]; Available from: http://privacyruleandresearch.nih.gov/pdf/clin_research.pdf.
- [17] Rhodes, ET; Laffel, LM; Gonzalez, TV; Ludwig, DS. Accuracy of administrative coding for type 2 diabetes in children, adolescents, and young adults. *Diabetes Care.*, 2007, Jan; 30(1), 141-3.

- [18] Seyfried, L; Hanauer, DA; Nease, D; Albeiruti, R; Kavanagh, J; Kales, HC. Enhanced identification of eligibility for depression research using an electronic medical record search engine. *Int J Med Inform.*, 2009, Dec; 78(12), e13-8.
- [19] Berger, JRS; Jackson, JA; Hanauer, DA; Petersen, K; Chenowrth, CE. editor. *Active Surgical Site Infection Surveillance in Ambulatory Surgery Centers Using Electronic Medical Record Search Engine*. 19th Annual Society for Healthcare Epidemiology of America (SHEA) Meeting; 2009, March 19-22, 2009, San Diego, CA.

Chapter 8

INSTANT MESSAGING: STANDARDS, PROTOCOLS, APPLICATIONS, AND RESEARCH DIRECTIONS

Bazara I.A. Barry^a and Fatma M. Tom^b

Mathematical Sciences and Information Technology Research Unit (MITRU),
Faculty of Mathematical Sciences – University of Khartoum - Sudan

Abstract

Instant messaging has brought an effective and efficient real-time, text-based communication to the Internet community. In addition, most instant messaging applications provide extra functions such as file transfer, contact lists, and the ability to have simultaneous conversations, which strengthens the reliance of wider sectors of users on these applications. In this chapter we explore the various attempts to create a unified standard for instant messaging. We show the efforts of organizations such as the Internet Engineering Task Force (IETF) in this regard, in addition to some proprietary solutions. We also shed some light on the different types of protocols that are used to implement instant messaging applications. Furthermore, the practical uses of instant messaging are highlighted alongside the benefits that will be reaped by organizations adopting the technology. We dedicate some parts of this chapter to review current and future research in the field. Various research trends and directions are discussed to show the impact of instant messaging on users, businesses and the decision making process. This chapter provides an attempt to strengthen the theoretical background behind instant messaging and presents the topic in a systematic way.

Keyword: Instant Messaging, XMPP, SIMPLE, CMC, SIP.

1. Introduction

Today, Instant Messaging (IM) is one of the most important Internet applications, and people are using it for personal, social, educational and business reasons. Instant messaging is a method of communication that enables users to share digitally-based information such as

^a E-mail address: baazobarry@hotmail.com.

^b E-mail address: fatuma_76@hotmail.com

text, audio, and video instantly with each other and monitor the availability of a list of users in realtime over a network of computers, such as the Internet. It is considered as a type of Computer Mediated Communication (CMC) [1][2] which defines any communicative transaction that occurs through the use of two or more networked computers. Starting as a casual application, mainly used by teenagers and college students, IM systems now connect Wall Street firms and Navy warships [1].

IM predates the Internet. Its systems have been around since the applications *talk* and *write* which were used for live text communication between different users of a single multi-user computer running the Unix operating system. IM usage increased with the early implementations of the Massachusetts Institute of Technology (MIT) Project Athena Zephyr notification system, and Internet Relay Chat (IRC) which was started at University of Oulu in Finland [7]. Before the Internet became popular, ordinary people could connect and communicate with each other using online services such as America Online (AOL) and CompuServe.

In the 1980s, local area networks (LANs) became popular. Administrators of LANs (or sometimes even users) could broadcast short messages to other users on the same network instantly. In the 1990s, web-based chat rooms started to provide general users instant communication capabilities. In a chat room, a group of people can type in messages that are seen by everyone in the room. Instant messaging combines the capability of e-mail and chat rooms. It allows realtime communication like chat rooms, while maintaining the personal atmosphere and privacy of e-mail. Instant messaging was introduced to Internet users in 1996 [4] with the introduction of ICQ, a free instant-messaging utility that could be used by a wide range of Internet users. Recently, extremely high prevalence of the Internet led IM to enormous rise in popularity.

Across the business organizations, the benefits of the IM technology and realtime communication have helped IM to evolve from a tool used by normal users into communication tool among the employees in their organizations and an improving tool for the decision making process.

A fundamental issue faced by organizations is designing standardized IM protocols and architecture to scale with large number of concurrent users.

A survey report from the Radicati Group suggests that 85% of businesses use public IM services but only 12% use security-enhanced enterprise IM services and IM-specific policies [1]. For the importance of security in business and decision making organizations, adequate security mechanisms that are capable of securing communications of IM from the potential threats must be ensured.

This chapter is organized as follows. Section 2 explores the fundamentals of instant messaging and its architectures. Section 3 sheds some light on instant messaging protocols and their classifications. Section 4 addresses the issue of providing secure instant messaging services. Section 5 lists some important areas where instant messaging can be used and relied upon. Section 6 highlights some of the conducted research in instant messaging. Section 7 concludes the chapter.

2. Instant Messaging Fundamentals

In this section we explore some of the underlying functionalities and mechanisms to show how instant messaging systems work. We start with the path that is taken by an instant message from its sender to its recipient and the components involved in delivering instant messages. Then we discuss the various architectures that differentiate between instant messaging systems.

2.1. How Instant Messaging Works

Usually, instant messaging users install a software application on their personal computers to work as a client to IM servers, which are software programs that enable IM clients to access IM features in an IM service. The IM client helps users to register their unique usernames and passwords and use these credentials to connect to servers in order to receive various services.

Once the client is logged on using the right credentials, it sends its connection information such as its IP address and port number to the IM server. The IM server creates a temporary file that contains the client's connection information and the list of the client contacts and checks to see if any of the users in the client contact list are currently logged on. If the server finds any of the client contacts logged on, it sends the connection information of that user back to the client. The server also sends the client connection information to the logged on users on the client contact list. Such connection information allows IM clients to deliver messages to the intended machines either directly or through servers. Figure 1 shows a standard instant messaging communications model that involves a server and clients.

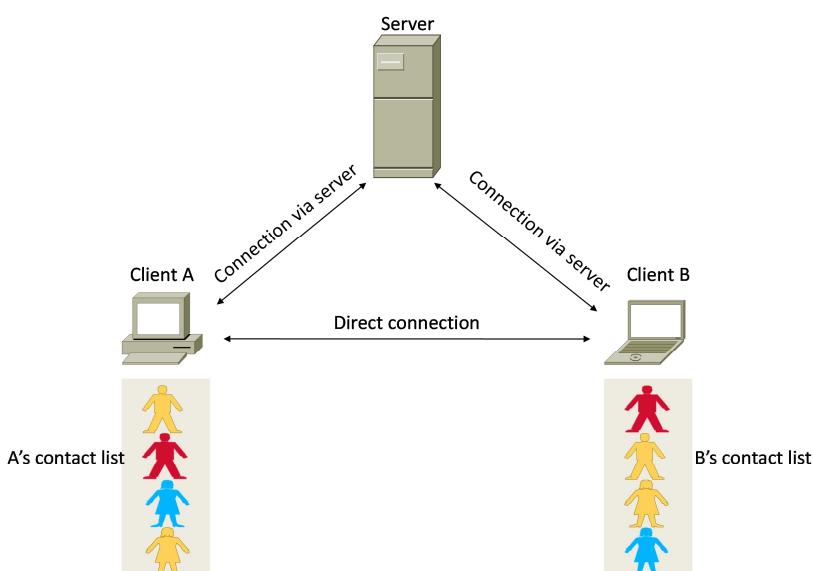


Figure 1. Standard instant messaging communications model.

For instant messaging over the Internet, an important role is played by Internet Service Providers (ISPs). ISPs deliver client connection information to the appropriate IM servers. They also deliver user messages between clients in a session alongside other data.

In addition to sending and receiving instant messages, most IM clients support other features such as creating chat rooms with selected contacts, sharing images, videos, files, and links to favorite websites with other users, and sending and receiving messages on mobile devices such as cell phones. Moreover, IM clients can provide information regarding users' presence (whether or not a user is logged on to an IM server) and users' availability (whether or not a user is willing to send or receive messages). Consequently, IM clients allow for the creation of block lists (lists of user IDs explicitly barred from getting the current user's presence and availability information) and allow lists (list of user IDs allowed to send messages to the current user and which can track the user's presence and availability information). Figure 2 and Figure 3 show some IM clients with varying features and capabilities.



Figure 2. Write IM application.

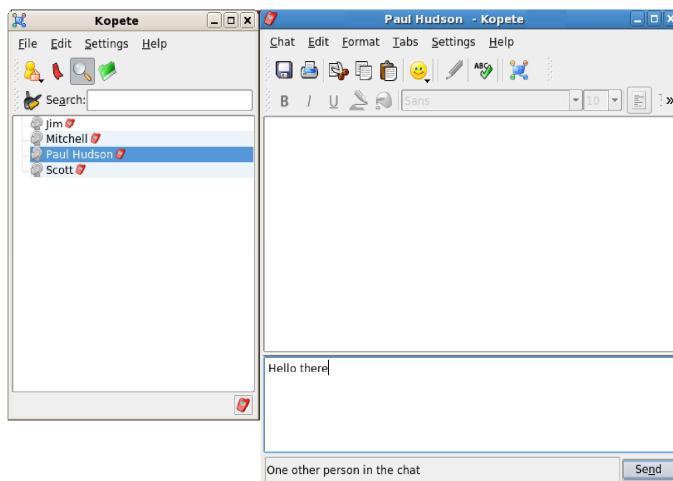


Figure 3. Kopete IM application.

2.2. Architectures of Instant Messaging Systems

Most IM systems use a *client-server* architecture to send and receive messages and files. The client-server architecture is the architecture used for normal IM operations. In this scheme, messages among users are typically relayed through a server. While an IM server appears to be a single entity to a client, it may be a group of servers controlled by a single IM service provider (e.g. AOL), or a collection of servers from independent IM service providers. If user A wants to communicate with user B, both must log into the same IM service. Messages from A to B will be delivered by the server depending on B's privacy settings [7]. Figure 4 shows a typical client-server architecture.

In the client-server architecture, two approaches are available: symmetric and asymmetric. In a symmetric architecture, each server performs identical functions, such that a client need not distinguish which server it contacts to engage in an activity with [5].

In an asymmetric approach, each server is dedicated to a particular activity such as logging in, discovering other users on the network, maintaining a chat room, or forwarding an instant message [5]. The main factor that determines which of these two approaches to choose is the scalability of the IM system with growing numbers of users.

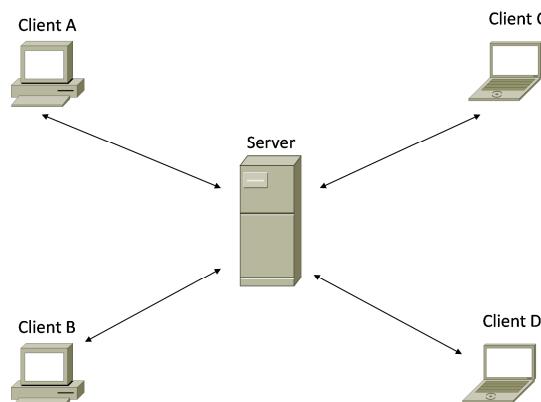


Figure 4. Client-server architecture.

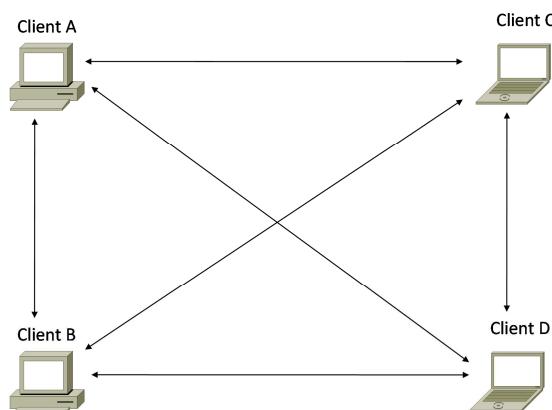


Figure 5. Peer-to-peer architecture.

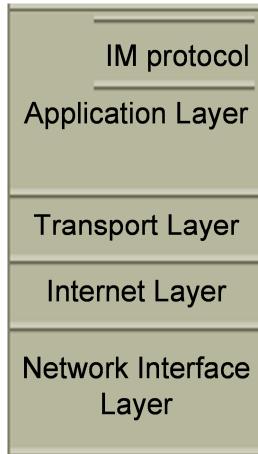


Figure 6. IM protocol in a TCP/IP reference model.

On the other hand, some IM systems use *peer-to-peer* architecture mostly for non-text sessions such as voice and video sessions. Unlike client-server, peer-to-peer end users can transfer files or video messages to one another directly without the aid of centralized server(s), as shown in Figure 5. Since there are no file storage requirements and only the users' bandwidth is needed for file transfers it is a cost-effective, scalable option.

3. Instant Messaging Protocols

IM protocols appear in the application layer in the TCP/IP model as shown in Figure 6, and are used by instant messaging clients and servers to communicate with each other over the network. IM clients use protocols to take full advantages of the IM services provided by servers. Depending on the nature of the IM application and the transferred data, some IM protocols connect directly to a server over a Transmission Control Protocol (TCP) connection, whereas others use User Datagram Protocol (UDP).

To develop IM service, the developer is required to select which protocol to use. In this section we explore some IM open standard and proprietary protocols.

3.1. Open Standard Protocols

A protocol is defined as open standard if its standards, rights of use, and design details among others are publicly available. The availability of open standard protocols allows developers who have a domain name and a suitable Internet connection to start up their own IM server, and even give away free accounts to be used by people with IM client applications. Unfortunately, most Applications of presence and instant messaging currently use independent, non-standard and non-interoperable protocols developed by various vendors. [8].

The Internet Engineering Task Force (IETF) which is the protocol engineering and development arm of the Internet, carry out the standardization of Internet protocols by its various working groups. One of these working groups, which is the Instant Messaging and

Presence Protocol (IMPP) working group, defines a standard protocol so that independently developed applications of instant messaging can interoperate across the Internet [8]. It defines the minimal set of requirements of namespace and administration, scalability, access control, network topology, message format, reliability, performance and security considerations that IM protocol must meet.

Furthermore, with the growing need for interworking between diverse instant messaging protocols, the IETF published Common Profile for Instant Messaging (CPIM) specification [9] that defines common semantics and data formats for IM that meet the requirements specified by the IMPP to facilitate the creation of gateways between instant messaging services to allow the interoperation between wide ranges of IM systems. The CPIM gateways can relay common payloads that are carried by diverse instant messaging protocols.

In the following we shed some light on two open standard IM protocols, namely, XMPP and SIMPLE.

A. XMPP

XMPP was initially given the name Jabber which was inspired by the Jabber open source community that was found by Jeremie Miller. In January 1999 Jeremie Miller released the source code for the initial version of the Jabber server. In 2002 the Jabber protocol was renamed to the neutral name Extensible Messaging and Presence Protocol (XMPP). XMPP is much more than just IM. It is an open Extensible Markup Language (XML) protocol for near-real-time messaging, presence, and request-response services. The basic syntax and semantics were developed originally within the Jabber open-source community, mainly in 1999 [10].

Starting from 2002, the IETF XMPP working group has worked on core features and extensions of XMPP to provide instant messaging and presence functionalities. An important feature of XMPP is that it supports the useful feature of CPIM gateways which allows XMPP clients to communicate using different protocols.

XMPP is responsible for message exchange between clients and servers by using XML *stream*, which is a container for the exchange of XML elements between any two entities over a network [10]. To start an XML stream, the opening tag `<stream>` is used to allow sending any number of other XML elements through the stream over a period of time, and the closing tag `</stream>` is used to end the XML stream. XML elements also use opening and closing tags. The attributes of the stream element are: to (should be used only in the XML stream header from the initiating entity to the receiving entity, and must be set to a hostname serviced by the receiving entity.), from (should be used only in the XML stream header from the receiving entity to the initiating entity, and must be set to a hostname serviced by the receiving entity that is granting access to the initiating entity), id (should be used only in the XML stream header from the receiving entity to the initiating entity, and is acting as a unique identifier created by the receiving entity to function as a session key for the initiating entity's streams with the receiving entity, and must be unique within the receiving application), xml (to define the default language of the stream) and version (to signal support for the stream-related protocols). Figure 7 shows a simple example of an XML stream with A as the initiating entity and B as the receiving entity.

```

A: <?xml version='1.0'?>
<stream:stream
    to='example.com'
    version='1.0'
B: <?xml version='1.0'?>
<stream:stream
    from='example.com'
    id='someid'
    xmlns='jabber:client'
    version='1.0'
    ... encryption, authentication, and resource binding ...
A: ... message ...
B: ... message ...
A: </stream:stream>
B: </stream:stream>

```

Figure 7. A simple XML stream.

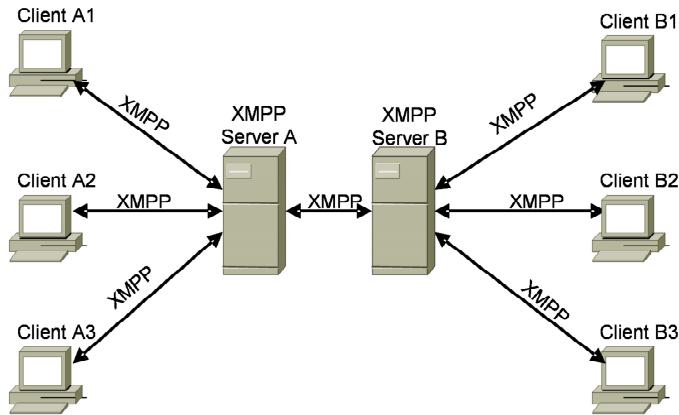


Figure 8. Sending messages over XMPP networks.

When sending messages between two clients connecting to the same XMPP server, the server sends the message directly to the recipient. However, if the sending clients is connected to another XMPP server, the sending XMPP server sends the message to the recipient's XMPP server then the message is sent to the recipient. Figure 8 illustrates the concept.

B. SIMPLE

SIMPLE, which stands for Session Initiation Protocol for Instant Messaging and Presence Leveraging Extensions, is an open standard instant messaging and presence protocol. As can be gathered from its name, SIMPLE is based on the more famous Session Initiation Protocol (SIP).

Session Initiation Protocol (SIP) is an application-layer control protocol that can establish, modify, and terminate multimedia sessions such as Internet telephony calls [11]. The mechanisms provided by SIP are more useful for presence applications and for session-oriented communication applications than for instant messaging [12]. For IM, SIP defines two modes, namely, the *page* mode (which sends messages without establishing a session)

and the *session* mode (which starts with establishing a session before sending instant messages). The IETF endeavored to develop standard IM protocol based on SIP that meets the requirements of the IMPP working group, and as a result they ended up with SIMPLE. The progress of the SIMPLE working group has been quite slow, mostly due to some complexities in the SIP protocol [13]. Most SIMPLE specifications are still ongoing work available as Internet Drafts.

3.2. Proprietary Protocols

Unlike open standard protocols, a proprietary protocol is owned by a single organization or individual. Developers of proprietary protocols keep protocol specifications and details away from public access and enforce restrictions so only the protocol's inventors or licensees are able to develop client software that depends on the protocol. There are also financial and legal issues associated with the use of proprietary protocols which adds even more restrictions.

Many of IM protocols are proprietary. Two examples are Microsoft Notification Protocol (MSNP) and RVP which were developed by Microsoft in 1999 and 1997 respectively and are used by Windows Live Messenger. Other examples of Proprietary protocols are Open System for Communication in Realtime (OSCAR) protocol, Skype Protocol, Talk to OSCAR protocol (TOC), TOC2 protocol, Yahoo! Messenger Protocol (YMSG) and Gadu-Gadu which is a Polish instant messaging protocol and is considered the most popular IM service in Poland.

4. IM Security

Software developers usually concentrate on the features that attract end-users rather than security-related issues when designing and implementing their software products. The software industry is plagued by such practices which treat security as an afterthought when planning and producing a new product. With the prevalence of IM and the millions of users benefiting from its services, it is very likely that attackers and creators of malicious programs will take advantage of the situation and exploit vulnerabilities in IM systems to infect a large sector of Internet community. In this section, we discuss some of the security features in IM systems and shed light on the threats surrounding them.

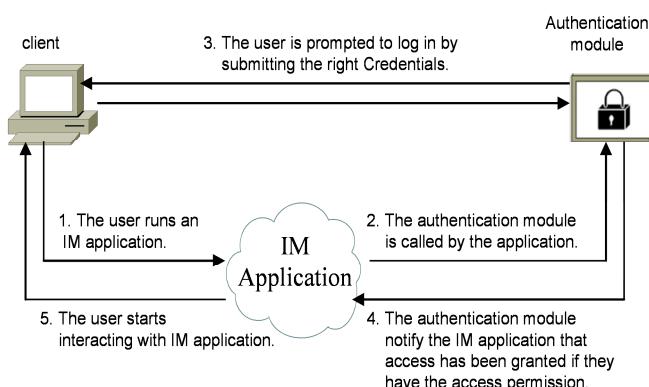


Figure 9. Authentication process.

4.1. Authentication in IM

Authentication is defined as the process of verifying an identity claimed by or for a system entity [17]. In the context of IM, a system entity could be represented by a user. IM applications are supposed to provide a security mechanism to authenticate users when they use IM clients to access servers. Furthermore, the IM application has to make sure that users are granted permission to use the appropriate system resources. For instance, Bob is not supposed to access Alice's list of contacts. The authentication process is usually based on a username and password that are chosen by a user when registering for the first time. The username is distributed by the user to all his/her contacts, whereas the password remains confidential. Figure 9 shows a typical authentication process that is used by an IM application.

A mechanism that is used for authentication is Single Sign-On (SSO), by which the user authenticates himself/herself only once and is automatically logged into Service Providers (IM server) as necessary without requiring further manual interaction [14].

An authentication method that is used by IM is Simple Authentication and Security Layer (SASL) which provides a generalized method for adding authentication support to connection-based protocols [10] and is published by IETF. SASL is supported by XMPP for authentication.

On the other hand, it is important to provide confidentiality for instant messages since a great deal of IM is happening over public and untrusted networks such as the Internet. Secure Socket Layer (SSL) is a protocol for managing the security of a message transmission on the Internet and it was developed by Netscape Communications Corporation. Later, the IETF used version 3.0 of SSL to develop a standard protocol, which became the Transport Layer Security (TLS) protocol. TLS protocol allows client/server applications to communicate in a way that is designed to prevent eavesdropping, tampering, or message forgery [15]. It works at the transport layer to authenticate the endpoints and encrypt messages between the authenticated entities. XMPP uses TLS to encrypt and secure the XML Stream across the network, and SIMPLE protocol relies on TLS as well.

Another way for securing IM is Off-the-Record (OTR) which allows clients to have private conversations over the network by providing strong authentication and encryption methods. Furthermore, OTR uses temporary per-message encryption keys, which makes compromising previous conversations impossible even if the attackers manage to capture a currently-used key.

Mannan and van Oorschot designed an Instant Messaging Key Exchange protocol (IMKE) [7] to ensure the confidentiality, integrity, and authentication of client-server and client-client communications. Their scheme is based on Password Authentication and Key Exchange (PAKE).

A common threat in IM environments is when attackers harness computer programs to sign up automatically for unlimited accounts to consume resources of IM servers. IM servers overcome this threat using the technique known as Completely Automated Public Turing test to tell Computers and Humans Apart (CAPTCHA). CAPTCHA is a challenge-response test that prompts the creator of the IM account to enter a required string of characters based on an image displayed on the screen. Usually, the image is displayed in a manner that only humans can recognize. A sample of CAPTCHA test is shown in figure 10.

Type the characters you see in the picture below.



Figure 10. CAPTCHA test.

4.2. IM Security Threats

IM clients provide the ability of transferring files between users in addition to exchanging instant messages. Such feature puts IM users at the risk of being infected by malicious programs that are hidden in the transferred files. In the following we explore some of the forms of these malicious programs.

A. Worm

An annoying self-replicating computer program that uses a network to copy itself to other machines. An important feature of worms is that they do not need a host program to be active and they can perform the replication without human intervention. In the context of IM, worms can infect user machines in different ways. For instance, a worm can be sent to an IM user in a file using automatic file transfer request from an infected client when initiating a conversation. Another way is to send malicious Uniform Resource Locators (URLs) of web pages as a text message encouraging the recipient to open these web pages.

Once detection is available for a particular worm, infected files can be stopped at the gateway. In the case of IM, however, detection software does not currently monitor traffic at the gateway level. If a worm starts to spread using IM, it cannot be stopped before it reaches the user's computer [16].

B. Trojan Horses

A computer program with a benign appearance and malicious intent. It tricks the user into executing it and usually opens a backdoor through which an attacker can access the system [13]. Trojan Horses exploit the ability of IM to transfer files by configuring the IM to share all files in the system with full access to every one using the IM port. Consequently, a backdoor is opened for hackers to access resources on the exploited machine. Firewalls can protect from Trojan Horses by blocking the port number which will block the traffic as whole. However, this is inappropriate for IM unless customized.

C. Denial-of-Service (dos) Attack

An attack that aims at overloading the system or service which it is directed at to deny legitimate users access to the service [13]. Attackers cause DoS attacks in the IM system by flooding the victim with unwanted messages which consume a large amount of CPU power

causing the IM client to crash, hang, or become unstable. IM clients may protect themselves from DoS by ignoring certain users using the ignore list.

D. Man-in-the-Middle (mitm) Attack

An attack that aims at listening to a conversation between two communicating parties, capturing messages exchanged in a communication, modifying messages before retransmission, or even sending messages to fool a communicating party into believing that they come from a genuine sender. The open standard solutions provide better protection against man-in-the-middle attacks, whereas proprietary IM solutions offer virtually no protection against them because the data are usually sent both unencrypted and unauthenticated, making reading and producing messages easy for an adversary [13].

E. Buffer Overflow Attack

An attack that tries to insert data that exceed the capacity of an IM application such as extremely long messages and huge files. Such attack can corrupt or overwrite existing data in the buffer. The IM Buffer overflow vulnerabilities have been found in all four of the major proprietary instant messaging solutions [13].

Recent versions of all major IM clients include an option to employ anti-virus software which can be launched automatically on every IM file download [7] which can protect IM clients from a variety of threats.

5. Practical Uses of IM

The instant and flexible nature of IM has resulted in a growing popularity of its applications in many sectors. In this section, we explore some of the practical uses of IM in various fields.

5.1. Business Uses

Recently, IM has been playing an important role in the business area as it enhances the business process performance by eliminating the office visits, voice mail, and calls. It facilitates collaboration between employees within the organization and eases contacting remote employees and connecting business partners using text, voice, and video-based means. IM is considered as a collaborative tool for tasks that require rapid exchange of information such as allowing employees located on different offices to conduct meetings and discussions, transfer documents, receive business support, solve problems, and respond to manager tasks quickly. Geographically separated workers can work on the same project using IM tools. The sensitivity of the messages exchanged between employees using IM tools in business compels managers to implement security measures against threats and unauthorized access especially when using a public network such as the Internet.

The research area Computer Supported Cooperative Work (CSCW) was first coined by Irene Greif and Paul M. Cashman. CSCW deals with approaches for collaborative activities

and their coordination that can be supported by computer systems. Instant Messaging tools are one of the subcategories of CSCW as a secure collaborative application [18].

5.2. Educational Uses

IM technology allows potentials of communication that offer advantages over other forms of communication in the educational field. IM tools can help students and learners to enhance many aspects of their educational experience. Universities can adopt IM applications to connect students with lecturers and librarians. Such a feature allows for further questions, discussions, and clarification that go beyond traditional means. It also allows students to communicate with each other during lectures about the lecture topic without interrupting the lecturers. Furthermore, students can communicate with each other to collaborate on team projects and assignments and have discussions about course material.

5.3. Social Uses

Most IM servers provide free, unlimited accounts for users. Therefore, obtaining an IM account and getting in touch with others is easy and inexpensive. With the falling costs of Internet access, users can take advantage of the convenience of IM tools to contact remote friends and relatives or even know new people.

6. Research on IM

IM has been subjected to many researches over the years. In this section we explore some of these researches and their results.

Osterman Research - which provides market research, cost modeling, benchmarking and related services to vendors and customers of messaging, collaboration and Web-related products, technologies and services - conducted a research which showed that IM was not just about instant messaging, but rather it was about the entire unified communications package. In that research, desktop videoconferencing showed the highest growth rate, increasing by 145% since 2007, whereas Web conferencing increased by 87% exceeding 2007's projected 37% growth.

In 2007, some researchers [3] used a quasi-experimental longitudinal research design to test how providing instant messaging to selected workgroups at a Fortune 500 company impacted employees' attitudes and work behaviour. Results from that research suggested that IM use had a positive effect on improving productivity with participants citing reductions in voice mail and phone tagging, and convenience of knowing whether colleagues were online and available to communicate, as well as increased productivity served by back-channel communications conducted via IM.

Another study by researchers at Ohio State University and University of California suggested that IM conversations were briefer than the telephone, e-mail, and face-to-face conversations. Also, it found that the employees were quite strategic in their use of IM as they checked their colleagues status before interrupting them which concluded that workers who used IM on the job reported less interruption than colleagues who did not.

Other researchers discussed the use of instant messaging system as a platform for collaborative applications with high security requirements. As an example for such applications, they defined scenarios of decision processes, which include a discussion and a subsequent e-voting. They presented a simple group-oriented protocol which is suitable to be included in their scenarios of discussion and decision in small groups [18].

On the other hand, an experiment using 44 teams in the United States was conducted. The results showed that teams using e-mail were more effective in terms of generating ideas than teams using instant messaging. There were no significant differences between the two communication methods, in terms of task difficulty, playfulness, and ease of use [19].

Another study explored the advantages and disadvantages of IM in a medical network environment. The authors concluded that IM might be, for many people, a convenient way of chatting over the Internet. They illustrated as well how an IM protocol could serve in the best way medical services and provide great flexibility to the involved parties. In addition, the directory services and presence status made it very attractive to medical applications that need to have real time and store and forward communication [20].

Another conducted research [6] examined uses of IM in a high-tech firm to illustrate how knowledgeable workers used this new tool to collaborate with co-workers. The study found that workers at the firm relied heavily on CMC for communication both within and outside the organization. They came up with a result that IM was used extensively to exchange work-related messages, coordinate and arrange meetings, and inquire about colleagues' availability for discussion.

7. Conclusion

In this chapter we have made an attempt to address the issue of instant messaging systematically. We started with introducing the basic concepts behind instant messaging and a brief historical background to show the progress of its systems. A whole section was dedicated for a discussion on the fundamentals of instant messaging including how it works and its various underlying architectures. Since protocols form the corner stone in the operation of instant messaging, a thorough discussion was given on the different types of IM protocols and their role in service provisioning. Thereafter, we shed some light on the issue of IM security since most IM applications operate on the public, insecure Internet. A mention was made of some of the major uses of instant messaging in various fields and areas. Finally, we concluded with the research that was conducted in instant messaging and its findings to help decision makers in terms of making informed decisions.

References

- [1] Mannan, M. & van Oorschot, P. C. (2006). "A Protocol for Secure Public Instant Messaging," in 10th Financial Cryptography and Data Security International Conference, 20-35.
- [2] Kucukyilmaz, T., Cambazoglu, B. B., Aykanat, C. & Can, F. (2008). "Chat mining: Predicting user and message attributes in computer-mediated communication," *Information Processing and Management*, vol. 44(4), 1448-1466.

- [3] Shaw, B., Scheufele, D. A. & Catalano, S. (2007). "The role of presence awareness in organizational communication: An exploratory field experiment," *Behaviour & Information Technology*, vol., 26(5), 377-384, September.
- [4] Huang, A. H. (2007). "*Theoretical Foundations and a Framework for Instant Messaging Research*," in International Conference on Business and Information, Japan,.
- [5] Jennings, R. B., Nahum, E. M., Olshefski, D. P., Saha, D., Shae, Z. & Waters, C. (1996). "A Study of Internet Instant Messaging and Chat Protocols," *IEEE Network*, vol. 20(4), 16-21, July-Aug.
- [6] Quan-Haase, A., Cothrel, J. & Wellman, B. (2005). "Instant messaging for collaboration: A case study of a high-tech firm," *Journal of Computer-Mediated Communication*, vol. 10(4).
- [7] Mannan, M. (2005). "Secure Public Instant Messaging," *master's thesis*, School of Computer Science, Carleton University, Ontario, Canada, August.
- [8] Day, M., Aggarwal, S., Mohr, G. & Vincent, J. (2000). "Instant Messaging / Presence Protocol Requirements," RFC 2779, *IETF Network Working Group*, February.
- [9] Peterson, J. (2004). "Common Profile for Instant Messaging (CPIM)," RFC 3860, *IETF Network Working Group*, August.
- [10] Saint-Andre, P. (2004). "Extensible Messaging and Presence Protocol (XMPP): Core," RFC 3920, *IETF Network Working Group*, October.
- [11] Rosenberg, J., Schulzrinne, H., Camarillo, G., Johnston, A., Peterson, J., Sparks, R., Handley, M. & Schooler, E. (2002). "SIP: Session Initiation Protocol," RFC 3261, *IETF Network Working Group*, June.
- [12] Campbell, B; Rosenberg, J; Schulzrinne, H; Huitema, C; Gurle, D. (2002). "Session Initiation Protocol (SIP) Extension for Instant Messaging," RFC 3428, *IETF Network Working Group*, December.
- [13] Salin, P. (2004). "Mobile Instant Messaging Systems - A Comparative Study and Implementation," *master's thesis*, Dept. *Computer Science and Engineering*, Helsinki University of Technology, Espoo, Finland.
- [14] Pashalidis, A. & Mitchell, C. J. (2003). "A Taxonomy of Single Sign-on Systems," in 8th Australasian Information Security and Privacy Conference, Wollongong, 249-264.
- [15] Dierks, T. & Rescorla, E. (2008). "The Transport Layer Security (TLS) Protocol Version 1.2," RFC 5246, *IETF Network Working Group*, August 2008.
- [16] John, R. (2007). *Practical Internet Security*. Vacca: Springer.
- [17] Shirey, R. (2000). "Internet Security Glossary," RFC 2828, *IETF Network Working Group*, May.
- [18] Meletiadou, A. & Grimm, R. (2009). "Using Instant Messaging Systems as a Platform for Electronic Voting," in E-Technologies: *Innovation in an Open World*, vol. 26, Berlin: Springer, 12-24.
- [19] Huang, A. H., Hung, S. & Yen, D. C. (2007)."An exploratory investigation of two internet-based communication modes," *Computer Standards & Interfaces*, vol., 29(2), 238-243, February.
- [20] Sachpazidis, I., Ohl, R., Kontaxakis, G. & Sakas, G. (2006). "TeleHealth networks: Instant messaging and point-to-point communication over the internet," Nuclear Instruments and Methods in Physics Research Section A: *Accelerators, Spectrometers, Detectors and Associated Equipment*, Vol. 569(2), 631-634.

Chapter 9

INSTANT MESSAGING COMMUNICATION: SELF-DISCLOSURE, INTIMACY, AND DISINHIBITION

Joshua Fogel*

Department of Economics, Brooklyn College, Brooklyn, NY, USA

Abstract

Individuals use instant messaging to communicate. This chapter reviews the empirical research from scholarly journals on what is known about the topics of self-disclosure, intimacy, and disinhibition with regard to general instant messaging. The search terms of “(instant message OR instant messaging) AND (self-disclosure OR intimacy OR disinhibition)” were searched in the databases of Medline, PsycINFO, CINAHL, and Business Source Premier from January 1990 to June 2009. Seven articles were reviewed from studies of adolescents and college students. Instant messaging use is associated with self-disclosure, intimacy, and disinhibition. Research is needed to determine if these associations apply to adults too.

Keywords: instant messaging, self-disclosure, intimacy, disinhibition, relationships

Introduction

There are over one billion users of instant messaging worldwide. Users typically send an average of 53 messages each day (Radicati & Khmartseva, 2009). Teenagers and young adults more often send instant messages to friends than do adults (Jones & Fox, 2009). Those who use instant messaging are those who multi-task (Shiu & Lenhart, 2004).

The Internet allows for a number of ways for individuals to communicate. This includes communication through an online forum, blog, e-mail, chat and instant messaging. In a study of content posted on six online forums, support forums had more than 50% of the content of first messages with high self-disclosure while discussion forums had less than 2% of the

* E-mail address: joshua.fogel@gmail.com. Phone: (718) 951-3857, Fax: (718) 951-4867. Correspondence: Joshua Fogel, PhD Brooklyn College of the City University of New York, Department of Economics, 218A, 2900 Bedford Avenue, Brooklyn, NY 11210, USA.

content of first messages with high self-disclosure and 15% with low self-disclosure (Barak & Gluck-Ofri, 2007). Gender differences exist for self-disclosure with female college students more likely to self-disclose than males (Punyanunt-Carter, 2006). Also, among adolescents, males are more likely than females to self-disclose sexual content on the Internet (Chiou, 2006).

With regard to self-disclosure on blogs, bloggers who self-disclose believe that this helps them better manage relationships (Lee, Im, & Taylor, 2008). Both personal profile information content anonymity and also visual photo anonymity are not associated with self-disclosure on blogs (Qian & Scott, 2007). With regard to e-mail, females are more likely than males to use intimacy markers such as affectionate sign-offs (Colley & Todd, 2002). With regard to chat, motivation for chatting is related to self-disclosure content. For example, if the motivation is to form a relationship, there is a greater amount of self-disclosure than if there would be motivation for entertainment purposes or for information (Cho, 2007).

In this chapter, a comprehensive review of self-disclosure and also the topics of intimacy and disinhibition that can be related to online self-disclosure are reviewed for instant messaging use. This review is conducted on all the data-based articles published in peer reviewed journals on the topic of instant messaging from January 1990 to June 2009.

Method

Inclusion and Exclusion Criteria

Criteria for inclusion and exclusion were determined a-priori before searching in the relevant databases. Inclusion criteria for the reviewed studies included that they were either qualitative or quantitative articles with empirical data about the association of instant messaging with either: 1) self-disclosure, 2) intimacy, or 3) disinhibition. Exclusion criteria for the reviewed studies included that they were: 1) from non-peer reviewed journals, 2) theoretical articles, 3) contained anecdotal information, and 4) were not written in English.

Search Strategy

On June 9, 2009, a number of databases were searched for all the relevant studies from the year of 1990 to that date. The search strategy consisted of the two sets of terms: (instant message OR instant messaging) AND (self-disclosure OR intimacy OR disinhibition). Both the subject headings and text words in the titles and abstracts were searched for these terms.

Databases searched used the Ebsco interface and included Medline, PsycINFO, CINAHL, and Business Source Premier. Also, a search was done by reading the relevant retrieved articles to determine if any other relevant articles were possibly quoted.

Results

The search retrieved 2 hits with Medline, 7 hits with PsycINFO, 2 hits with CINAHL, and 2 hits with Business Source Premier. These hits were not all unique, as there was overlap

in the articles retrieved from these databases. Overall 7 different articles were deemed relevant and were included in the review.

The Table summarizes some details about the 7 reviewed articles. The articles were published from the years of 2002 to 2009. Only 3 of 7 articles include information about when data were collected. Five studies were about adolescents and 2 were about college students. The studies were conducted in three countries of the United States, Canada, and the Netherlands. Three studies had longitudinal study designs while 4 had cross-sectional study designs.

Table. Characteristics of Instant Messaging Studies Reviewed

Study Reference	Year Data Collection	Study Sample	Study design
Gross et al. (2002)	not provided	130 seventh grade students from southern California in United States	Longitudinal
Hu et al. (2004)	not provided	123 college students in northeastern United States	Cross-sectional
Schouten et al. (2007)	not provided	1,340 elementary and secondary schools in the Netherlands	Cross-sectional
Valkenburg & Peter (2007)	2004	794 elementary and secondary schools in the Netherlands	Cross-sectional
Dimmick et al. (2007)	not provided	286 college students in the midwestern United States	Cross-sectional
Blais et al. (2008)	2001-2002	884 adolescents from rural and urban Canadian schools	Longitudinal
Valkenburg & Peter (2009)	2006	812 adolescents from the Netherlands	Longitudinal

Below are brief summaries about all the included and reviewed articles. They are organized by separate categories of adolescents and college students and then the studies are summarized in order of year of publication.

Adolescents

Gross, Juvonen, and Gable (2002) studied 130 seventh grade students from southern California from a middle to higher socioeconomic status community. The average age was 12 years, included 62.3% females, and included 40% minorities with the largest minority group of 17.5% Asian American. Participants completed measures in school and also that night and the following two consecutive nights before going to sleep. Participants spent on average 28.85 minutes daily using instant messaging and exchanged messages with an average of 2.68 different people daily. Popular topics of use included friends (58%), gossip (51%), and “boyfriend/girlfriend stuff” (50%). Popular reasons for use included “hanging out with a friend” (92%) and boredom (74%). The relationships with those whom they instant messaged were typically those that they initially met at school and developed a long-term friendship. Only 12% of those whom they instant messaged were those that they initially met first online. These relationship initiation patterns did not differ by gender. In the multivariate regression analysis, both daily social anxiety and daily loneliness were associated with instant messaging use with those whom they typically did not have a close relationship.

Schouten, Valkenburg, and Peter (2007) surveyed 1,340 adolescents from 6 elementary and secondary schools in the Netherlands. The average age was slightly above 14 years and included 49% females. Race/ethnicity information was not provided. The authors tested a model to understand self-disclosure during instant messaging communication. Participants with greater disinhibition were more likely to self-disclose during instant messaging communication. Both a sense of controllability during instant messaging communication and also perceiving that during instant messaging communication there are reduced nonverbal cues were associated with greater disinhibition during instant messaging communication. This model also worked in the separate analyses by gender for both males and females.

Valkenburg and Peter (2007) surveyed 794 adolescents from 6 elementary and secondary schools in the Netherlands. The average age was slightly above 13 years, included 49% females, and consisted only of those who were of White race/ethnicity. Instant messaging use with MSN had a significant positive correlation ($r=0.32$) with communication with pre-existing offline friends. Analyses were also performed as part of the overall category of online communication that included instant messaging as part of that overall category. These results included an association of online communication with closeness to friends for those who communicated with pre-existing friends while no such association of closeness existed for those who communicated with strangers.

Blais, Craig, Pepler, and Connolly (2008) surveyed 884 adolescents from rural and urban Canadian schools at baseline and at one-year later during 2001 and 2002. The average age was 15 years, included 54% females, and included 24% who identified as non-White race/ethnicity. Instant messaging was measured by use of ICQ instant messaging, which the authors report that at the time of their data collection was an extremely popular instant messaging source. Multivariate analyses were conducted for instant messaging at baseline as a possible predictor for a number of variables measuring friendship quality one year later. Each model included covariates of age, sex, the outcome variable at baseline and also use of chat rooms at baseline. Instant messaging use was significantly associated with outcome variables one year later of higher levels of a) commitment, b) trust and communication, and c) intimacy and companionship. Instant messaging at baseline was not significantly associated with the outcome variable one year later of alienation and conflict. Also, similar analyses for romantic relationships were conducted for the sub-sample of 610 adolescents who reported being involved in romantic relationships in secondary school. Instant messaging use at baseline was significantly associated with outcome variables one year later of higher levels of a) commitment, b) trust and communication, and c) intimacy and companionship. Instant messaging use at baseline was not significantly associated with the outcome variable one year later of alienation and conflict.

Valkenburg and Peter (2009) surveyed 812 adolescents from the Netherlands at two time periods of six months apart in 2006. The ages ranged from 10 to 17 years and included 50% females. Race/ethnicity information was not provided. Average instant messaging use was 1 hour and 22 minutes daily at baseline and at follow-up was 1 hour and 23 minutes daily. Correlational analyses showed significant positive correlations of instant messaging use at baseline with intimate online self-disclosure at both baseline and follow-up. Also, there were significant positive correlations of instant messaging use at baseline with quality of friendships at follow-up but not at baseline. Instant messaging use at follow-up had significant positive correlations with intimate online self-disclosure at both baseline and follow-up. No significant correlations occurred for instant messaging use at follow-up and

quality of friendships whether at baseline or follow-up. In their structural equation modeling analyses, instant messaging use at baseline was significantly associated with intimate online self-disclosure at follow-up after including quality of friendships in the model. Also, the relationship of instant messaging use at baseline with quality of friendships at follow-up was mediated by intimate online self-disclosure. These patterns also occurred for different adolescent age subgroups.

College Students

Hu, Wood, Smith, and Westbrook (2004) surveyed 123 college students in northeastern United States. The average age was slightly above 21 years and included 46% females. Race/ethnicity information was not provided. The most favorite place for using instant messaging was at home as compared to a computer lab or at a job. The instant message software was active (i.e., “on”) for an average of 10 hours daily while actual use of instant messaging was about 2 hours daily. Increased instant messaging use was associated with increased perceived verbal intimacy (measuring conversation content and self-disclosure), perceived affective intimacy (measuring feelings of proximity, understanding, and trust), and perceived social intimacy (measuring interpersonal relationships between friends and within marriages).

Dimmick, Ramirez, Wang, and Lin (2007) surveyed 286 college students in the midwestern United States. The average age was slightly above 20 years, included 55% females, and consisted of 81% of those from White race/ethnicity. The larger the sub-network size (proportion of participant’s overall network with whom participant used a specific communication technology) and sociability gratifications (relationship maintenance) were significantly associated with frequency of instant messaging use. Also, for instant messaging, both sub-network size and intimacy were significantly associated with sociability gratifications and also sub-network size was significantly associated with gratification-opportunities (the relative ability of the media to obtain gratification through effective communication).

Conclusions

Self-disclosure

One of the reviewed studies showed an association of instant messaging use with intimate online self-disclosure (Valkenburg & Peter, 2009a). This finding is consistent with what is previously known that general computer-mediated communication stimulates online self-disclosure (Valkenberg & Peter, 2009b).

Intimacy

Five of the reviewed studies (Blais et al., 2008; Dimmick et al., 2007; Gross et al., 2002; Hu et al., 2004; Valkenburg & Peter, 2007) focused specifically on intimacy or intimacy-related topics. Specific intimacy topics findings included an association of instant messaging

use with general intimacy (Blais et al., 2008) and also verbal intimacy, affective intimacy, and social intimacy (Hu et al., 2004). Intimacy-related topics involved closeness and relationships. Findings included an association of instant messaging use with closeness to friends for pre-existing friends (Valkenburg & Peter, 2007) and also to either no closeness (Valkenburg & Peter, 2007) or even social anxiety (Gross et al. 2002) for strangers or to those without a close friendship. Relationship maintenance was also associated with frequency of instant messaging use (Dimmick et al., 2007). These studies overall show that instant messaging is associated with increased intimacy for friends.

Disinhibition

One of the reviewed studies showed an association of instant messaging use with disinhibition (Schouten et al. 2007). Although more research would be useful to further understand disinhibition with regard to specifically instant messaging use, this one study suggests that the presence of disinhibition is not necessarily occurring among all instant messaging users. There needs to be a sense of either controllability or knowledge that there are reduced nonverbal cues for this disinhibition to occur.

Limitations and Future Research

All of the reviewed studies were specifically for studies that focused on adolescents or young adults (i.e., college students). There are apparently no studies about instant messaging and the topics of self-disclosure, intimacy, and disinhibition for adults. As adults use instant messaging too, it would be very useful to study these topics among adults.

References

- Barak, A., & Gluck-Ofr, O. (2007). Degree and reciprocity of self-disclosure in online forums. *CyberPsychology & Behavior*, **10**, 407-417.
- Blais, J. J., Craig, W. M., Pepler, D., & Connolly, J. (2008). Adolescents online: The importance of Internet activity choices to salient relationships. *Journal of Youth and Adolescence*, **37**, 522-536.
- Chiou, W.-B. (2006). Adolescents' sexual self-disclosure of the Internet: Deindividuation and impression management. *Adolescence*, **41**, 547-561.
- Cho, S. H., (2007). Effects of motivations and gender on adolescents' self-disclosure in online dating. *CyberPsychology & Behavior*, **10**, 339-345.
- Colley, A., & Todd, Z. (2002). Gender-linked differences in the style and content of e-mails to friends. *Journal of Language and Social Psychology*, **21**, 380-392.
- Dimmick, J., Ramirez, A., Jr., & Wang, T. (2007). 'Extending society': The role of personal networks and gratification-utilities in the use of interactive communication media. *New Media & Society*, **9**, 795-810.
- Gross, E. F., Juvonen, J., & Gable, S. L. (2002). Internet use and well-being in adolescence. *Journal of Social Issues*, **58**, 75-90.

- Hu, Y., Wood, J. F., Smith, V., & Westbrook, N. (2004). Friendships through IM: Examining the relationships between instant messaging and intimacy. *Journal of Computer-Mediated Communication*, **10**(1). Retrieved October 7, 2009 from <http://jcmc.indiana.edu/vol10/issue1/hu.html>
- Jones, S., & Fox, S. (2009). *Generations online in 2009*. Washington, DC: Pew Internet and American Life Project. Retrieved October 7, 2009 from <http://www.pewinternet.org/Reports/2009/Generations-Online-in-2009.aspx>
- Lee, D.-H., Im, S., & Taylor, C. R. (2008). Voluntary self-disclosure of information on the Internet: A multimethod study of the motivations and consequences of disclosing information on blogs. *Psychology & Marketing*, **25**, 692-710.
- Punyanunt-Cater, N. M. (2006). An analysis of college students' self-disclosure behaviors on the Internet. *College Student Journal*, **40**, 329-331.
- Qian, H., & Scott, C. R. (2007). Anonymity and self-disclosure on weblogs. *Journal of Computer-Mediated Communication*, **12**, 1428-1451.
- Radicati, S., & Khmartseva, M. (2009). *Email statistics report, 2009-2013 - executive summary*. Palo Alto, CA: Radicati Group. Retrieved October 7, 2009 from <http://www.radicati.com/wp/wp-content/uploads/2009/05/email-stats-report-exec-summary.pdf>
- Schouten, A. P., Valkenburg, P. M., & Peter, J. (2007). Precursors and underlying processes of adolescents' online self-disclosure: Developing and testing an "Internet-attribute-perception" model. *Media Psychology*, **10**, 292-315.
- Shiu, E., & Lenhart, A. (2004). *How Americans use instant messaging*. Washington, DC: Pew Internet and American Life Project. Retrieved October 7, 2009 from http://www.pewinternet.org/~/media/Files/Reports/2004/PIP_Instantmessage_Report.pdf
- Valkenburg, P. M., & Peter, J. (2007). Preadolescents' and adolescents' online communication and their closeness to friends. *Developmental Psychology*, **43**, 267-277.
- Valkenburg, P. M., & Peter, J. (2009a). The effects of instant messaging on the quality of adolescents' existing friendships: A longitudinal study. *Journal of Communication*, **59**, 79-97.
- Valkenburg, P. M., & Peter, J. (2009b). Social consequences of the Internet for adolescents. *Current Directions in Psychological Science*, **18**, 1-5.

Chapter 10

DIGITAL WATERMARKING FOR IPR PROTECTION OF MULTIMEDIA CONTENTS

Sarabjeet Singh Bedi¹ and Shekhar Verma²

¹M. J. P. Rohilkhand University, Bareilly, India

²Indian Institute of Information Technology, Allahabad, India

Abstract

Digital Watermarking technology has been proposed for the implementation of Digital Right Management (DRM) system by establishing ownership right, ensuring authorized access and content authentication to protect the Intellectual Property Rights (IPR). The existing basket of technologies like cryptography secure the multimedia data only during storage or transmission and not while it is being consumed. Digital Watermarking provides an answer to this limitation as the watermark continues to be in the data during its usage. A watermark is designed to permanently reside in the original data, and extraction of this watermark provides the protection of IPR. When the watermark is permanently embedded into digital data at the one hand it may be used for checking whether the data have been modified. On the other hand, the detection of the watermark affects the way it is used in practical application. In watermarking applications, watermark extraction raise security issues and need to be protected from several standard data manipulations and modifications.

The goal of this chapter is to address the theoretical and practical aspects related to watermarking and the issues related to imperceptibility, robustness and security problems in digital contents. The tradeoff between major requirements in watermark embedding is elaborated and examined the existing solutions proposed for the same. The chapter elucidates various aspects of digital watermarking for types of multimedia signals and attacks. The techniques of digital image watermarking in spatial and transform processing domain have also been reviewed in this chapter. The chapter concludes with observations and future directions for researchers to design more robust and secure digital watermarking schemes to address the emerging region of real life applications like medical, telemedicine, insurance, defense, mobile communication and entertainment media, where the need for authentication is often high. This would lead to a basis for design of media security systems for protection of the IPR for digital multimedia contents.

1. Introduction

1.1. Information Security and Issues

Information security aspects come into role when it is essential to protect information as it is being shared during transmission or storage from an opponent who may present a threat to confidentiality, authenticity, integrity, access control, and availability. The need for information security has been termed as security attack, mechanism, and services [1].

The standard manipulation and modifications during flow of concerned information is constitutes an attack. The aim of attack is to destroy the security services. The attacks have been categorized in terms of passive and active according to the nature of the attack. Active attacks involve modification of data stream or the creation of a false stream. While passive attacks involve monitoring of transmissions and categorized as release of message contents and traffic analysis. The security mechanism is designed to detect, prevent and recover the information from such security attacks. The services are intended to counter security attacks and make use of one or more security mechanism to provide the service.

In the information security course of action the information is to be transmitted from source to destination through a communication channel. A logical information channel is established by defining a route through the communication channel from source to destination and by the cooperative use of communication protocol. A trusted third party may be needed to achieve secure transmission. In this regard at the sender's end the secret information is encoded with message. The transmitted message can only decoded, if receiver has secret information.

The process describes the three issues for the design of a particular security service algorithm. According to first issue the algorithm should generate additional secret information such that an opponent cannot defeat its purpose. Development of schemes for the distribution and sharing of the secret information is described in the second issue. Whereas the last issue focused on specification of a protocol to be used by the two principles that make use of the security algorithm and the secret information to achieve a particular security service [1].

On the basis of the information security process, various techniques have been proposed as security mechanisms to provide services and prevent security attacks.

1.2. Techniques for An Information Security

The techniques for information security have two components. First component is security-related transformation on the information to be sent. Second component describe the secret information shared by the sender and receiver, which is unknown to the receiver. Cryptography [2] is a technique having both components and provides information security [2]. The purpose of cryptography is to make information unreadable except by people who are authorized to see it. Security solution for such dilemmas is encryption and decryption. Encryption is based on algorithms that scramble information into unreadable form. Decryption is the process of using the same algorithm to restore the

scrambled information to its original form. The study of information security includes not only cryptography but also traffic security, importance of privacy and protection of intellectual property rights, whose essence lies in hiding information [3]. Therefore, protecting private information and intellectual property during digital communication has becomes an important issue. These issues have been addressed in sub discipline of information hiding, called steganography.

Steganography is the method of hiding information into perceivable information sources, such as images, audio, and video data. Steganography is sometimes confused with cryptography [4]. In contrast to cryptography, which is about protecting the content of messages, setganography is about concealing their existence. The goal of cryptography is to make data unreadable by a third party, where as the goal of steganography is to hide the data from a third party. However, steganography has a number of disadvantages as well. Unlike encryption, it generally requires a lot of overhead to hide a relatively few bits of information. Once a steganographic system is discovered, it is rendered useless. This problem, too, can be overcome if the hidden data depends on some sort of key for its insertion and extraction [1]. To cater these needs the other form of steganography named digital watermarking has been introduced.

Digital watermarking is the process of embedding information into digital multimedia contents such that the embedded information (watermark) can be extracted later. In contrast to other data hiding techniques (like cryptography, stegnography, digital signatures, fingerprinting, as described earlier) digital watermarks are transparent signature, which are inserted as noise with original data and seek to be robust and secure. The watermarking techniques embed the information of watermark signal into original signal and may also employ cryptography as well. The cryptography tool adds an extra security during information exchange to watermarking.

Digital watermarking has been considered as new field of research, which includes the area of computer science, Information Technology, cryptography, communication and Information Security. Therefore the scientist, researchers, professionals and especially multimedia content providers and distributors showed keen interest in this new emerging field.

2. Digital Watermarking Technology

DWM technology has been proposed for the implementation of Digital Right Management (DRM) system for Intellectual Property Right (IPR) protection. A watermark is designed to reside permanently in the original data, and extraction of this watermark provides the protection of IPR. When the watermark is permanently embedded into digital data at the one hand it may be used for checking whether the data have been modified. On the other hand, the detection of the watermark affects the way it can be used in practical application.

2.1. Digital Watermarking System Model

The digital watermarking system model shown in fig. 1 is described using three basic functional components as the generation of watermark, insertion of the watermark and extraction of the watermark from the watermarked signal [4]. The details of these components are as follows:

2.1.1. Watermark Generation

The generating function, f_g of watermark image data, I_w is to be added to the original image data, I_o depending on key, k . The key may be used as secret or public type. Therefore watermark generation function can be used to protect the watermark and make it secure for the purpose of authentication. This operation can be represented as:

$$I_g = f_g(I_w, k, I_o)$$

Where I_g – generated watermark, f_g – generating function, I_w - watermark data, k – key, I_o – Original Image (optional).

2.1.2. Watermark Embedding

The insertion function f_i embeds the generated watermark data, I_g into original data, I_o and forms the watermarked data, I_r . The insertion function can modify the original data according to watermark data I_g and key k . The insertion is performed in such a way that modification is perceptually similar to the original data. This allows insertion of controlled amount of “distortion” in the original data that can be represented as

$$I_r = f_i(I_o, I_g, k)$$

Where I_r – watermarked image, f_i – Insertion function, I_o – Original Image, I_g - generated Watermark, k – key (optional).

2.1.3. Watermark Extraction

The extraction function, f_e extracts the watermark data from received watermarked data, I'_r with the use of corresponding key, k , and original data, I_o . This operation can be represented as

$$I'_w = f_e(I_o, I'_r, k)$$

Where I'_w – Extracted watermark; f_e – Extraction function; I_o – Original Image; I'_r – received watermarked image k – key.

At the receiver side the received watermarked data, I'_r is to be checked for the existence of watermark data I_g . The watermark extraction proves the ownership, while watermark detection only verifies the ownership.

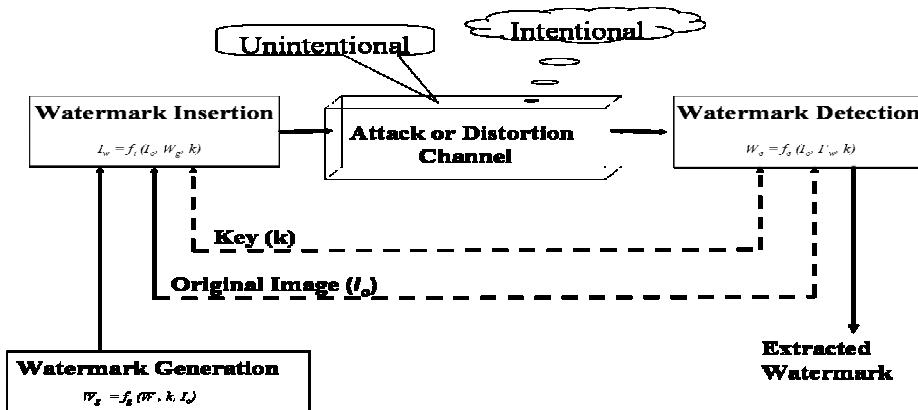


Figure 1. Digital Watermarking System.

The received watermarked data, I'_r may be the indistinct form of watermarked data, I_r . In this regard image quality distortion is measured using methods like Peak Signal to Noise Ratio (PSNR), and Normalized Cross Correlation (NCC).

2.2. Digital Watermarking System Requirements

The watermarking system has the following requirements.

2.2.1. Perceptibility

The inserted watermark should be perceptually transparent. Watermark should not create visible artifacts in marked data until it is compared with original data. By taking advantage of the psycho visual and psycho auditory properties, effective watermarking schemes can be designed to embed transparent watermarks [5].

2.2.2. Robustness

The robustness of watermarking system means that the watermark must be resistant to the distortion introduced during either normal use or a deliberate attempt to remove the watermark. The robustness is assessed by measuring the detection probability of the watermark and the bit error rate for a set of criteria that are relevant for the application.

2.2.3. Payload Capacity

It is the amount of information required to store the watermark in the original data which would not degrade the quality of image. “Watermarking granularity” is a term used to refer to the number of bits that are actually needed to represent the entire watermark in the image.

2.2.4. Watermark Security

Securing the watermark prevents unauthorized users from access and modifications. Two types of secrecy are defined. In the first type, unauthorized user cannot read and detect the

watermark, while in second type of secrecy; only the detection is possible, wherever the reading of hidden data is not possible without the secret key.

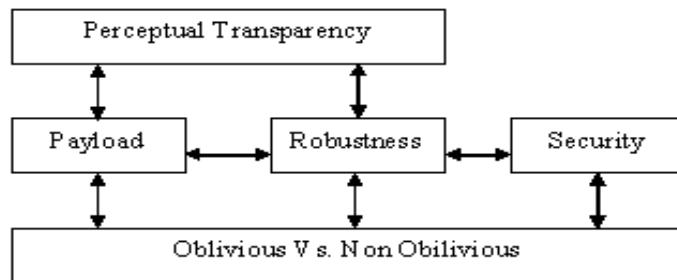


Figure 2. Common Dependencies among the Basic Requirements of Digital Watermarking System.

2.2.5. Need of Original Data

Original data is required depending on the application and type of watermarking method used. The availability of original data does not only help in detection of the watermark, it is also useful in inversion of distortion.

The common dependencies among the basic requirements of watermark are in conflict to a certain degree. Increasing the payload of the watermark reduces the robustness or increase the perceptual impact of that watermark [6]. The oblivious watermarking is more robust and is able to tolerate higher payload for similar perceptual impact as compared to non-oblivious watermarking schemes. However, the requirement of original data limits the application of oblivious watermarking.

2.3. Types and Applications

2.3.1. Digital Watermarking in Different Media

Watermarking techniques for digital media can be divided into four categories according to the type of documents to be watermarked viz. text, image, audio and video. Text watermarking methods embed a watermark in a formatted document by altering either the appearance or the position of a text element. However due to limited payload capacity of text and binary nature of text document, text watermarking is limited as compared to images and video watermarking [7]. The watermarking for digital images either, insert the watermark directly into the original image data or into some transformed version of the original image data to take advantage of perceptual property. The audio watermarking methods have focused on either direct watermarking of the audio signal or bit stream embedding in which the audio is represented in a compressed format. Video watermarking is described for video sequences consist of a series of consecutive and equally time spaced still images. Generally, image watermarking techniques are directly applicable to video sequences. However, similarity between adjacent frames introduces additional design constraints.

2.3.2. Types of Digital Watermarking

The broad categorization of the digital watermarking types is shown in Figure 2. The watermark based on imperceptible can be visible or invisible in digital media. A visible watermark is embedded in visual contents of digital media in such a way that they are visible along with the content in viewed while an invisible watermark is designed to be transparent to the observer and detected using signal processing techniques.

Depending on robustness, the invisible watermark can be classified into three main categories. A robust watermark is designed to resist attacks that do not seriously affect the quality and value of the original image. The second category does not tolerate any tampering that modifies the complete integrity of the image, is named fragile invisible watermark. The third category is semi-fragile watermark, which is capable of tolerating some degree of the change to a watermarked image, such as the addition of quantization noise from lossy compression.

The watermarking can also be classified on the basis of embedding of watermark in original data. The embedding can be done in different processing domains. One approach is to transform the original image into a transform domain representation and embed the watermark data therein. The second approach is to directly embed in the pixel intensity values of the original data to make the information imperceptible.

Watermarking extraction methods are referred as non-oblivious (non-blind) and oblivious (blind) watermarking according to the necessity of an original data [8]. The non-oblivious watermarking requires the original data for extraction; whereas the original data is not required in oblivious watermarking.

2.3.3. Digital Watermarking Applications

Digital watermarking techniques have applications in a wide range. Some applications of digital watermarking are as follows:

- a) Content Authentication: Content authentication requires detection of change or modification in the data under consideration. The presence of watermark ensures the prevention of modification in data anyway.
- b) Ownership Verification: The embedded data in original data can be used as a proof of ownership. In such cases the watermark must be very robust to unintentional attacks.
- c) Fingerprinting: To trace the source of illegal copies, the owner can use a fingerprinting technique. This requires the owner to embed different watermark information onto copied of the data provided to different customers.
- d) Copy Control: Copy control is to prevent the copy of digital document from permitted number of copies to authorized party. The watermark insert in original document indicates the number of copies permitted for copying and the watermark will be modified to indicate remaining number of copies.

Besides these, digital watermarking is also applicable in other areas like Broadcast Monitoring, Medical Safety and Mobile Communication Networks.

3. Security Issues in Digital Watermarking

In watermarking applications, watermark extraction raise security issues and need to be protected from several standard data manipulations and modifications. These standard manipulation and modifications are called attacks. The aim of attacks is to destroy the watermark. In the adamant of information security field, there is rapid growth of attacks which cause of illegal use of copyright and malpractice of IPR of digital contents. It may be practical impossible to design a system pompous to all forms of attack and new methods to defeat watermarking system, which will be invented in time. But certainly knowledge of common attack is a requirement for the design of improved system.

An extensive list of attacks and the related countermeasures are described in [9] where attacks are broadly categorize in to two groups named intentional and unintentional attacks. The intentional attacks are performed by users who know that a watermark is present and moreover the method with which they are embedded. These attacks use different methods to confuse the software so the watermark cannot be removed. The temper proofing, forgery, illegal copying are the focused area of such attacks. The attack performed by a user without specific intent to remove the watermark is called unintentional attacks. The causes of such attacks are due to format conversion, analog to digital or digital to analog conversion, and images when being prepared for publication.

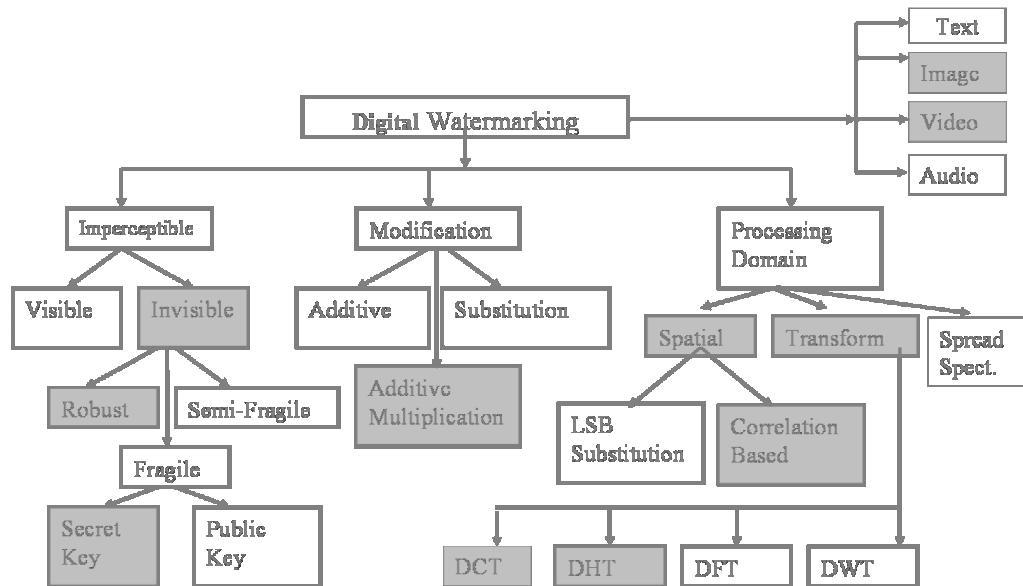


Figure 3. Types of Digital Watermarking (Gray colored boxes indicate the extensive research area in watermarking).

In [10] the attacks are grouped in to four categories: removal, geometrical, cryptographic, and protocol attacks. The removal attack is to remove the embedded watermark without cracking the security of the watermark embedding technique. This category includes de-noising, vector quantization, demodulation and collusion attacks [11]. In contrast to removal attacks, geometric attacks do not actually remove the embedded watermark itself, but intended to distort the watermark detector synchronization with the embedded information.

Cryptographic attacks aims to remove the embedded watermark by cracking the security mechanism for watermarking scheme. Such attacks are based on concepts of Brute force search [12] of key space, statistical averaging watermarking images. The protocol attack is based on concept of invertible watermarks [9] in which the attacker subtract his own watermark data and claims to be the owner of the watermarked data.

The possible attacks against watermarks are wide and varied. In order of merit, a watermarking technique must handle unintentional attacks viz. common image processing operations. The detailed categorization of attacks is shown in Figure 3. The gray colored boxes in figure indicated the most extensive areas in research.

4. Digital Watermarking Techniques for Images

4.1. Techniques for Images in Spatial Domain

The digital watermarking technique in the spatial domain is to embed the watermark into the pixels intensity values of the pixels of the original image directly. It is most appropriate for fragile watermarking techniques [13, 14] where the watermark is not robust to image processing operations. Spatial domain watermarking methods have larger capacity to embed maximum data. Since every pixel of watermark in watermarked image has specific location, therefore it is finest for content authentication.

A watermarking technique in spatial domain must fulfill the following properties.

- a) The watermarking system should be able to detect any changes made in a marked image after marking.
- b) Watermarking should not alter the quality of the image in a large extent.
- c) The detector should be able to locate the alteration made to an image.
- d) The watermark should be detectable through the correct key.
- e) The marking key should be difficult to be extracted from the marked image without the correct key.

The technique [15] based on LSB substitution are preferred for watermark insertion. The cause of preference of lowest order bit modification is due to its visual insignificance. LSB Substitution method is easy but it has some drawbacks. These methods are not robust to noise and lossy compression. An even better attack would be to simply set all the LSB bits to “1” fully defeating the watermark with negligible impact on the original image data.

The basic LSB substitution has be improved by using a pseudo random number generator to determine the pixel to be used for embedding based on a given key. After embedding watermark, the watermark is recovered using PN sequence and watermark location [4].

In the correlation based watermarking techniques, embedding is performed by adding a PN pattern to the luminance value of the pixels. These patterns consist of the integers {-1, 0, 1} where the energy is almost uniformly distributed. The patterns do not correlate with the host image contents.

Despite the several advantages of watermarking techniques in spatial domain, it has limitation of robustness. The watermarking in spatial domain is less resilient to common

image processing operations. Therefore it is necessity to make a watermark more secure and resistant to common image processing operations in spatial domain.

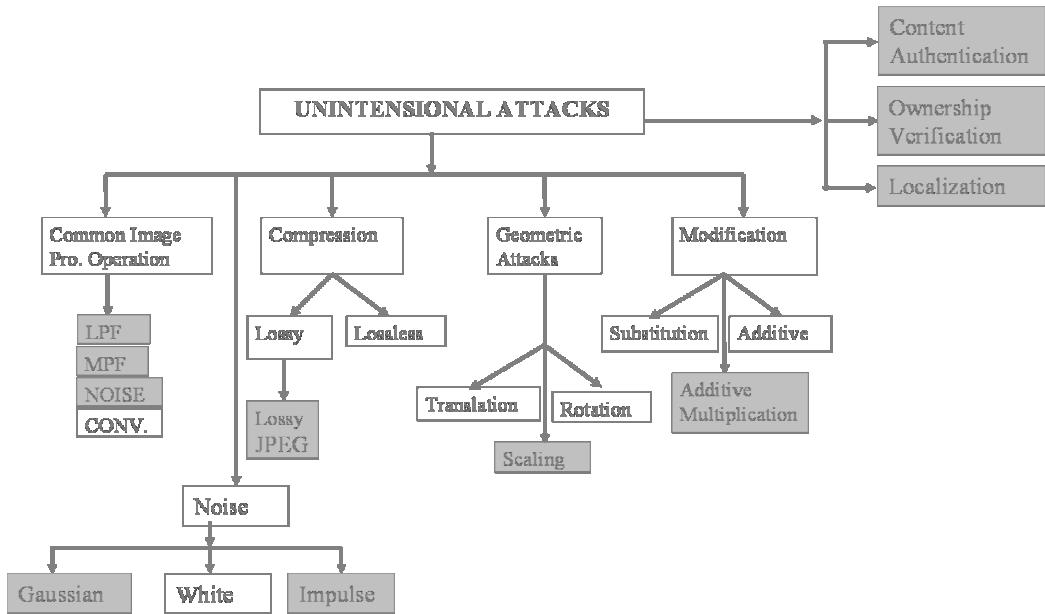


Figure 4. Types of Unintentional Attacks and its focused area (Gray colored boxes indicate the extensive research for robust watermarking system).

4.2. Techniques for Images in Transform Domain

The limitations of the watermarking in spatial domain are addressed in transform domain. The aim of the Digital Watermarking in transform domain is to insert the maximum possible watermark signal without perceptually affecting image quality, so that the watermark must remain present as imperceptible and robust. In transform domain techniques, the watermark is embedded in the transformed coefficients according to the perceptual significance of the transform coefficient. Therefore, the watermark is irregularly distributed over the entire image. After inverse transform, the watermarked image becomes perceptually identical to the original image. Watermarking in an appropriate transform domain improves the capacity and robustness to geometric distortions and JPEG compression.

The work in describes the human visual system for image and video applications and particular in the area of source coding or compression. The three properties of the human visual system are described for image coding are frequency sensitivity; luminance sensitivity; and contrast masking.

Once the image characteristics are identified which satisfies HVS, then the watermark is embedded in transform coefficient according to the perceptual significance of the coefficient in the transform domain. A number of methods [16, 17] have been described to embed the watermark data in block based transform domain. The method in [18] proposed the embedding of the watermark into selective modified middle frequencies of Discrete Cosine Transform (DCT) coefficients of original image. The technique in [19] embeds the watermark

using the Discrete Fourier Transform (DFT) that is invariant to image manipulation due to rotation, scaling and translation. Discrete Wavelet Transform (DWT) is used to embed the watermark in frequency domain [20]. The just noticeable distortion (JND) feature of HVS is applied in [21] and uses DCT. The embedding techniques are described in Discrete Hartley Transform (DHT) with features of HVS. However, these techniques have certain limitations. Many of these techniques are not robust in spite being in consonance with the HVS system due to their embedding procedure. Embedding in block based transform domain coefficients leads to these perceptual problems as each block has different genre and perceptual properties. Watermark insertion in an appropriate transform domain improves the capacity and robustness to geometric distortion and JPEG compression.

5. Observations

The following observations and suggestions have been made to design any watermarking technique for invisible, robust and secure watermarks, useful for content authentication and copyright protection

- The difference between the original image and the embedded watermarked image should be perceptually invisible.
- An invisible watermark must be undetectable by an unauthorized user.
- The owner of the image is the only one who holds the secret key.
- The insertion of invisible watermark should not degrade the quality of original image.
- Insertion of watermark requires minimum human intervention.
- Extracted watermark must identical to the original one.
- For high quality images the amount of individual pixel modification should be as small as possible.
- Common signal processing operations should not make any effect on watermark signals.

The design of watermarking technique is based on the application area. However fragile watermarking is suggested to design in spatial processing domain and make it useful for content authentication and protection of forgery, tempering. Where as robust watermarking techniques are advisable to design in transform processing domain and make it useful for copy and copyright protections.

6. Conclusion

Digital watermarking is an effective technique for embedding a discerning signature in digital multimedia data that remains in the data while it is being consumed or manipulated. It provides security and protection of IPR as it can be proved to exist in the data through its retrieval. It is expected to have huge commercial potential when it gets widely deployed in consumer electronic devices.

Digital watermarking technology and its applications expected to have huge potential in consumer electronics industry. Digital watermark technology is in use with consumer electronic devices like digital still camera, digital video camera, set-top box (STB), DVD players, MP3 players, etc., for various applications like providing controlled access, preventing illegal replication and watermark embedding. Digital watermarking technology has the ability to provide the solutions for the protection of digital contents transmission in various fields like: medical, insurance, banking, finance, entertainment industry, e-commerce and e-governance.\

References

- [1] William Stallings, "Cryptography and Network Security: Principles and Practice", *third edition*, Pearson Education India, 2003, ISBN: 8178089025.
- [2] Bruce Schneier, "Applied Cryptography: Protocols, Algorithms, and Source code in C", Second ed., John wiley and sons Inc., USA, 1996, ISBN: 0-471-12845-7.
- [3] Menezes, A; Oorschot, P; Vanstone, S. "Handbook of Applied cryptography", CRC Press, Florida, USA, 1997.
- [4] Swanson, MD; Kobayashi, M; Tewfik, AH. "Multimedia data-embedding and watermarking technology", Proceeding of IEEE Conference, June, 1998, vol. 86, page 1064-1087.
- [5] Podilchuk, C; Zeng, W. "Image-adaptive watermarking using visual models", *IEEE Journal On Selected Areas in Commun.*, May, 1998, vol. 16, page 525-539.
- [6] Cox. IJ; Miller, ML; Bloom, JA. "Digital watermarking", Publisher: Morgan kaufffuan, 2002, ISBN 1-55860-714-5.
- [7] Young-Won Kim and Il-Seok Oh, "A survey on text watermarking techniques", Proceeding of Conference, Korea Information Science Society, Korea, Aug., 2002, vol.14, no.1, page 34-39.
- [8] Braudway, G; Magerlein, KA; Mintzer, F. "Protecting publicly available images with a visible image watermark", Proceeding of International Conference on Electronic Imaging, *SPIE*, Feb., 1996, vol. 2659, page 126-133.
- [9] Craver, S; Memon, N; Yeo, B. "Resolving Rightful Ownerships with Invisible Watermarking Techniques: Limitations, Attacks and Implications", *IEEE Journal on Selected Areas in Communications*, May 1988, vol 16, no 4, page 573-586.
- [10] kutter, M; Voloshynovskiy, S; Herrigel, A. "The watermark copy attack", Proceeding of Security and Watermarking of Multimedia Contents II, *SPIE*, San Jose, USA, Jan., 2000, vol. 3971, page 371-380.
- [11] Voloshynovskiy, S. "Generalized watermark attack based on watermark estimation and perceptual remodulation", *Proceeding of IS&T/SPIE, 12th Annual Symposia on Electronic Imaging*, San Jose, CA, 2000, vol. 3971, page 358-70.
- [12] Li, YN; Li, CH. "Robust image watermarking algorithm based on predictive vector quantization", *First International Conference on innovative computing, information and control, ICICIC-06*, China, 2006, vol. 6, page 491-494.
- [13] Barreto, PSLM; Kim, HY. V-Rigmen, "Toward seeure public key blockwise fragile authentication watermarking", *Proc of IEEE Magz.*, April, 2002, vol. 149, no. 2, page. 57-62.

- [14] Lelik, MU; Sharma, G; E-Saber, Tekalp, AM. "Hierarchical watermarking for secure image authentative with localization", *IEEE Transaction on Image processing*, Jun., 2002, vol. 11, no. 6, page 585-595,
- [15] Johnson, NF; Katezenbeisser, SC. "A survey of steganographic Techniques", *Information Technique for teganography and digital watermarking, S. C. Katezenbeisser et al.*, Eds. Northwood, MA: Artec House, Dec. 1999, page 43-75.
- [16] Piva, A; Barni, M; Bartolini, F; Cappellini, V. "DCT-based watermark recovering without resorting to the uncorrupted original image," Proceeding of IEEE Int. Conference on Image Processing,, *ICIP 97*, Santa Barbara, CA, , Oct. 1997, page 520-527.
- [17] Podilchuk, C; Zeng, W. "Perceptual watermarking of still images", Proceeding of 1st IEEE Workshop on Multimedia Signal Processing", Princeton, NJ, June, 1997, page 363-368.
- [18] Hsu, CT; Wu, JL. "Hidden digital watermarks in images", *IEEE Transaction on Image Processing*, Jan. 1999, vol. 8, page 58-68.
- [19] Joseph, JK; O' Ruanaidh, Pun, T. "Rotation, Scale and Translation Invariant Digital Image Watermarking", *IEEE Transaction on Signal Processing*, 1998, vol. 66, no. 3, page 303-317.
- [20] Dugad, R; Ratakonda, K; Ahuja, N. "A new wavelet-base for watermarking image", *Proceeding of International Conference on Image Processing*, 1998, vol. 2, page 2119-2123.
- [21] Cox, IJ; Miller, ML. "A review of watermarking and the importance of perceptual modeling", Proceeding of SPIE Electronic Imaging: Storage and Retrieval for Image and Video Databases V, San Jose, CA, Feb. 1997.

Chapter 11

PATHWAY SEARCH ENGINE FOR EXPRESSION PROTEOMICS

Consuelo Marín Vicente, David M. Good and Roman A. Zubarev*

Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Stockholm, Sweden

Abstract

Proteomics is a high-throughput technology for obtaining information on the identity and the expression levels of proteins in a biological sample. Modern mass spectrometry (MS) combined with liquid chromatography (LC) now routinely yields data on >1000 proteins per hour of analysis. However, interpretation of this high-throughput information has until recently been performed in a reductionist way, with emphasis on a few regulated proteins. Increasingly, expression proteomics data are being interpreted by hypothesis-free analyses of activation levels of signalling pathways. The bioinformatics tool performing this pathway analysis was named Pathway Search Engine (PSE). PSE is a hypothesis-generating tool whose predictions are to be tested and validated by complementary (non-mass-spectrometric) techniques. Typically, the PSE consists of a mass spectrometry data analysis module that converts raw LC-MS data into protein identities and their abundances, and a key node analysis module that maps the proteins onto known signalling pathways, performing an upstream search and assigning each identified regulatory molecule (“key node”) a preliminary score, with the pathway score being the sum of key node scores. Finally, the post-processing module performs statistical analysis of the group of identified key nodes or pathways and determines the degree of activation for each, as well as providing statistical significance through computation of the associated p-value.

Introduction

Mass spectrometry-based proteomics is a large-scale proteome analysis commonly used to study different biological samples. With modern instrumentation, single analysis of a full proteome sample provides a large volume of information. For instance, a 2 hour long LC-MS

* E-mail address: Roman.Zubarev@ki.se. Prof. Roman A. Zubarev, Head of Molecular Biometry, Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Scheelles väg 2, A3:5, Stockholm, Sweden, 171 77.

run can yield qualitative (ID) and quantitative (abundance) composition of > 2000 proteins. For a typical human tissue sample, this number represents 20-40% of the total expressed proteome. Comparative proteomics attempts to identify differences between the proteomes of a case and control, *i.e.* two biological systems in different states. Because of the high proteome plasticity, within hours of affecting or stimulating cells, hundreds of proteins can exhibit differences in their relative abundances. For instance, in a recent seminal paper, all 1260 proteins in a cancer cell line treated by a drug have changed their expression levels significantly (*i.e.* by a factor of 2 or higher) over a 48 h period [1]. This is why the old reductionist paradigm that focuses on one or few “key” proteins is to be less and less relied upon in proteomics. The one-by-one consideration of protein expression levels is currently being replaced by a new type of analysis that reduces the number of independent entities to a manageable level. Since proteins are not produced by a cell in an individual manner but expressed collectively, patterns of simultaneously expressed proteins can be used as such entities. The proteins linked into these patterns usually belong to the same pathway, either signaling or metabolic. The identification of activated pathways in a case *versus* control analysis is an important element of the new type of data processing called Pathway Analysis (PA). PA is a method of rationalizing biological knowledge that attempts to go far beyond heat maps and gene ontology classification [2]. The pathway identification is performed by a software tool known as the Pathway Search Engine (PSE) [3]. PSE also reveals the key regulatory elements of the identified pathway, and these “key nodes” can be affected (*e.g.* by drugs) to switch on or off the flow of information along a given branch of the pathway. One should remember that the PSE output is by necessity a prediction rather than the ultimate truth, and as every prediction it must be validated by orthogonal means, *i.e.* by methods based on physically different processes than those used to generate the predictions themselves. This validation can be performed by biochemistry, *e.g.* as Western blots, immunofluorescence, or functional assays. Despite the hypothetical nature of the PSE output, it is important to provide a quantitative element to it, *i.e.* to supplement each identified activated pathway and/or key node by a factor representing the degree of activation. This brings about the necessity of introducing universal units of key node activation, as the pathway activation score can be a combination of the activation scores of all involved key nodes. Another challenge is to provide realistic p-values, *i.e.* assess the probability that the key node or pathway is activated at all. The degree of activation and the p-values are linked only indirectly, as the former are objective parameters of the studied biological system while the latter are largely determined by statistical variation of the experimental data.

In order to provide reliable, unbiased pathway analysis, a PSE is required to integrate three elements: i) the pathway database, ii) the mapping tool converting the input data (protein IDs and their abundances) into the elements of pathways, and iii) the scoring system for the output of pathway mapping [4]. The scoring system, which is the heart of any search engine, is designed to recognize differences and provide good separation between heterogeneous samples (*e.g.* case-control) and homogeneous ones (*e.g.* case-case or control-control). At the same time, the scoring system must be robust in relation to statistical noise always present in the data. Below we review how these challenges are met by current efforts.

Input data for PSE. The first step of pathway analysis is the generation of quantitative proteomics data. This is currently performed by two methods: 1) relative quantification, and 2) “absolute” quantification. In the first approach, proteomes or proteome digests of the

appropriately labeled Case and Control are mixed together and then subjected to LC/MS/MS, with each peptide identified. Quantification is typically performed by measuring the ratio of the abundances of two differently labeled but otherwise identical peptides, one coming from the Case proteome and another one from Control (Figure 1A). Labeling is performed either *in vivo* as in SILAC[5] or *in vitro* as in ICAT, iTRAQ, and a number of similar techniques [6, 7]. Using appropriate chemistry, the multiplexing degree is extended from two to eight [7]. The common denominator in these approaches is that the output is normalized in such a way that the abundances of all Control peptides (or proteins) are set to unity. The same normalization was employed in work [1] where quantification was performed by fluorescence microscopy.

Figure 1A

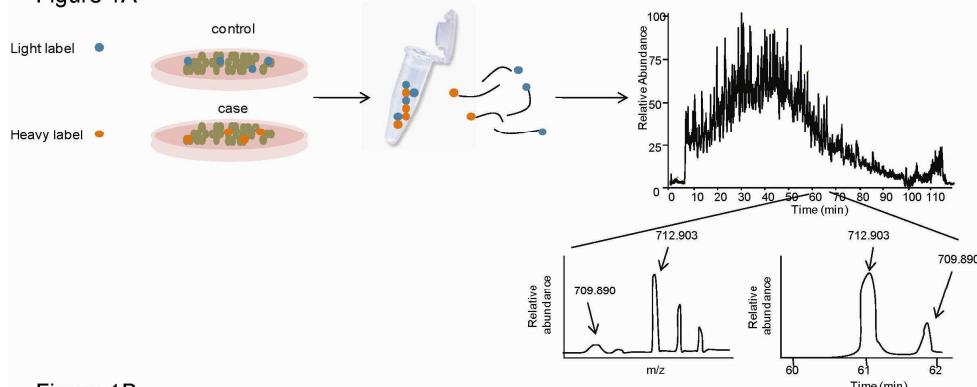


Figure 1B

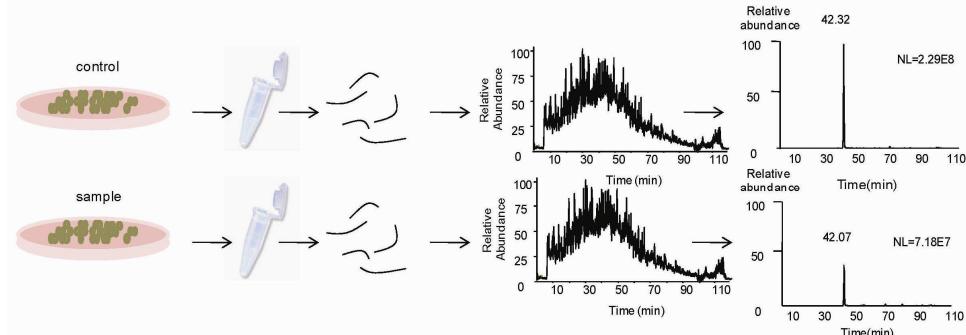


Figure 1. A) Label-based proteomic assay. The cellular samples are *in vivo* or *in vitro* labeled by isotopically different chemical groups. The strategy is shown, in which the Case and Control cells are labeled *in vivo* by isotopically different amino acids (as e.g. in the SILAC approach [5]). The Case and Control cells are mixed, the proteome is extracted and digested with trypsin, after which the peptides are analyzed by mass spectrometry. The peptides with isotopically different amino acids elute in LC at the same retention time, but the molecular ions possess different masses, and their abundances are measured to provide quantification. B) Label-free proteomic assay. Case and Control cells are extracted and digested independently and analyzed by mass spectrometry separately. The ratio of the areas under the chromatographic peaks of for the same identified peptide is considered for quantification.

The alternative method is the so called label-free approach (Figure 1B), in which Case and Control samples are run separately in unadulterated form, and quantification is performed by calculating the area under the chromatographic peak of each identified peptide. This

calculation includes all charge states and the whole isotopic distribution (*i.e.* performing charge, isotope, and chromatographic peak deconvolutions is required). Assuming similar ionization efficiencies for the most abundant peptides of each protein (a valid assignment for many proteotypic peptides [8]), protein abundance is taken to be the sum of the peptide abundances. This approach gives not only the Case/Control ratios, but also the abundance values for each protein. These values are proportional to the protein concentrations. Since the proportionality factor usually remains unknown, here we called this method “absolute” quantification, as opposed to truly absolute measurements that yield the concentrations. To make Case and Control measurements comparable, the total protein content in both samples are normalized by the same value. Usually, label-based relative measurements provide smaller experimental uncertainties of the Case/Control ratios and can reduce the overall experimental time, but the “absolute” measurements give additional dimension to measurements in the form of abundance of each individual protein, although determined with some uncertainty.

Often in literature, the list of detected proteins is filtered to leave only statistically significant up- and down-regulated molecules. Such an approach has two disadvantages. First, statistical significance can be determined by experimental uncertainties if the latter are larger than biological variations, which is often the case with cell lines. Therefore, valuable signal is inevitably lost or reduced when regulated proteins are removed from the list because their abundance measurements fail to meet the significance criterion. Second, the criterion for regulation is itself poorly defined in quantitative terms. There are still debates whether a certain fixed factor, *e.g.* 2, can be used as a threshold for fold-regulation. Our opinion is that for an unbiased pathway analysis, *all* protein abundances must be used, and the statistical significance test is better performed on the final result (pathway or key-node) instead of the input protein abundance data. Therefore, a typical input for pathway analysis consists of a list of protein IDs, common for both Case and Control, and two lists of the respective abundances, one for Case and another for Control (Figure 2).

Mapping on Pathways. A number of different pathway databases are currently available to researchers, including public efforts [9-12]. A number of commercial companies, *e.g.* Ingenuity, Ariadna Technologies, BioBase, *etc*; also provide such databases. The TRANSPATH database from BioBase currently encompasses 169 major signaling pathways and thousands of sub-pathways. These pathways and sub-pathways are composed from extensive knowledge of genes, molecules, and their interactions extracted from many thousands of diverse biomedical publications [13].

In principle, mapping a list of proteins on pathways is a straightforward task, and “direct mapping” can be performed by software tools supplementing all pathway databases. In this analysis, the pathway encompassing the highest number of regulated proteins (“hits”) wins (Figure 3). Sometimes the number of hits is normalized by the size of the pathway. Direct mapping, as discussed in reference [3] is a straightforward approach that appeals by its simplicity, but it sometimes fails to correctly identify the relevant pathway. Because of the branching nature of signaling pathways, the detected protein abundance changes are separated in time and space (in terms of the network graph) from the regulatory elements of the signaling pathway that triggered these changes. Therefore, relevant pathway analysis should perform a sort of (time, space)-deconvolution to pinpoint the correct cause for the proteome alteration. Such deconvolution is performed in the form of key node analysis, in which an algorithm searches the network upstream from each of the detected proteins to arrive to the

closest “key node”. Key nodes are regulatory molecules in the pathway, also known as bottleneck molecules, as they are typically found in intersections of the pathways [3, 13, 14, 15]. Each found key node receives a score reflecting the number of proteins from the input list linked to it, the distance to each of these proteins (longer distances give lower score), as well as the abundance of these proteins. In TRANSPATH, this task is performed by a tool called ExPlain [13, 14]. The output of the key-node analysis is typically a list of key nodes, common for Case and Control, and two lists of scores, individual for Case and Control.

Key-node score post-processing. By comparing the score differences between Case and Control for the same key nodes, one can reveal which key nodes are activated. Mapping of these key nodes on the pathways will correctly reveal the activated pathway. However, the task of key node score comparisons is still a matter of research. The first question to address is whether score ratios or score differences should be considered. In mRNA array analysis, it is customary to consider ratios [16]. The same approach is widely used in proteomics [17, 18, 19]. However, large fold-changes of low-scoring key node may be less relevant biologically than moderate fold-changes of a highly-scoring key node. Besides, low-scoring key nodes may be subject to statistical variations if their scores result from few proximal, low-abundant proteins (as opposed to highly abundant but distant proteins). Thus, absolute score difference captures a different dimension of pathway activation than fold changes do. Previously, score differences are shown to possess significant predictive power in pathway analysis [20]. But unlike fold changes that are unitless, absolute score differences require introduction of units. These units should be universal enough to provide unbiased activation degree for different Case-Control comparisons. Universal score units are a major issue that have so far not been solved completely satisfactory.

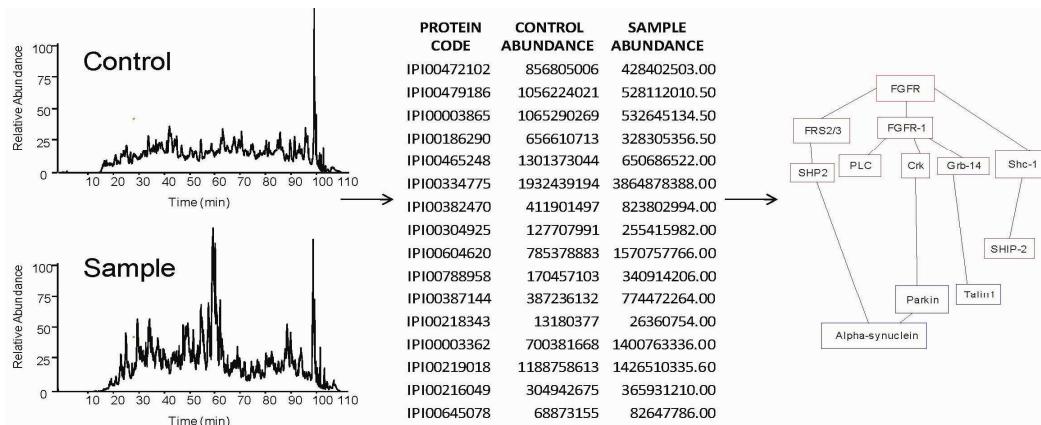


Figure 2. Generic pathway analysis workflow. After analyzing by LC/MS the Case and Control proteomes, relative abundance of each protein is determined. The proteins are mapped onto known pathways using available databases of protein-protein and protein-molecule interactions. Upon scoring each candidate pathway, the highest-scoring pathway is selected. Intermediate steps, such a key node analysis, and the final step of statistical evaluation are explained in the text.

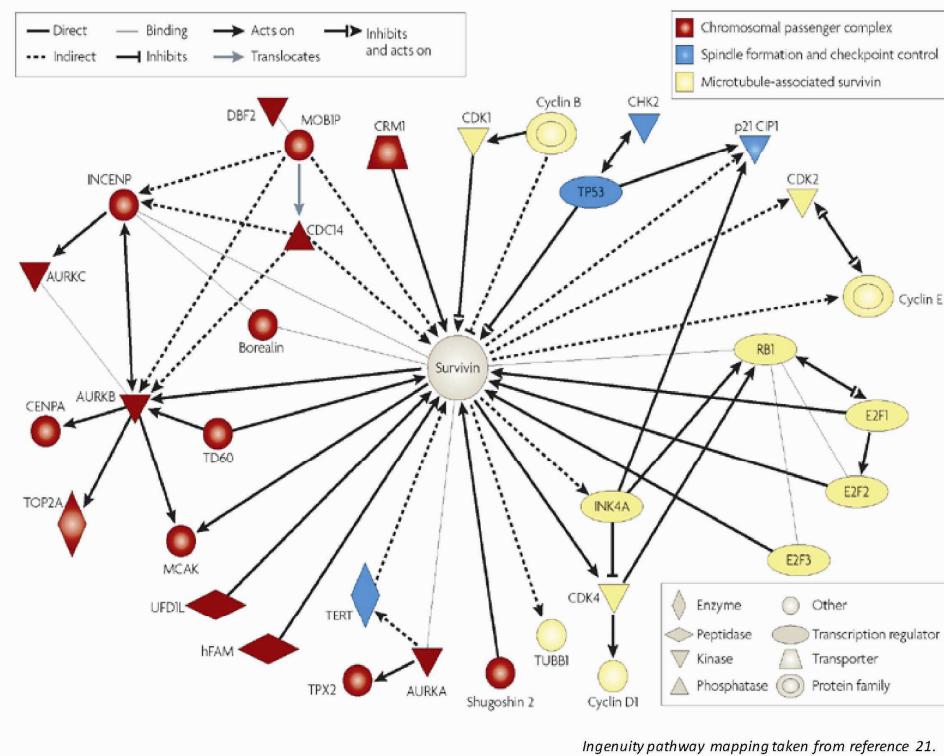


Figure 3. An example of direct mapping of “regulated” proteins into a pathway database. The pathway accommodating the largest number of these molecules is chosen as the answer. Adopted from ref. 21. The deficiencies of usage of only “regulated” molecules for pathway analysis and direct mapping are discussed in the text.

The final part of post-processing is statistical evaluation. One evaluation method is to get for each key node a p-value, *i.e.* the *a priori* probability of this key node to be found activated by pure chance. Providing realistic p-values is a non-trivial task, as experimentally obtained distributions of fold-changes tend to be non-Gaussian [16]. One useful approach is to employ a decoy dataset derived from *e.g.* a comparison of two biological replicates of the same nature (Case-Case or Control-Control). The distribution of score differences (or fold-changes) from the decoy can help in estimation of the false discovery rate (FDR), which is an alternative to the p-value evaluation parameter. The threshold for FDR is usually taken at 0.05.

In many cases key node analysis is the end point of the pathway analysis routine. Since many classical signaling pathways overlap to a significant degree and cross-talk between pathways is not uncommon, a list of activated key nodes often provides better representation of the biological process under investigation than a list of activated pathways. Yet if the latter is desired, key nodes can be directly mapped onto pathways, with key node scores serving as weighting factors. The pathway score can be defined as a sum of the scores of constituent key nodes.

Validation. Since PSE predictions are of hypothetical nature, they must be validated. As a good practice, validation should be performed by a complementary technique to the one used

for prediction generation (*i.e.* mass spectrometry). Good validation tools include immunochemistry, fluorescent microscopy, *etc.* [15].

Conclusion

PSE is a powerful emerging tool that extends proteomics analysis way beyond the reductionist's focus on a few molecules or genes, and gives a more complete picture of the biological processes under study. Equally important is that PSE utilizes for its predictions *all information* available to date in the form of signaling pathway databases. Pathway analysis is not, however, the final step in investigation of biological samples. The key node regulation revealed by PSE can be used to build predictive quantitative models of diseases and other biological processes.

Despite recent progress, much needs to be done to develop a fully capable PSE. One of the unresolved issues is the universal units of key node activation. Another issue is the statistical model of the pathway analysis process that would allow for determination of realistic p-values. Finally, there is a great need for more quantitative proteomics datasets to be made publically available, for testing and fine tuning of next-generation pathway search engines.

Acknowledgments

This work was supported by the Knut and Alice Wallenberg Foundation, European Union (consortium PredictAD) as well as the Swedish research council (grant 621-2007-4410). CMV is supported by a postdoctoral grant from the Spanish Ministry of Science and Innovation (MCINN). DMG is a recipient of a Wenner-Gren postdoctoral fellowship (2010).

References

- [1] Cohen, AA; Geva-Zatorsky, N; Eden, E; Frenkel-Morgenstern, M; Issaeva, I; Sigal, A; Milo, R; Cohen-Saidon, C; Liron, Y; Kam, Z; Cohen, L; Danon, T; Perzov, N; Alon, U. Dynamic proteomics of individual cancer cells in response to a drug. *Science.*, (2008), 322, 1511-1516.
- [2] Schilling, CH; Schuster, S; Palsson, BO; Heinrich, R. Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol. Prog.*, 1999, 15, 296-303.
- [3] Zubarev, RA; Nielsen, ML; Fung, EM; Savitski, MM; Kel-Margoulis, O; Wingender, E; Kel, A. Identification of dominant signaling pathways from proteomics expression data. *J. Proteomics.*, 2008, 71(1), 89-96.
- [4] Marin-Vicente, C; Zubarev, RA. Search engine for proteomics: Fact or fiction. *G.I.T. Laboratory journal Europe.*, 2009, 13, 10-11.
- [5] Ong, S; Blagoev, B; Kratchmarova, I; Kristensen, D; Steen, H; Pandey, A; Mann, M. Stable Isotope Labeling by Amino Acids in Cell Culture, SILAC, as a simple and accurate approach to expression proteomics. *Mol. Cell. Proteomics.*, 2002, 5, 376-386.

- [6] Gygi, SP; Rist, B; Gerber, SA; Turecek, F; Gelb, MH; Aebersold, R. Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nat. Biotechnol.*, 1999, 17, 994-999.
- [7] Ross, PL; Huang, YN; Marchese, JN; Williamson, B; Parker, K; Hattan, S; Khainovski, N; Pillai, S; Dey, S; Daniels, S; Purkayastha, S; Juhasz, P; Martin, S; Bartlet-Jones, M; He, F; Jacobson, A; Pappin, DJ. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using aminereactive isobaric tagging reagents. *Mol. Cell. Proteomics.*, 2004, 3, 1154-1169.
- [8] Ono, M; Shitashige, M; Honda, K; Isobe, T; Kuwabara, H; Matsuzaki, H; Hirohashi, S; Yamada, T. Label-free quantitative proteomics using large peptide data sets generated by nanoflow liquid chromatography and mass spectrometry. *Mol. Cell. Proteomics.*, 2006, 5(7), 1338-47.
- [9] Krieger, CJ; Zhang, P; Mueller, LA; Wang, A; Paley, S; Arnaud, M; Pick, J; Rhee, SY. and Karp, PD. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, 2004, 32, D438-D442.
- [10] Chowbina, SR; Wu, X; Zhang, F; Li, PM; Pandey, R; Kasamsetty, HN; Chen, JY. HPD. an online integrated human pathway database enabling systems biology studies. *BMC Bioinformatics.*, 2009, 8, 10.
- [11] Kono, N; Arakawa, K; Ogawa, R; Kido, N; Oshita, K; Ikegami, K; Tamaki, S; Tomita, M. Pathway projector: web-based zoomable pathway browser using KEGG atlas and Google Maps API. *PLoS One.*, 2009, 4(11), e7710.
- [12] Geer, LY; Marchler-Bauer, A; Geer, RC; Han, L; He, J; He, S; Liu, C; Shi, W; Bryant, SH. The NCBI BioSystems database. *Nucleic Acids Res.* 2009; Oct 23. [Epub ahead of print].
- [13] Krull, M; Pistor, S; Voss, N; Kel, A; Reuter, I; Kronenberg, D; Michael, H; Schwarzer, K; Potapov, A; Choi, C; Kel-Margoulis, O; Wingender, E. TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.*, 2006, 34, D546-51.
- [14] Kel, A; Voss, N; Jauregui, R; Kel-Margoulis, O. & Wingender, E. Beyond microarrays: Finding key transcription factors controlling signal transduction pathways. *BMC Bioinformatics.*, 2006, 7 (Suppl. 2):S13.
- [15] Ståhl, S; Fung, E; Adams, C; Lengqvist, J; Mörk, B; Stenerlöw, B; Lewensohn, R; Lehtio, J; Zubarev, R; Viktorsson, K. Proteomics and pathway analysis identifies JNK signaling as critical for high linear energy transfer radiation-induced apoptosis in non-small lung cancer cells. *Mol. Cell. Proteomics.*, 2009, 8(5), 1117-29.
- [16] Brody, JP; Williams, BA; Wold, BJ; Quake, SR. Significance and statistical errors in the analysis of DNA microarray data. *PNAS*. 2002, 99(20), 12975-12978.
- [17] Koehler, CJ; Strozyński, M; Kozielski, F; Treumann, A; Thiede, B. Isobaric peptide termini labeling for MS/MS-based quantitative proteomics. *J. Proteome Res.*, 2009, 8(9), 4333-41.
- [18] Bouyssié, D; Gonzalez de Peredo, A; Mouton, E; Albigot, R; Roussel, L; Ortega, N; Cayrol, C; Burlet-Schiltz, O; Girard, JP; Monsarrat, B. Mascot file parsing and quantification (MFPaQ), a new software to parse, validate, and quantify proteomics data generated by ICAT and SILAC mass spectrometric analyses: application to the proteomics study of membrane proteins from primary human endothelial cells. *Mol. Cell. Proteomics.*, 2007, 6(9), 1621-37.

- [19] Pan, C; Kora, G; Tabb, DL; Pelletier, DA; McDonald, WH; Hurst, GB; Hettich, RL; Samatova, NF. Robust estimation of peptide abundance ratios and rigorous scoring of their variability and bias in quantitative shotgun proteomics. *Anal. Chem.*, 2006, 78(20), 7110-20.
- [20] Willingale, R; Jones, DJ; Lamb, JH; Quinn, P; Farmer, PB; Ng, LL. Searching for biomarkers of heart failure in the mass spectra of blood plasma. *Proteomics.*, 2006, 6(22), 5903-14.
- [21] Altiori, DC. Survivin, cancer networks and pathway-directed drug discovery. *Nat. Rev. Cancer.*, 2008, 8(1), 61-70.

Chapter 12

BIOMEDICAL LITERATURE ANALYSIS: CURRENT STATE AND CHALLENGES

Maurice H.T. Ling^{1,2}, Christophe Lefevre³ and Kevin R. Nicholas^{2,3}

¹School of Chemical and Life Sciences, Singapore Polytechnic, Singapore

²Department of Zoology, The University of Melbourne, Australia

³Institute for Technology Research and Innovation, Deakin University, Australia

Abstract

Advances in molecular biology tools and techniques from the end of the last century had shifted the focus of biomedical research from the study of individual proteins and genes to the interactions within an entire biological systems. At the same time, advanced tools generates large sets of experimental data which required collaborations of groups of biologists to decipher. This resulted in a need to have a diverse research knowledge. However, the amount of published research information in the form of published articles is increasing exponentially, making it difficult to maintain a productive edge. Biomedical literature analysis is seen as a means to manage the increased amount of information – to gather relevant articles and extract relevant information from these articles. We review the central (information retrieval, information extraction and text mining) and allied (corpus collection, databases and system evaluation methods) domains of computational biomedical literature analysis to present the current state of biomedical literature analysis for protein-protein and protein-gene interactions and the challenges ahead.

1. Introduction

With rising emphasis in genomics, transcriptomics and proteomics from the end of the last century, the focus of biomedical research is shifting from the study of individual proteins and genes to entire biological systems, such as tissues or whole organisms. Experimental techniques used to study them, like mass spectrometry and microarrays, often generates large data sets. A group of biologists must then collaborate to make sense of this large set of experimental data, and often requires connections with research areas outside their own core competencies, which exists as published literature in various research areas. This has resulted

in a need to be versed in research areas other than the researcher's own specialty. In addition, the amount of information, in the form of published articles, is increasing exponentially, making it difficult for a researcher to keep abreast with relevant literature manually [1], even on specialized topics.

Due to these changes, literature processing tools are becoming essential to researchers [2] as it was estimated that only about 20% of biological knowledge exist in structured formats, such as in databases, while the remaining 80% are in natural language document [1, 3]. They include targeting relevant papers, known as information retrieval; identifying gene or protein or chemical entities; identifying abbreviations; extracting facts from the literature, known as information extraction; and in some instances, generating hypotheses. This review shall briefly examine the historical roots and current state of biomedical literature analysis. The computational procedures of 30 systems will be briefly described to illustrate some of the current methods used, and its related areas of importance before defining the objectives and organization of this thesis.

2. Brief History of Biomedical Literature Analysis

Don Swanson initiated interest in biomedical literature analysis by analyzing publications semi-automatically and suggested links between separate areas of research, such as fish oil and Raynaud's syndrome [4], migraine and magnesium [5] in the mid-1980s.

At around the same time, the First Message Understanding Conference (MUC-1) was held in 1987, which explored formats for recording information in documents. In 1989, MUC-2 concentrated on template filling of information and formulated the details of precision and recall measures, which is still in use today. MUC-3 (1991) and MUC-4 (1992) were centered on compiling and completing template from information in terrorist reports, and therefore, had no direct bioinformatics relevance but benefited improved techniques.

In 1992, the First Text Retrieval Conference (TREC-1) was initiated by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense and used the idea of challenge evaluation tasks to tease out the state of the art of that time. Both TREC-1 and MUC-5 in 1993 were greatly influenced by the Tipster program (a U.S. Government program) which emphasized on evaluation-driven research [6], setting the tone for future conferences. TREC-2 in 1993 was critical for providing the baseline performance for main tasks, which was then expanded to having different tracks for various tasks in future TRECs. Each track in TREC was started based on interests and notably, a genomic track was started in 2003 [7] with a subsequent TREC in 2006. In short, MUC and TREC conferences had significantly advanced the field of information retrieval and extraction by their challenge tasks due to rigorous use of systematic common evaluations [8].

The earliest work in text mining for genomics was by Timothy Leek [9]. Fukuda et al. [10] pioneered protein name recognition (named entity recognition) in text in the 3rd Pacific Symposium on Biocomputing. Craven and Kumlein [11] and Blaschke et al. [12] independently published the first work on recognition of relationships between entities (proteins, genes, and small molecules). By 2000, the focus had shifted to the recognition of relationships between entities (proteins, genes, and small molecules), with Shatkay and Wilbur [13] and GENIES [14] as one of the first systems. Following GENIES, the field of biomedical literature analysis for information retrieval and information extraction was

extremely active with numerous systems being developed (reviewed in later sections). The Pacific Symposium on Biocomputing between the years 2001 and 2004 included predominantly presentations on various aspects of biomedical literature analysis and other related conferences, such as Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), saw an increase in the number of similar presentations. Specific workshops for biomedical literature analysis, such as BioLink 2004, were run in that period. It could be said that the years between 1998 and 2004 were the golden age of biomedical literature analysis.

By the end of 2004, the general emphasis of biomedical literature analysis had diminished at these conferences. This might be due to the asymptotic performance of various systems and a serious lack of benchmark data, such as biomedical corpora tagged for various purposes [15]. This might also meant that conventional means of analysis and technology transfer from more traditional fields, such as computational linguistics for understanding general text, had reached its maximum. The changes were more likely due to the rising interest in other fields, such as microarray and sequence analyses. The Fourth Asia Pacific Bioinformatics Conference (APBC 2006) saw more than half of the posters in the area of microarray and sequence analyses while only about 5% in biomedical literature analysis. However, the golden age of biomedical literature analysis of 1998 and 2004 had left us with preliminary resources and techniques that could be collated for other purposes.

Interestingly, although the field of biomedical literature analysis arose from Don Swanson's work in the mid-1980s [4, 5], hypothesis generation (text mining) was not adopted into mainstream biomedical literature analysis. This suggested that the field of biomedical knowledge is still largely uncharted with gems for many explorers to find in the years to come. With that optimistic thought, we shall explore the current areas of research in biomedical literature analysis.

3. Current Areas of Research

Although the primary utility of biomedical literature processing is obtaining the relevant research articles (information retrieval; IR), extracting facts (information extraction; IE), and in some instances, drawing new hypotheses (text mining; TM) as shown in Figure 1, there are five concentrations of research efforts instead of the mentioned three. The other two are Named Entity Recognition (NER) and Abbreviation Recognition (AR), which are more domain specific [2]. Comparatively, IR, IE and TM tend to be less domain-specific.

Before focusing on each of the core research areas, it is crucial to understand some commonly used performance measures. Systems are typically measured in terms of precision (number of correct predictions divided by the total number of predictions) and recall (number of correct predictions divide by the total number of correct predictions in the test set) [2]. Precision (P) and recall (R) can be combined into a single F-score, defined as the harmonic mean of precision and recall, $2PR/(P+R)$ [16, 17]. A more extensive treatment of evaluation strategies will be given in Section 1.8.3.

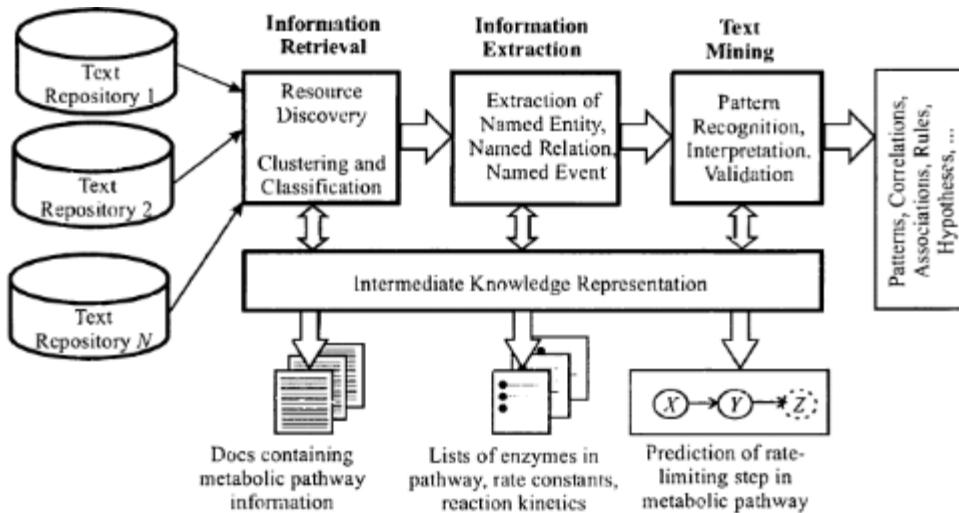


Figure 1. Stages of biomedical literature processing showing examples of possible output from each stage. Adapted from [16].

3.1. Information Retrieval: Finding the Papers

Information retrieval (IR) tools aim to identify text (full text, abstracts, sentences) pertaining to a certain topic of interest, which might be user-defined (ad hoc IR) or pre-defined categories, as in text categorization [18], or a hybrid of both [19]. The best-known biomedical IR system, PubMed, is an ad hoc IR system, which uses Boolean logic and a vector space model, and is available as a programmatic interface, PubMed EUtils.

Both Boolean logic and the vector space model are established IR methodologies [20]. Boolean logic builds on keyword queries where multiple keywords are chained using Boolean operators, such as “(mouse AND cytokines) NOT interferon”. In the vector space model, each document is represented by document-term vector, calculated by a frequency-based weighing scheme, which is then compared to the query vector [21, 22]. This had also been used for thematic analysis of the AIDS literature [23].

Building on PubMed's IR system, more advanced tools like Textpresso [24] and MedMiner [25] employed named entity recognition (see Section 3.3) to identify text for protein and gene names. Shatkay et al. [26] used named entity and terms recognition before presenting the results to a machine-learned classifier to score biomedical text. Results from an IR system are usually presented as a long list which gives a poor overview. Based on the idea of index creation, words that are not commonly used, also known as discriminatory words, had been used by MedlineRanker [27] to rank retrieved citations. Recent developments had attempted to either summarize search results into other representations using Gene Ontology [28, 29] or to represent pair-wise searches as network [30], as it had been shown that biomedical ontologies can significantly enhance the quality of document clustering [31] which can be used as input for text mining. Biological ontologies had also been used to restrict the amount of returned documents in PathBinderH [32] and to assist in document navigation [33]. Query term matching [21] and expansion [34] is an important aspect and a current obstacle in biomedical IR due to multiple synonyms describing the same entity. For

example, 'yeast', also known as 'baker's yeast', can be both '*Saccharomyces cerevisiae*' or '*Schizosaccharomyces pombe*' or their abbreviated forms, '*S. cerevisiae*' and '*S. pombe*' respectively. Although matching of query terms to biomedical vocabularies like Medical Subject Headings [30], UMLS [35, 36] and OWL-DL ontology [37] is possible, the query term expansion is still largely an open problem [38].

3.1.1. Brief Descriptions of Information Retrieval Systems

MedMiner [25] interrogates PubMed and GeneCard. User query of GeneCard database obtained a list of associated genes, which could be filtered with a user defined list of genes of interest and used to query PubMed. Results from PubMed were analyzed at sentence-level and only sentences with at least one gene and keyword remained. Final results were presented to the user after clustering based on rules or keywords.

Textpresso [24] was an example of a system able to extract multiple types of information from biomedical literature [38]. Text was POS tagged with Brill tagger [39] trained on *C. elegans* literature but was not further used. Instead, the same text was tagged by a set of 33 tags, forming the Textpresso Ontology, consisting of 14500 Regular Expressions. Exhaustive indexing on the ontology was carried out to facilitate text retrieval. A further work [19] expanded on Textpresso by automatic clustering of the results into document categories using phrase clustering and support vector machines. Pharmspresso [40] was also built on Textpresso [24] for studying the relationship between genetic variation and the variation in drug response phenotypes. When tested on 45 human-curated abstracts, Pharmspresso [40] identified 78% of target genes, 61% of target polymorphisms, and 74% of target drug concepts.

PathBinderH [32] intersected PubMed search results and NCBI Entrez Taxonomy. Both search results and taxonomical terms were user-defined. Only results that were indexed in the correct taxonomy were returned to the user.

Shatkay et al. [26] used MedPost [41] and YamCha [42] to identify terms in the source text to train a classifier to identify the required articles.

3.2. Abbreviation Recognition: Aliasing the Names

Growth in biomedical terminology parallels the growth of biomedical literature and many of these biomedical terminologies have abbreviations. There are two foreseeable uses in collecting terminologies and their abbreviations. Firstly, abbreviations can aid information retrieval by providing a means to expand search terms to cover terminological variants that have the same abbreviation. Secondly, a dictionary of terminologies and abbreviation may facilitate text processing of multi-word terms as described in the previous section. Abbreviation Recognition (AR) is pair recognition of a terminology (may be a phrase or an entity) and its corresponding abbreviation from free text.

Most of the progress in AR was made between the years 2001 and 2004, and had mostly adopted rules-based techniques with a statistical score. Liu and Friedman [43] were probably the only group that used mainly statistical co-location to determine abbreviations and their phrases. The main drawback of AR by statistics is the need for a large collection of text, known as a corpus (corpora for plural).

Rules-based methods used the knowledge of how abbreviations were being formed and had been well described by Jeffrey Chang (Stanford University's Abbreviation Server) [44, 45], which demonstrated 97% precision at 22% recall (F-score = 0.36) and 95% precision at 75% recall (F-score = 0.88). AbbRE [46] and Schwartz and Hearst [47] both used pattern matching rules and achieved 96% precision with 70% recall (F-score = 0.81), and 96% precision with 82% recall respectively (F-score = 0.88). SaRAD system [48] reported 95% precision with 85% recall (F-score = 0.9). AcroMed [49] reported 97% precision with 72% recall (F-score = 0.83) while ARGH [50] reported 96% precision with 93% recall (F-score = 0.94). The results (abbreviations and their full form) of Stanford University's Abbreviation Server, SaRAD, AcroMed and ARGH were available [51]. Acromine [52] used the observation that abbreviations can usually be expanded to their full form (reversing the idea of reducing full forms into abbreviations) reported precision of 99% with 82% to 95% recall (F-score = 0.89 – 0.97). Sohn et al. [53] uses a similar long-form to abbreviation matching algorithm to AbbRE [46] reported 96.5% precision with 83.2% recall. MBA [54] used text alignment for acronym-type abbreviations and statistics for non-acronym-type abbreviations. It reported 88% recall and 91% precision (F-score = 0.89). BIOADI [55] exploited a set of textual features to describe the properties of potential abbreviation pairs and reported 96% precision with 85% recall (F-score = 0.90). Torii et al. [56] performed a meta-study to compare the results of a number of AR systems and found that they generally agree with each other in terms of results. In addition, Kuo et al. [55] evaluated Sohn et al. [53], Schwartz and Hearst [47] and BIOADI [55] using 2 different corpora and found that the performance of the systems agree with each other in terms of precision, recall and F-score.

3.3. Named Entity Recognition: Identifying the Players

The goal of Named Entity Recognition (NER) is to find entities, the names of physical or abstract objects, in a given text [57]; in particular, the names of chemicals, proteins and genes. Essentially, NER asks the question “What makes a name a name?” This question is generally answered by recognizing words that may refer to entities, followed by identifying the entities in question uniquely. NER is currently one of the most difficult tasks in biomedical text mining [38] and solving this problem will allow for more complex text mining tasks to be addressed [58] as it is a prerequisite for information extraction and advanced IR [38, 59, 60]. NER could also be expanded into recognizing other medically important terms, such as names of diseases [61].

One of the main reasons for the difficulty is the high degree of variations in terms that are not explicitly reflected in biomedical ontologies [62]. It is common that biological entities can have several names, for example, PTEN and MMAC1 refers to the same entity [2]. It was estimated that one-third of biological terms are variants [63]. In addition, biological and chemical entities may have multi-word names and variants of the names. For example, “Peroxisome Proliferator Activated Receptor”, “Peroxisome proliferator-activated receptor” and “Peroxisome-proliferator-activated receptor” refer to the same entity. Liu et al. [64] did a comprehensive study on this area. With new genes and proteins being discovered and named in the genomic era, it can be implied that there is no complete dictionary of named biological entities; hence, NER by simple text matching will not suffice [2, 65]. Due to the potential utility and complexity of the problem, NER has held the attention of many researchers and it

is not surprising that much work in biomedical NER had focused on recognizing gene and protein names in text.

The approaches can be classified as either lexicon-based, rules-based, statistics-based, or combinations of these means. Krauthammer et al. [66] adapted BLAST algorithm to identify entities from text with 78.8% precision and 71.7% recall ($F\text{-score} = 0.75$). A good example of a rules-based NER system is AbGene, which was based on a Brill tagger trained on 7000 manually-tagged sentences using a 'Gene' tag. It achieved a 85.7% precision and 66.7% recall ($F\text{-score} = 0.75$) [67]. In contrast, GAPSCORE, also a rules-based system, examines the appearance, morphology and context of the word before applying a classifier trained using these features. It achieved 74% precision and 81% recall ($F\text{-score} = 0.77$) in inexact matches, and 59% precision and 50% recall ($F\text{-score} = 0.54$) in exact matches [68]. On the other hand, Hanisch et al. [58] used a dictionary approach and achieved 95% precision and 90% recall ($F\text{-score} = 0.92$) while Egorov et al. [69] achieved 98% precision and 88% recall ($F\text{-score} = 0.93$). A further study report using curated dictionary demonstrated good performance, $F\text{-score}$ between 0.8 to 0.9, across different organisms [70]. VTag [71], McDonald and Pereira [72] and ABNER [73] employed conditional random field, a statistical approach and achieved precisions between 58.2% to 85.4% and recall between 53.9% and 79.8%. Zhou et al. [74] combined several approaches using voting strategy to achieve a $F\text{-score}$ of 0.83. Other groups had also attempted combinations of approaches to improve precision [75-78]. Jimeno et al. [61] demonstrated that although a curated dictionary method for NER is still superior ($F\text{-score}$ of 0.59 to 0.69) compared to statistical methods ($F\text{-score}$ of 0.28 to 0.57), voting strategy may provide a better performance than any single method ($F\text{-score}$ of 0.54 to 0.83). Li et al. [79] reported an $F\text{-score}$ of 0.743 by chaining 2 conditional random field models for named entity recognition followed by classification.

It is clear from recent developments that a dictionary approach (solely or in combination) tends to outperform lexicon-based or statistics-based approaches [61, 80, 81]. However, it is debatable how well NER must perform before rendered useful in biomedical text mining [82] as previous studies illustrated that performance (in terms of $F\text{-score}$) of biomedical information extraction in approximately equal to that of biological NER [50, 83, 84].

It is unlikely automated biomedical NER will approach that of human experts in the near future in terms of precision. However, current biomedical NER systems can be useful in providing an initial list of genes and protein names for human curation if precision is critical in the context of application.

A close relative to NER is an area of study known as gene normalization (GN). The main purpose of GN is to standardize a set of gene names using a thesaurus of gene synonomous and homologies [85] which will be useful in many aspects of biomedical literature analysis. GN is currently an area of active research and had warranted a specialized track in the latest BioCreative II challenge workshop in 2007 [86-88].

Intuitively, AR is an easier task than NER, which is supported by the observation that current AR systems uniformly perform consistently better compared to current NER system in term of their $F\text{-score}$ [2]. Similarly to NER, AR techniques has been used to create lists of abbreviations, such as ADAM [89], which can be used for other applications.

3.4. Information Extraction: Getting the Facts

The most common aim of biomedical information extraction (IE) is finding relationships between two entities [90, 91], in this case, usually either genes, proteins or metabolites. The relationship in question may range from very general, like any form of biochemical association, to very specific, such as regulatory activation. In contrast to IR, IE systems tends to be less ad hoc but are more targeted towards specific relationships. Another difference is the granularity of text. IR systems identifies text of interest whereas IE systems work within the text to identify facts of interest, which can be subsequently verified by a curator reading the paper of interest. Biomedical information extraction are currently being developed by three different ways [92]; co-occurrence, template matching, and natural language processing.

3.4.1. Co-occurrence

Co-occurrence is fundamentally statistical and based on the tenet that multiple occurrences of the same pair of entities suggests that the pair of entities are related in some way [22, 93] and the confidence of such relatedness increases with more co-occurrences. In practice, most systems used a frequency-based scoring technique [38, 93-95] to ensure that the co-occurrence of two entities is higher than random chance [2]. Being a statistical probability, co-occurrence of entities within the same text alone will not give any insights into the type and nature of the relationship [96] unless it is used downstream to IR systems which pre-identified the type of relationships of interest [94, 97, 98]. Despite so, the advantages of co-occurrence techniques over natural language processing (NLP) are simplicity, easy to implement and efficient over large amounts of data [99].

Two of the most successful implementations of co-occurrence methods are PubGene [100] and CoPub Mapper [101]. Both had been shown to co-relate well with microarray results. A recent tool, PPI Finder [3], attempted to map queries into Gene Ontology [28, 29] before co-occurrence analysis and reported that only 28% of the co-occurred pairs in PubMed abstracts appeared in any of the commonly used human protein-protein interaction databases (HPRD, BioGRID and BIND). In addition, only 69% of the known protein-protein interactions in HPRD showed co-occurrences in the literature [3] suggesting large proportions of experimentally validated protein-protein interaction may not be reported in the literature.

A variant of co-occurrence which bootstrapped on PubMed had also been suggested [30, 102] and is commonly known as co-citation. Common experiences in using PubMed suggested that if two terms were used together with an 'AND' clause to search PubMed, it would return the intersection of the results compared to when each term was used separately. Translated statistically, the size of the intersection (proportion of co-cited documents) increases with more relatedness in the pair of entities used to interrogate PubMed. However, co-citation had not been evaluated against co-occurrence measures.

3.4.1.1. Brief Descriptions of Co-occurrence Systems

PubGene [100] used the idea that if 2 entities were mentioned in the same article, there will be some relationship, no matter how remote. By this, it simply counted the number of articles with occurrences of 2 entities in question and used the count as a relative strength of relatedness. If there was 1 article in 10 million mentioning both gene entities, there would be

60% chance of a true relationship between them. This was increased to 71% with 5 or more articles in 10 million. A study testing PubGene's count based co-occurrence on statistical testing framework using Poisson distribution demonstrated that 1 co-occurred protein-pairs in 900 thousand abstracts is generally significant (p-value of less than 1%) [303] suggesting that PubGene's criteria of 1 co-occurred protein-pairs in 10 million abstracts is generally significant.

CoPub Mapper [101] calculated the article number-normalized occurrence of any 2 entities (number of articles with both concepts divided by total number of articles, divided by the product of number of articles with one concept each divided by the total number of articles), which was termed as mutual information measure. Mutual information measure was then converted to logarithmic scale and normalized on a scale of 0 to 100, known as the scaled log transformed relative score. The confidence of relationship between 2 entities was directly proportional to the scaled log transformed relative score.

MedInfoText [103] aims to extract the relationships between gene methylation and cancer from biomedical literature. It uses Lucene-based full text search engine for Perl, Plucene (<http://search.cpan.org/dist/Plucene/>), for term indexing. The relationships are extracted based on co-occurrences of terms in abstracts and sentences by measuring association rule interestingness [104] using support and confidence measures.

3.4.2. Natural Language Processing / Template Matching

There is significant overlap between Natural Language Processing (NLP) methods and Template Matching methods as both share many common components, such as ontologies and substantial use of Regular Expressions. Template Matching methods mainly uses the grammatical structure of English sentence to construct templates, which can be done either manually [105] or automated [106, 107], and using these templates to extract specific information from text. An example has shown extracting mutation information solely by regular expressions achieved 85% precision [108]. BioIE [109] uses rules with template matching. However, sole use of template matching is not widespread in biomedical text analysis. Instead, template matching is usually used either implicitly within NLP methods or explicitly to process the outputs of NLP. The main principles of NLP will be described to facilitate discussions into the tools that employed NLP.

Natural Language Processing (NLP) covers all aspects and stages of processing natural human speech or text into forms usable by automated systems. In the case of biomedical NLP, it can be reasonably assumed that only machine-accessible text in English language forms the majority of source materials. Hence, this section only covers the processing of English text. Given the long history of NLP, a large volume of text had been written and this section can only provide a broad coverage of the general techniques used in NLP, namely, tokenization, part of speech (POS) tagging, and shallow parsing.

The text is first broken up into its constituent atoms, known as tokens, by a process of tokenization. While the granularity of tokens may vary from chapters to phonemes (atomic units of sound), the most common form of tokenization for NLP is to break down each sentence into words and punctuations. Generally, in English text, words in a sentence are delimited by whitespace(s), except words prior to a punctuation, like a comma. This feature poses the main challenge of tokenization – distinguishing punctuations (especially period) signaling the end of either a sentence or phrase and that being part of the previous token, like

in shorthand (for example, Mr., Dr.). Another challenge is the expansion of common contractions, such as “you’ve”, which is usually done using a dictionary approach.

Part of Speech (POS) Tagging is the process of annotating a series of tokens, presumably a sentence, with semantics information (that is, the role of each token in a sentence) with a set of tags, such as Penn Treebank Tag Set [110]. There are two main approaches to POS tagging, rule-based and probabilistic. Probabilistic taggers estimate the probability of a sequence of POS tags for a given sequence of words, based on a probability model, such as the Hidden Markov Model [111, 112]. On the other hand, a rule-based tagger [39] uses contextual rules to assign tags to ambiguous words. Contextual rules, often known as context frame rules, are rules suggesting a tag based on the tag(s) before and after the unknown word. This is usually followed by morphological rules which looks at the appearance of the word. For example, words ending with “ing” is likely to be verbs.

POS Tagging can be seen as a reduction scheme to map a potentially infinite amount of words into a small and definite set of tags, to facilitate further processing. Based on the sequence of POS tags, the source text is then broken up into non-overlapping phrases. This process is chunking, also known as shallow parsing, where phrases are tagged by a small number of grammatical phrase tags, such as, Noun Phrase, Verb Phrase, Prepositional Phrase, Adverb Phrase, Subordinate Phrase, Adjective Phrase, Conjunction Phrase, and List Marker. Chunking is generally useful as a preprocessing step for information extraction as the main effect of NLP is to process unstructured data (in the form of human text) into a more structured form (POS annotated phrases), suitable for information extraction [14, 113-115] or phrase extraction in itself, can be used for medical concept identification [116]. One of the most important output of chunking is the subject-verb-object(s) tuples. In linguistic typology, the English language, together with more than 75% of all languages in the world is classified under the subject-verb-object (SVO), also known as the agent-verb-object (AVO) typological system [117]. A sentence can be processed into more than one SVO tuples, which are useful for extracting relationships between entities [118-120]. In contrast to shallow parsing (chunking), full parsing or complete parsing is much more computationally intensive but has been shown to be useful in a real world biological application [121].

3.4.2.1. Brief Descriptions of Natural Language Processing/Template Matching Systems

GENIES [14] used GenBank and SwissProt to tag genes and protein names in text before processing using MedLEE [122], a specialized text processor for biomedical literature, which used rules to parse text into structured frames. The original lexicon in MedLEE was enlarged in GENIES.

MedScan [123] used a biomedical lexicon to tag text before tokenization and stemming words into their infinite form. Tokens were syntactically processed by a parser based on active chart parser algorithm [124]. Syntactic tokens were processed into semantics structured based on an established method [125].

MeKE [126] used Gene Ontology and LocusLink as basis for constructing function names and gene names ontologies to tag text. A pattern recognizer was trained to recognize sentences or phrase describing gene functions by sentence alignment. Extracted sentences were classified by the probability of containing protein-function relationships by a Naïve Bayes classifier.

PreBIND [94] collected a list of non-redundant protein names from the NCBI RefSeq database, which was used to scan for protein names in text. Extracted term features from text (protein names, words, adjacent words) were used to train a linear support vector machine for identifying protein-protein binding interactions using yeast data.

Arizona Relation Parser [84] specialized a Brill tagger [39] with 100 PubMed abstracts and GENIA corpus [127]. POS tagged text was processed by a hybrid shallow parser which allowed for multilevel n-ary branching of up to 24-nary, meaning POS tags or phrases up to 24 tags or phrases away could be linked. Information from each allowed combination was extracted by rule templates.

BioRAT [128] used GATE [129] to provide utilities to perform POS tagging and information extraction. Text was POS tagged to remove uninformative words, such as determinant verbs, before passing through a series of template matching to extract protein-protein interactions using a set of handwritten templates and a manually curated list of entities forming the gazetteers. An evaluation between information extraction from abstracts and full text demonstrated that 58.7% of the interactions were extracted from full text but the precision from full text was 51.3%, 3.82% lower than that of abstracts (55.07%).

Chilibot [130] used TnT POS tagger [131] trained on GENIA corpus [127] followed by CASS for shallow parsing. Chunked text was used to construct named interactions among either biological concepts, genes, proteins, or drugs. It also showed that the connectivity of molecular networks extracted from the biological literature follows the power-law distribution.

GIS [132] consists of two modules: gene information screening and gene-gene relation extraction. In gene information screening, documents were downloaded from PubMed and informative sentences were selected before tagged using a domain-specific lexicon built from online dictionaries and suggestions by biomedical researchers. Gene-gene relation extraction had a identifier for gene-gene relations, presumably trained by machine learning methods. Identified relations were then evaluated to be positive, cooperative or negative.

MedTAKMI [119] was built on TAKMI, which used standard text processing (tokenization, POS tagging, shallow parsing) to process text into [133] subject-verb-object(s) tuples as an intermediate form for further information extraction. Besides using a dictionary of protein names, the difference between TAKMI and MedTAKMI was not clear. A number of tools for information extraction from subject-verb-object(s) tuples has been described [119] but none was evaluated. Nevertheless, MedTAKMI could be used as a curator's tool.

Karopka et al. [134] used GATE [129] and modified the ANNIE gazetteer of GATE to tag for gene names before tokenizing the sentences and POS tagging by GATE. Gene relations were extracted by 34 manually-written grammar rules in JAPE (Java Annotation Patterns Engine) language, which is essentially template matching.

Cooper and Kershenbaum [97] overlapped the results of text processing, and graphical and statistical analysis to extract protein-protein interactions. TALENT text mining system [135] was used for extracting protein-protein interactions from text by a method previously established for extracting relationships between noun phrases [136]. For each pair of proteins, a 3-hop neighbourhood graph was defined and the coherence of the graph, defined as “the ratio of the number of edges present to the possible number of edges”, was calculated but the threshold coherence for a positive result was not obvious. Positive results from both methods, led to an improvement of precision from 62% to 74%.

Santos et al. [118] used a shallow parser, CASS, to extract protein names (Wnt pathway proteins) from text by statistical comparison with a Wnt pathway corpus to avoid maintenance of a list of protein names. At the same time, they used Link parser to process text to subject-verb-object(s) tuples before full parsing to extract interactions between the Wnt pathway components. However it was not clear which POS tagger was used before shallow parsing by CASS nor which grammar was used for Link parser.

Jang et al. [137] avoided the problem of multi-word protein names and complex sentences by simplification of sentences – protein names and specific noun phrases were substituted by pre-defined words, and parenthesis phrases which does not contain entity names were removed. This simplified sentence is POS tagged by a Brill tagger [39] trained on GENIA corpus [127], then shallow parsed. Protein-protein interactions were extracted by Regular Expression parsing of shallow parsed sentences.

CONAN [138] combined several known tools and used a set of decision criteria to evaluate the output of these tools and achieved 53% precision and 52% recall using LLL corpus [139]. CONAN used Krauthammer et al. [66], AbGene [67] and NLProt [140] for NER, and MuText [141] and PreBind [94] for interaction extraction.

GeneLibrarian [29] is a PubMed text summarization tool that uses a list of gene names as input, instead of keywords. It consists of 2 modules: GeneCluster and GeneSum. GeneCluster clusters the list of given gene names using Gene Ontology while GeneSum obtains text from PubMed based on the clusters and process the text linguistically by natural language processing methods. Finally, a 9-state finite state machine (FSA) is used to perform text summarization based on the part-of-speech tags of the processed text.

Rinaldi et al. [121] used a dependency parser, Pro3Gres [142], to parse processed text. It reported precision of 96% and recall of 63%. The source text were initially analyzed for specific terms, such as entity names, before splitting into sentences by MXTERMINATOR [143] and tokenized using Penn Treebank tokenizer [110]. The tokenized text was POS tagged by MXPOST [144] and lemmatized by morpha [145]. After which, the text was processed against GENIA Ontology [127] before shallow parsed by LTCHUNK [146]. The chunks were parsed for dependencies by Pro3Gres.

Feng et al. [147] aims to extract chemical-CYP3A4 interactions from biomedical text. They had used a combined rule-based and dictionary-based method to identify chemical names in text and had used GATE [129] for POS tagging and information extraction. An evaluation with 100 abstracts demonstrated 87.4% recall and 92.3% precision for chemical name identification and 85.2% recall and 92.0% precision for the extraction of chemical-CYP3A4 interactions.

Muscorian [148] used the 2-layered generalization-specialization paradigm suggested by Novichkova et al. [123] and achieved 85% precision and 30% recall on binding and activation relationships. A manually curated dictionary of entity names were assembled for abbreviation of multi-word entity names [45] in the abstracts before processing into subject-verb-object structures using MontyLingua [149], a generic text processing engine [150] formally used to process scientific text [151, 152]. This is followed by specific data mining from the subject-verb-object structures.

E3Miner [153] aims to extract the interactions between ubiquitin-protein ligase (E3) and its target proteins from biomedical text. E3 was identified using a specially constructed POS tagger and shallow parser. The target proteins were then identified using a rule-based method

before processing for Gene Ontological terms. Using a set of 47 abstracts, a precision of 97% and a recall of 74% was indicated.

PIE [154] used a 2-phase method: a term co-occurrence based method to identify potential abstracts that may contain protein-protein interactions, followed by NLP processing using a POS tagger trained on GENIA corpus [127] to extract protein-protein interactions. PIE was tested on BioCreAtIVe I corpus [155] and reported 84% precision.

Barnickel et al. [156] used artificial neural networks for semantic role labelling of sentences at a rate of 25 to 390 milliseconds per sentence for relationship extraction. It reported a precision of 71% with 43% recall.

Jiao and Wild [157] used a set of maximum entropy based learning models for POS tagging, NER, dependency parsing, and relation extraction to extract cytochrome P-450 protein and chemical interactions. It reported an overall precision of 68.4% and recall of 72.2%.

3.4.3. Applications of Biomedical Information Extraction

There are two main schools of thought in current biomedical IE, one school (call it the “Specialist” for further argument) takes the view that biomedical texts are specialized text (very much like the use of Legalese in legal documents) requiring highly domain-specific tools. This opinion had sparked off the development of biomedical-specific POS tag sets (such as SPECIALIST tag set [158]), POS taggers (such as MedPost [41]), ontologies and NLP systems (such as MedLEE [159]). Another school (call it the “Generalist”) takes the view that biomedical texts are not sufficiently specialized to require a re-development of existing tools but either re-use or adapt generic NLP tools for biomedical text processing. This triggered the use of generic NLP systems, such as TAKMI [133], Link Grammar [160], and GATE [129], for biomedical IE.

Regardless of opinions, the focus of biomedical IE has been on a few types of relationships, namely, physical protein-protein interactions (PPIs) [14, 94, 97, 161], non-physical PPIs [162-164], relationships between proteins and diseases and terms [101, 165-167], gene regulation [113, 168], protein phosphorylation [153, 169], alternate transcription [170], and functions of transcription factors [171].

Most of the earlier work in biomedical IE belongs to the Specialist's School and one of the significant contributions was the GENIES system [14], which modified MedLEE's lexicon preprocessor and parser. GENIES was only evaluated using one article and reported an overall precision of 96%. MedLEE [122] was also adapted to process pathology reports for breast cancer study [172]. The MeKE system used a lexicon of gene and protein names from LocusLink to construct an ontology for pattern matching within text into structured data [126]. Novichkova et al. [123] agreed that NLP can be used to process text into semantic structures and developed MedScan, which incorporated a biomedical lexicon. Further work by Daraselia demonstrated 91% precision with 21% recall ($F\text{-score} = 0.34$) in extracting protein interactions from text using MedScan [163]. The output of MedScan were assembled and constructed into a set of tissue-specific pathways, ResNetCore database [173]. The Arizona Relation Parser [84] attempted to improve MedScan's low recall by re-training Brill Tagger [39] with Brown Corpus, Wall Street Journal, PubMed abstracts, and added GENIA lexicon [127]. This was followed by a hybrid grammar, template matching and semantic filtering. It reported 89% precision with 35% recall ($F\text{-score} = 0.5$). GIS [132] uses a domain-

specific lexicon but instead of NLP, it employs a machine learning approach and reported 84% precision with 77% recall (F-score = 0.80). Jang et al. [137] had trained Brill tagger [39] on GENIA corpus [127] and a purpose-built protein-protein interaction extractor system which achieved 81% precision and 43% recall (F-score = 0.56). E3Miner [153] built a specialized POS tagger and shallow parser to achieve 97% precision and 74% recall (F-score = 0.84). PIE [154] used a GENIA [127] trained POS tagger and achieved 84% precision. Jiao and Wild [157] used separate maximum entropy based learning models for each component and reported 68.4% precision and 72.2% recall (F-score = 0.70). Barnickel et al. [156] used artificial neural networks for semantic role labelling of sentences for relationship extraction and reported 71% precision with 43% recall (F-score = 0.54).

In the Generalist's School, BioRAT [128] is one of the earliest systems to modify GATE [129] to extract protein-protein interactions and reported 48% precision with 39% recall (F-score = 0.43). TAKMI [133], originally developed to process customers call logs in IT helpdesks, was used to develop MedTAKMI [119] which uses term frequency to support search results. Karopka et al. [134] modified the ANNIE system of GATE [129] and modified precision and recall measures to account for partial correct extractions, and reported 92.8% precision with 30% recall (F-score = 0.45). Cooper and Kershenbaum [97] used a generic TALENT text mining system [135] with graphical and statistical approaches to mine for protein-protein interactions and reported 74% precision. A Link grammar parser [160] was used to mine the Wnt pathway and reported 90% precision with 64% recall (F-score = 0.75) [118]. Rinaldi et al. [121] used a myraid of text processing tools with GENIA Ontology [127] and reported 96% precision with 63% recall (F-score = 0.76). Feng et al. [147] used a purpose-built rule-dictionary hybrid for chemical name identification and GATE [129] for information extraction and reported 85% recall and 92% precision (F-score = 0.87). Muscorian [148] used MontyLingua [149], a generic text processing engine [150], in the 2-layered generalization-specialization paradigm [123] and achieved 90% precision and 30% recall (F-score = 0.45).

From the research results gathered from both school of thought (tabulated in Table 1), it is still not possible to demonstrate superiority of one approach over the other in terms of system performance. Intuitively, it might be easier to modify an existing system for a specific application than to develop one from scratch. In addition, it has not been demonstrated that systems developed from the Specialist's school of thought can be adapted to extract other biomedical relationship of interests as only a few systems have been designed to extract multiple relationships [38]. On the other hand, Generalists view generic NLP systems as a processing tool to convert unstructured text into structured forms, such as tuples; thus, is inherently more readily adapted for different problems. It is probably reasonable to comment that by adapting an existing system for use in biomedical text mining usually implies that the system has been used in a different context. For example, TAKMI was used in 3 different areas; analyzing customer call logs [133], generating frequently-asked-questions candidates [174], and in MedTAKMI [119].

However, adapting a generic system may require intense effort in formulating rules and templates (GATE and Link Grammar) for the specific problem domain or re-training parts of the system, especially POS tagger, which may require a prior manual tagging of training corpus [38]. For instance, Chilibot [130] used TnT tagger [131] trained on the GENIA corpus [127] but succeeded in using CASS parser (<http://www.vinartus.net/spa>), un-modified, for chunking. This might be an obstacle to adapt a previously adapted generic NLP system for a

biomedical problem to another biomedical problem. Moreover, there is no certainty of rewards in this effort as Miyao et al. [175] had shown that combining text processing components may be synergistic.

Table 1. Summary of performances of biomedical literature analysis systems. 'NG' means that the particular performance metric was not given in the study.

Specialist Systems				Generalist Systems			
Study	Precision	Recall	F-Score	Study	Precision	Recall	F-Score
[14]	0.96	NG	--	[128]	0.48	0.39	0.43
[163]	0.91	0.21	0.34	[134]	0.93	0.30	0.45
[84]	0.89	0.35	0.50	[97]	0.74	NG	--
[132]	0.84	0.77	0.80	[118]	0.90	0.64	0.75
[137]	0.81	0.43	0.56	[121]	0.96	0.63	0.76
[153]	0.97	0.74	0.84	[147]	0.92	0.85	0.87
[154]	0.84	NG	--	[148]	0.90	0.30	0.45
[157]	0.68	0.72	0.70				
[156]	0.71	0.43	0.54				

It is inherent in the process of evaluating systems using corpora that human experts are the only absolute performer. That is, human experts are performing at 100% precision and 100% recall. Therefore, it should be conceivable that learning from the output errors by artificial intelligence and machine learning methods could be used to improve information extraction systems. The first of such biomedical information extraction systems which uses support vector machines and neural networks to mimic human expert curation had surfaced [176] with precision ranging from less than 30% to more than 90% over 68 extraction tasks. In addition, biomedical information extraction had been shown to be able to improve curation efficiency of protein-protein interactions into database [177].

3.5. Text Mining: Finding Hypotheses

While the main premise of IR and IE is deductive reasoning (the conclusion is of no greater generality than the premises), text mining (TM) is fundamentally inductive reasoning (the conclusion is of greater generality than the premises). In other words, TM aims at finding or induce new information and hypotheses from existing knowledge from the literature. One of the pioneers of biomedical TM is Don Swanson who suggested in the mid-80s that there were connections between fish oil and Raynaud's syndrome [4], migraine and magnesium [5], arginine intake and the level of somatomedin C in blood [178]. This had triggered the advancement of biomedical IR/IE, NER and AR, which are all precursors to TM. The method Swanson used is essentially Hypothetical Syllogism (if p then q; if q then r; therefore, if p then r), which is an extension of Modus Ponens. In biomedical TM, it is commonly referred to as Swanson's ABC model [179]. Using this discovery model, Weeber et al. [180] had attempted to automate it and found new potential uses for thalidomide. More recently, the potential therapeutic use of turmeric on spinal cord injuries was suggested [181].

Despite its potential and history, biomedical TM is still at its infancy [182]. In order for hypothesis generation systems to be a standard tool of biologists, a fundamental question needs addressing – how to evaluate an untested set of hypotheses? [2] A way to circumvent this problem may be using statistical measurements from IR/IE to provide a means of prioritizing the potential of each hypotheses, as shown as Anne 2 [183]. In spite of this inherent problem, there may be use of TM to evaluate and score several possible hypotheses from experimental or clinical research [184]. TM is also known by other authors as “literature based discovery” [180, 185, 186] or “knowledge discovery” [187].

4. Related Areas of Importance

Notwithstanding the development in previously discussed areas, there are three other key areas that are important within the literature analysis pipeline, namely, corpora, which forms the gold standard for evaluating systems; databases, which may be used to evaluate systems or are themselves resulting from literature analysis systems; evaluation strategies, the definition of performance metrics and their calculations; assisted microarray analysis using output from biomedical text analyses; and visualization tools for viewing large interaction maps.

4.1. Corpora

A corpus (corpora for plural) is a collection of literature which has been either tagged, annotated or categorized for specific purpose(s). Essentially, a corpus is a defined data set of literature. The importance of corpora to literature analysis tools cannot be over-emphasized, analogously, it is as important as antibodies to protein studies. However, there are not many corpora of biomedical literature for various purposes as they often require manual annotations with high-level agreement among annotators [188, 189], known to be labour-intensive to create [38] and need to reflect a biologist's interpretation of the text [154]. The main value of corpora is that it provides a known finite source of positives which is essential for calculating recall measure and error analyses.

Categorically, the following biomedical corpus for different purposes are as follows: For protein and gene name recognition (NER), there are Yapex (used in [68]) and GeneTag [190], PennBioIE [71] corpora. For abbreviation recognition, there are Medstract [191] and AB3P [53] corpora. For part-of-speech tagging, there are GENIA [127, 192, 193], PennBioIE [71] and MedPost [41]. GENIA team had also expanded the annotation into biomedical events to reflect a biologist's understanding of the text [154]. For relationship extraction, there is a dataset used for Learning Logic in Language 2005 (www.cs.york.ac.uk/aig/lll/) [139, 194] and BioCreAtIvE corpus [155]. BioScope corpus [195] represents the first attempt to incorporate uncertainty or negative information into a corpus.

There is a general sentiment that progress in biomedical literature analysis suffers from the lack of corpora [15] which is relatively obvious when one starts to list down the corpora available for each purpose. There is no biomedical corpus for shallow parsing (chunking) or citation retrieval from PubMed (information retrieval) and minimal choices for relationship extraction. Moreover, testing using different corpora (if any) can result in F-score varying as

much as 19% [196]. Hence, researchers had resorted to compare their system output with that in curated databases [197] as these databases represent high-quality molecular interaction data [198].

4.2. Databases

The main database for biomedical literature is PubMed where most source materials for literature analysis work is derived from. Possibly, the largest repository for biochemical information is KEGG which provides links to GenBank and a number of other publicly available databases. Databases are repositories of source text (PubMed), curated tools for comparison (comparing system output against KEGG in Zhang et al., [197], Lee et al. [199] compared their system against SwissProt and Maguitman et al. [200] tested their system against Pfam), or are themselves the results from literature analysis, such as DIP [201]. It is generally true that databases can benefit from literature analysis efforts [202]. Large institutional initiatives, such as KEGG and BIND [203], which are mainly manually maintained and curated, cater to the general research community.

Bioinformatics has a section of the periodical catering to the publications of databases and Nucleic Acid Research releases issues periodically with a database focus (known as database issue). The latest edition of Nucleic Acid Research database issue (Volume 36) features 98 databases. These have almost become a de facto source of new databases. As the availability of corpora is scarce [204], a cursory knowledge of database availability might assist in evaluation efforts.

In terms of individual proteins, there are databases for proteins in specific organelles [205, 206]; prokaryotic proteins [207]; proteins of specific biochemical events [208-211]; proteins of specific domains or characteristics [212-222]; protein anomalies [223]; transcription factors [224, 225]; crystal structures [226]; proteomics resources [227] and specific classes of proteins, such as lectin [228] and centrosomal proteins [229].

Some databases focused on protein-protein interactions, which may be all types of interactions [201, 230-239] or specific interactions [240-244]. For genes, there are databases for genes of specific characteristics [211, 216, 245-256]; cleavage sites [257]; genetic variations [253, 258-262], promoters or regulatory elements [263-267]; genes of specific organisms [268] organelle genomes [206, 269]; comparative genomics [270]; entire genomic resource [271-274]; and expressed sequence tags [275-277].

Other databases includes those for managing experimental results [278-280]; haptens [281]; orientation of proteins on cellular membranes [282]; protein localization [283]; and therapeutically important pathways [284]. Currently, the largest and most extensive database for microarray and other high-throughput data storage is the Gene Expression Omnibus (GEO) [285].

Scanning the wide variety of databases, it is clear that the challenge is not the creation of databases but on the use of these databases, especially integrating them into a composite (federation) of biomedical databases and querying them [286-289].

4.3. Evaluation Strategies

During the development of a literature analysis tool, it is critical to have an estimation of the reliability, which are usually compared to a standard of desired results [92], usually in a form of tagged, annotated or categorized corpus. The most common evaluation strategy and metrics (measurements), such as precision and recall, originated from the Second Message Understanding Conference in 1989.

Given a corpus and a specific query, the results can be partitioned into true positives (TP; items correctly labeled as positive), false positives (FP; items incorrectly labeled as positive) true negatives (TN; items correctly labeled as negative), and false negative (FN; items incorrectly labeled as negative). With these four items, a few metrics can be established, the most common being precision, also known as positive predictive value, is defined as $TP/(TP+FP)$; recall is $TP/(TP+FN)$. Precision and recall are typically inversely related [290].

Precision and recall are commonly used because of their simplicity to evaluate against an established standard (annotated corpus). However, in the absence of a standard, precision can still be evaluated comparing the output of a system with its input. Karopka et al. [134] modified precision and recall measures to account for partially correct extractions. Precision and recall are important because the inverse of precision is a measure of false positives ($1 - \text{precision}$) of the system output and the inverse of recall measures false negatives ($1 - \text{recall}$) or proportion of lost information as a result of processing.

It is important to note that in IE, it is not possible to define true negatives (TN) as there is no theoretical bounds of the number of 'facts' that can be generated from a piece of text [291]. Therefore, a number of measures that required TN, such as accuracy $((TP+TN)/(TP+FP+TN+FN))$, error rate $((FP+FN)/(TP+FP+TN+FN))$; which is $1 - \text{accuracy}$, negative predictive value $(TN/(FN+TN))$, prevalence $((TP+FN)/(TP+FP+TN+FN))$, and specificity $(TN/(TN+FP))$ are impossible to calculate. In addition, receiver operating characteristics (ROC), which had been used extensively in evaluating system performance [292-295], cannot be calculated for IE as it requires specificity.

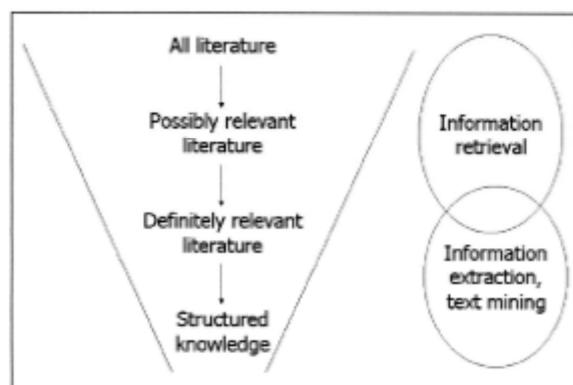


Figure 2. The funnel of knowledge-based information. Demonstrating how the structuring and understanding of knowledge progresses from all biomedical literature becomes more refined and relevant. With larger amounts of literature, information retrieval techniques are essential and once the relevant literature has been defined, information extraction and text mining are required [297].

Notwithstanding variations using different corpora for evaluating different systems and using different criteria for assigning results into each of the three bins (TP, FP, FN), it will be difficult to compare two systems each characterized by precision and recall. Given presence of a single decision parameter (non-categorical variable), it is possible to obtain and compare the respective relative operating characteristic curves (aROC) [296]. However, aROC is not possible for systems without a single decision parameter. Thus, precision (P) and recall (R) are reduced to a single F-score, defined as $2PR/(P+R)$, which is the harmonic mean of precision and recall [16, 17], and is always between 0 and 1 where 1 means that the system produces neither FP or FN. F-score assigns the same weight to both precision and recall, that is, both are equally important. A more general form of F-score allows for different weight be assigned to precision and recall [16]. Hirschman et al. [291] proposed a variant of F-score, simple matching coefficient (SMC), which is defined as $TP/(TP+FN+FP)$.

5. Challenges of Biomedical Literature Analysis

The challenge for the field of biomedical literature analysis is to manage and process large amounts of literature. The rate of publication of literature has exceeded the capacity to manually review it, and therefore, there is an increasing need for IE in this post-genomic era where the focus of biomedical research is shifting from the study of individual proteins and genes to entire biological systems.

Biomedical literature analysis (see Figure 2) consists of getting source text from text repository (information retrieval; IR), finding information of interests within the text (information extraction; IE) and in the process, may require the recognition of entities like proteins (named entity recognition; NER) and evaluation of abbreviations (abbreviations recognition; AR). Extracted information may either be deposited into specialized databases as structured knowledge or may support knowledge discovery or support hypothesis generation by text mining (TM). Testing and evaluation of each step in the analysis requires the presence of a defined data set (corpus) for certain metrics, like recall measure, to be estimated. Alternatively, evaluation can be done by comparing with a suitable, existing database. Inherent in this process is the definition of an appropriate evaluation strategy to allow for comparisons across similar systems.

The main repository of biomedical textual data is PubMed (MedLine) which provided only one means for searching relevant information. Although tools such as Textpresso [24] and MedMiner [25] represented an alternative, they are essentially bootstrapping on PubMed's IR. Despite its universal use, neither the implementation details (source code) of PubMed's IR engine is known [30, 298] nor is the engine thoroughly evaluated for precision and recall. All research to date that retrieved data from PubMed as source had assumed that the data is pristine but this is neither possible nor verified. As a result, certain research questions in biomedical literature analysis which depended on the accuracy of PubMed, such as, what does the collective research knowledge to date tells us about the proteomic and metabolomic differences between mouse and rat, is fundamentally impossible at this stage.

Information extraction is heavily reliant on precise NER [59] and AR. In spite of on-going debate as to the required performance of automated NER [82] and AR systems before being useful in biomedical literature analysis, it is clear that near-human precision is unlikely in the near future. As a result of non-optimal performance (human performance is assumed

optimal), IE is generally limited to a priori approach, that is, the user has to know the list of proteins or entities whose relationship he is interested in, as opposed to an a posteriori approach, such as finding relationships of possibly new entities or entities that are unknown to the user at time of search. Drawing analogies from genomic studies, a priori approach is analogous to microarray technology (the spots on the chip are pre-determined before actual experimentation) while a posteriori approach is likened to Massive Parallel Signature Sequencing (MPSS) [299]. Nevertheless, current or in future, improved NER and AR systems can be of great use in assisting human curators to assemble a near complete dictionary [300], such as ADAM [89]. Systems that learn from human curation efforts, in attempt to improve its performance, had recently surfaced [176]. A system that performed automated curation of extracted interactions from text at the graph-level has also emerged [301]. Another similar problem is the need to recognize variants of the same name, commonly known as gene normalization [86].

Relationships of paired entities by co-occurrence statistics usually requires large volumes of initial text as the term frequency of each entity (number of occurrences of a term divided by the total number of documents) is low, presumably less than 5% of the document set (corpus). This is likely to restrict its use on “small” corpus of less than 100000 and PubGene had used more than 10 million abstracts to generate its gene network [100]. There are three advantages of co-occurrence methods when compared to NLP. Firstly, co-occurrence is basically statistical correlation and is easier to understand by more biologists than NLP techniques. Secondly, most NLP systems work at the sentence level; thus, cannot extract relationships either spanning more than one sentence or in complex sentences, where information may be split across multiple SVO tuples. On the other hand, co-occurrence can be easily deployed on different granularity of text. But it is necessary to note that the fundamental assumption of co-occurrence is independent observations, which is assumable for abstracts as full text usually refer to prior papers (in its introduction and discussion), thus, independent observations cannot be assumed. Lastly, it is known that co-occurrence methods generally has a higher recall but lower precision as compared to NLP means [302], and given that IE by NLP generally suffer from poor recall, it may be possible to improve the overall performance (improving recall substantially while suffering a small decline in precision) by simultaneously employing both co-occurrence and NLP. This notion had been supported by a recent study [303] demonstrating that NLP extracted interactions is generally a proper subset of co-occurred pairs, suggesting that NLP can be used to annotate co-occurred pairs.

As previously described, biomedical IE is driven by the Specialist (biomedical text are highly domain-specific and require specially developed NLP tools) and Generalist (generic or existing non-biomedically focused NLP tools can be adapted for biomedical use) schools of thought. However, both directions require either formulation of rules and templates or re-training parts of the system. Both tasks are manually intensive, require manually tagged corpus [38]. Furthermore, there is no certainty in a better system and combining existing tools may be synergistic [175]. These approaches generally do not fall within the expertise of biologists which are the very people using the systems [204]. Early studies by Grover [304, 305] suggested that native generic NLP tools may be used in biomedical text. Recently, a study by Ling et al. [148] had used an un-modified, generic NLP system for biomedical literature analysis and reported comparable precision. Although it has generally been assumed that some modifications must be made to generic NLP systems (especially the POS tagger) for it to be used on biomedical text, further analysis by Ling et al. [306] revealed that POS

tagging accuracy may not negatively impact on the transformation to subject-verb-object structures due to complementary POS tag use in shallow parsing. However, Ling et al. [148] examined the extraction of 2 interactions from published abstracts and the extrapolation of these results to other interactions [24] requires further studies.

All biomedical literature analysis systems require some form of evaluation, either as performance metrics, such as precision and recall, or as statistical confidence of results (like in BLAST), in attempt to make evaluation across systems. However, this approach faces a number of challenges. Firstly, evaluation by performance often requires tagged corpora which are in severe shortage [15] and most available corpora do not provide programmatic tools to use them readily; hence, developers across the globe have to implement access routines in a particular computer programming language for each new corpus. Secondly, the current evaluation of individual systems make it difficult for comparison even though performance metrics may be known. This is due to different approaches used to obtain the performance metrics. For example, GENIES [14] was evaluated using only one paper; BioRAT [128] was evaluated against an existing database, DIP [201]; E3Miner [153] was evaluated against 47 abstracts. In order to be statistically sound, all systems should preferentially be evaluated against a common set of data, which was accomplished in challenges, such as TREC (trec.nist.gov/) [307], BioCreative and LLL (www.cs.york.ac.uk/aig/lll/) [194]. Alternatively, a common dataset can be established for communal use [196, 204, 308], like that of UCI Machine Learning Repository [309].

One of the main reasons this technology is slowly adopted by biologists is because they are not trained in computer science to integrate the tools effectively [310]. Therefore, it is necessary to present clear benefits of using these tools [2, 88, 204, 311-313]. It is almost a tradition in data analysis and mining to create a system that allows users to set their own parameters in accordance to the task at hand or to evaluate a system independent of meeting user needs [2]. However, the biologists using the system, who are generally clueless about the nature of each parameters, faces a daunting task of setting these parameters. Therefore, it is necessary to involve the biologists in the process of creating new tools or adapting or aggregating existing tools to help biomedical researcher to solve real world problems [2, 314, 315] as these needs remains unmet [316]. Hence, a recent trend is to use literature analysis to provide and update evidence data for Gene Ontology annotations [317-322], combining literature analysis with ontology information for query answering [323], extracting concepts from text [324] or to access the information needs of specific areas of research [315]. At the same time, literature analysis has also been used to group genes based on their functions [325], extracting medically important terms from text [61] and further the development of new ontologies [326].

6. Conclusion

The golden age of biomedical literature analysis of 1998 to 2009 had left us with a number of disjoint sets of tools: systems for specific purposes in the process, such as MedPost; systems for specific biological purposes, like microGENIES; various ontologies and lexicons, like Textpresso Ontology, GENIA Ontology; visualization tools, etc. Although it seems that technology transfer from more traditional fields, such as computational linguistics for understanding general text, had reached its maximum in that period, creative

use of these techniques, picking up and re-structuring the pieces left behind, and targeting the resulting systems to the actual needs of biologists, could bring forth the next golden age of biomedical literature analysis.

Acknowledgment

We wish to thank Professor Thomas Rindflesch, National Institute of Health, USA; Professor Jonathan Wren, Associate Editor for Bioinformatics, for his comments on improving the initial drafts.

References

- [1] Hunter, L. & Cohen, K. B. (2006). Biomedical language processing: what's beyond PubMed? *Molecular Cell*, **21**, 589-594.
- [2] Cohen, A. M. & Hersh, W. R. (2005). A survey of current work in biomedical text mining. *Briefings in Bioinformatics*, **6**, 57-71.
- [3] He, M., Wang, Y. & Li, W. (2009). PPI finder: a mining tool for human protein-protein interactions. *PLoS ONE*, **4**, e4554.
- [4] Swanson, D. R. (1986). Fish oil, Raynaud's syndrome, and undiscovered public knowledge. *Perspectives in Biology and Medicine*, **30**, 7-18.
- [5] Swanson, D. R. (1988). Migraine and magnesium: eleven neglected connections. *Perspectives in Biology and Medicine*, **31**, 526-557.
- [6] Prange, J. D. (1996). *Evaluation driven research: the foundation of the TIPSTER text program*. Tipster Text Program Phase II, May 6-8, 1996.
- [7] Hersh, W., Bhupatiraju, R. T. & Corley, S. (2004). Enhancing access to the Bibliome: the TREC Genomics Track. *Medinfo*, **11**, 773-777.
- [8] Hirschman L. (1998). The evolution of evaluation: lessons from the Message Understanding Conferences. *Information Processing and Management*, **37**, 383-402.
- [9] Leek TR. Information extraction using Hidden Markov Model. *Department of Computer Science*. University of California, San Diego (1997)..
- [10] Fukuda K., Tsunoda T., Tamura A., Takagi T. (1998). Toward information extraction: identifying protein names from biological papers. *Proceedings of the Pacific Symposium on Biocomputing (PSB'98)*: 705 - 716.
- [11] Craven M., Kumlien J. (1999). Constructing biological knowledge bases by extracting information from text sources. *Proc Int Conf Intell Syst Mol Biol*, 77-86.
- [12] Blaschke C., Andrade, M. A., Ouzounis, C. & Valencia, A. (1999). Automatic extraction of biological information from scientific text: protein-protein interactions. *Proc Int Conf Intell Syst Mol Biol*, 60-67.
- [13] Shatkay H. & Wilbur, W. J. (2000). *Finding themes in Medline documents: probabilistic similarity search*. IEEE Conference on Advances in Digital Libraries., pp. 183-192.
- [14] Friedman C., Kra, P., Yu, H., Krauthammer, M. & Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics*, **17**, S74-S82.

- [15] Leser, U. & Hakenberg, J. (2005). What makes a gene name? Named entity recognition in the biomedical literature. *Briefings on Bioinformatics.*, **6**, 357-369.
- [16] Natarajan, J., Berrar, D., Hack, C. J. & Dubitzky, W. (2005). Knowledge discovery in biology and biotechnology texts: a review of techniques, evaluation strategies, and applications. *Critical Reviews in Biotechnology*, **25**, 31-52.
- [17] Tsai, R. T., Wu, S. H., Chou, W. C., Lin, Y. C., He, D. & Hsiang, J. et al. (2006). Various criteria in the evaluation of biomedical named entity recognition. *BMC Bioinformatics*, **7**, 92.
- [18] Han, B., Obradovic, Z., Hu, Z. Z., Wu, C. H. & Vucetic S. (2006). Substring selection for biomedical document classification. *Bioinformatics*.
- [19] Chen, D., Muller, H. M. & Sternberg, P. W. (2006). Automatic document classification of biological literature. *BMC Bioinformatics*, **7**, 370.
- [20] Gerard, S., Edward, A. F. & Harry, W. (1983). Extended Boolean information retrieval. *Communications of the ACM*, **26**, 1022-1036.
- [21] Aronson, A. R., Bodenreider, O., Chang, H. F., Humphrey, S. M., Mork, J. G. & Nelson, S. J. et al. (2000). The NLM Indexing Initiative. *Proc AMIA Symp*, 17-21.
- [22] Wilbur, W. J. & Yang, Y. (1996). An analysis of statistical term strength and its use in the indexing and retrieval of molecular biology texts. *Comput Biol Med*, **26**, 209-222.
- [23] Wilbur, W. J. (2002). A thematic analysis of the AIDS literature. *Pacific Symposium on Biocomputing.*, **7**, 386-397.
- [24] Muller, H. M., Kenny, E. E. & Sternberg, P. W. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biology.*, **2**, e309.
- [25] Tanabe, L., Scherf, U., Smith, L. H., Lee, J. K., Hunter, L. & Weinstein, J. N. (1999). MedMiner: an Internet text-mining tool for biomedical information, with application to gene expression profiling. *Biotechniques.*, **27**, 1210-1214, 1216-1217.
- [26] Shatkay, H., Pan, F., Rzhetsky, A. & Wilbur, W. J. (2008). Multi-dimensional classification of biomedical text: toward automated, practical provision of high-utility text to diverse users. *Bioinformatics*, **24**, 2086-2093.
- [27] Fontaine, J. F., Barbosa-Silva, A., Schaefer, M., Huska, M. R., Muro, E. M. & Andrade-Navarro, M. A. (2009). MedlineRanker: flexible ranking of biomedical literature. *Nucleic Acids Res.*, **37**, W141-146.
- [28] Doms, A. & Schroeder, M. (2005). GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Research*, **33**, W783-786.
- [29] Chiang, J. H., Shin, J. W., Liu, H. H. & Chin, C. L. (2006). GeneLibrarian: an effective gene-information summarization and visualization system. *BMC Bioinformatics*, **7**, 392.
- [30] Simon, M. L., Patrick, M., Kimberly, F. J. & Jennifer, S. (2004). MedlineR: an open source library in R for Medline literature data mining. *Bioinformatics*, **20**, 3659.
- [31] Yoo, I., Hu, X. & Song, I. Y. (2007). Biomedical ontology improves biomedical literature clustering performance: a comparison study. *International Journal of Bioinformatics Research and Applications.*, **3**, 414-428.
- [32] Ding, J., Viswanathan, K., Berleant, D., Hughes, L., Wurtele, E. S. & Ashlock, D. et al. (2005). Using the biological taxonomy to access biological literature with PathBinderH. *Bioinformatics*, **21**, 2560-2562.

- [33] Baker, C. J., Kanagasabai, R., Ang, W. T., Veeramani, A., Low, H. S. & Wenk, M. R. (2008). Towards ontology-driven navigation of the lipid bibliosphere. *BMC Bioinformatics*, **9**, Suppl 1, S5.
- [34] Aronson, A. R. & Rindflesch, T. C. (1997). Query expansion using the UMLS Metathesaurus. *Proc AMIA Annu Fall Symp*, 485-489.
- [35] Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp*, 17-21.
- [36] Hersh, W., Price, S. & Donohoe, L. (2000). Assessing thesaurus-based query expansion using the UMLS Metathesaurus. *Proc AMIA Symp*, 344-348.
- [37] Witte, R., Kappler, T. & Baker, C. J. (2007). Enhanced semantic access to the protein engineering literature using ontologies populated by text mining. *International Journal of Bioinformatics Research and Applications.*, **3**, 389-413.
- [38] Jensen, L. J., Saric, J. & Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Review Genetics.*, **7**, 119-129.
- [39] Brill, E. (1995). Transformation-based error-driven learning and natural language processing: a case study in part of speech tagging. *Computational Linguistics.*, **21**, 543-565.
- [40] Garten, Y. & Altman, R. B. (2009). Pharmpresso: a text mining tool for extraction of pharmacogenomic concepts and relationships from full text. *BMC Bioinformatics.*, **10**, Suppl 2, S6.
- [41] Smith, L., Rindflesch, T. & Wilbur, W. J. (2004). MedPost: a part-of-speech tagger for bioMedical text. *Bioinformatics.*, **20**, 2320-2321.
- [42] Kudo, T., Matsumoto, Y. (2000). Use of support vector learning for chunk identification. *4th Conference on CoNLL-2000 and LLL-142*-144.
- [43] Liu, H. & Friedman, C. (2003). Mining terminological knowledge in large biomedical corpora. *Pacific Symposium on Biocomputing.*, **8**, 415-426.
- [44] Chang, J. T. (2003). *Using machine learning to extract drug and gene relationships from text.* pp. 183. Stanford University.
- [45] Chang, J. T., Schutze, H. & Altman, R. B. (2002). Creating an online dictionary of abbreviations from MEDLINE. *Journal of the American Medical Informatics Association.*, **9**, 612-620.
- [46] Yu, H., Hripcsak, G. & Friedman, C. (2002). Mapping abbreviations to full forms in biomedical articles. *Journal of the American Medical Informatics Association.*, **9**, 262-272.
- [47] Schwartz, A. S. & Hearst, M. A. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Symposium on Biocomputing.*, **8**, 451-462.
- [48] Adar, E. (2004). SaRAD: a Simple and Robust Abbreviation Dictionary. *Bioinformatics.*, **20**, 527-533.
- [49] Pustejovsky, J., Castano, J., Cochran, B., Kotecki, M. & Morrell, M. (2001). Automatic extraction of acronym-meaning pairs from MEDLINE databases. *Medinfo.*, **10**, 371-375.
- [50] Wren, J. D. & Garner, H. R. (2002). Heuristics for identification of acronym-definition patterns within text: towards an automated construction of comprehensive acronym-definition dictionaries. *Methods of Information in Medicine.*, **41**, 426-434.

- [51] Wren, J. D., Chang, J. T., Pustejovsky, J., Adar, E., Garner, H. R. & Altman, R. B. (2005). Biomedical term mapping databases. *Nucleic Acids Research.*, **33**, D289-293.
- [52] Okazaki, N. & Ananiadou, S. (2006). Building an abbreviation dictionary using a term recognition approach. *Bioinformatics.*, **22**, 3089-3095.
- [53] Sohn, S., Comeau, D., Kim, W. & Wilbur, W. J. (2008). Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics.*, **9**, 402.
- [54] Xu, Y., Wang, Z., Lei, Y., Zhao, Y. & Xue, Y. (2009). MBA: a literature mining system for extracting biomedical abbreviations. *BMC Bioinformatics.*, **10**, 14.
- [55] Kuo, C. J., Ling, M. H. T., Lin, K. T. & Hsu, C. N. (2009). *BIOADI: A machine learning approach to identifying abbreviations and definitions in biological literature*. 9th International Conference on Bioinformatics., Singapore.
- [56] Torii, M., Hu, Z. Z., Song, M., Wu, C. H. & Liu, H. (2007). A comparison study on algorithms of detecting long forms for short forms in biomedical text. *BMC Bioinformatics.*, **8** Suppl 9, S5.
- [57] Franzen, K., Eriksson, G., Olsson, F., Asker, L., Liden, P. & Coster, J. (2002). Protein names and how to find them. *International Journal of Medical Informatics.*, **67**, 49-61.
- [58] Hanisch, D., Fluck, J., Mevissen, H. T. & Zimmer, R. (2003). Playing biology's name game: identifying protein names in scientific text. *Pacific Symposium on Biocomputing.*, 403-414.
- [59] Krauthammer, M. & Nenadic, G. (2004). Term identification in the biomedical literature. *Journal of Medical Bioinformatics.*, **37**, 512-526.
- [60] Proux, D., Rechenmann, F., Julliard, L., Pillet, V. V. & Jacq, B. (1998). Detecting Gene Symbols and Names in Biological Texts: A First Step toward Pertinent Information Extraction. *Genome informatics Workshop on Genome Informatics.*, **9**, 72-80.
- [61] Jimeno, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R. & Rebholz-Schuhmann, D. (2008). Assessment of disease named entity recognition on a corpus of annotated sentences. *BMC Bioinformatics.*, **9** Suppl 3, S3.
- [62] Nenadic, G., Spasic, I. & Ananiadou, S. (2004). Mining biomedical abstracts: What is in a term? In: K.Y. Su, (ed), *Natural Language Processing - IJCNLP 2004*. Springer., Berlin, 797-806.
- [63] Jacquemyn, C. (2001). *Spotting and discovering terms through natural language processing*, MIT Press, Cambridge, MA.
- [64] Liu, H., Hu, Z. Z., Torii, M., Wu, C. & Friedman, C. (2006). Quantitative assessment of dictionary-based protein named entity tagging. *J Am Med Inform Assoc.*, **13**, 497-507.
- [65] Tsuruoka, Y. & Tsujii, J. (2004). Improving the performance of dictionary-based approaches in protein name recognition. *J Biomed Inform.*, **37**, 461-470.
- [66] Krauthammer, M., Rzhetsky, A., Morozov, P. & Friedman, C. (2000). Using BLAST for identifying gene and protein names in journal articles. *Gene.*, **259**, 245-252.
- [67] Tanabe, L. & Wilbur, W. J. (2002). Tagging gene and protein names in biomedical text. *Bioinformatics.*, **18**, 1124-1132.
- [68] Chang, J. T., Schutze, H. & Altman, R. B. (2004). GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics.*, **20**, 216-225.
- [69] Egorov, S., Yuryev, A. & Daraselia, N. (2004). A simple and practical dictionary-based approach for identification of proteins in Medline abstracts. *J Am Med Inform Assoc.*, **11**, 174-178.

- [70] Hanisch, D., Fundel, K., Mevissen, H. T ., Zimmer, R. & Fluck, J. (2005). ProMiner: rule-based protein and gene entity recognition. *BMC Bioinformatics.*, **6**, Suppl 1, S14.
- [71] Ryan, T. M., Winters, R. S., Mark, M., Yang, J., Peter, S. W. & Fernando, P. (2004). An entity tagger for recognizing acquired genomic variations in cancer literature. *Bioinformatics.*, **20**, 3249.
- [72] McDonald, R. & Pereira, F. (2005). Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics.*, **6** Suppl 1, S6.
- [73] Settles, B. (2005). ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics.*, **21**, 3191-3192.
- [74] Zhou, G., Shen, D., Zhang, J., Su, J. & Tan, S. (2005). Recognition of protein/gene names from text using an ensemble of classifiers. *BMC Bioinformatics.*, **6**, Suppl 1, S7.
- [75] Hatzivassiloglou, V., Duboue, P. A. & Rzhetsky, A. (2001). Disambiguating proteins, genes, and RNA in text: a machine learning approach. *Bioinformatics.*, **17**, Suppl 1, S97-106.
- [76] Hou, W. J. & Chen, H. H. (2004). Enhancing performance of protein and gene name recognizers with filtering and integration strategies. *Journal of Biomedical Informatics.*, **37**, 448-460.
- [77] Majoros, W., Subramanian, G. & Yandell, M. (2003). Identification of key concepts in biomedical literature using a modified Markov heuristic. *Bioinformatics.*, **19**, 402-407.
- [78] Finkel, J., Dingare, S., Manning, C. D., Nissim, M., Alex, B. & Grover, C. (2005). Exploring the boundaries: gene and protein identification in biomedical text. *BMC Bioinformatics*, **6**, Suppl 1, S5.
- [79] Li, L., Zhou, R. & Huang, D. (2009). Two-phase biomedical named entity recognition using CRFs. *Comput Biol Chem.*, **33**, 334-338.
- [80] Kou, Z., Cohen, W. W. & Murphy, R. F. (2005). High-recall protein entity recognition using a dictionary. *Bioinformatics.*, **21** Suppl 1, i266-273.
- [81] Fundel, K., Guttler, D., Zimmer, R. & Apostolakis, J. (2005). A simple approach for protein name identification: prospects and limits. *BMC Bioinformatics.*, **6** Suppl 1, S15.
- [82] de Bruijn B., Martin J. (2002). Getting to the (c)ore of knowledge: mining biomedical literature. *International Journal of Medical Informatics.*, **67**, 7-18.
- [83] Gaizauskas, R., Demetriou, G., Artymiuk, P. J. & Willett, P. (2003). Protein structures and information extraction from biological texts: the PASTA system. *Bioinformatics.*, **19**, 135-143.
- [84] Daniel, M. M., Hsinchun, C., Hua, S., Byron, B. M. (2004). Extracting gene pathway relations using a hybrid grammar: the Arizona Relation Parser. *Bioinformatics.*, **20**, 3370.
- [85] Crim, J., McDonald, R. & Pereira, F. (2005). Automatically annotating documents with normalized gene lists. *BMC Bioinformatics.*, **6**, Suppl 1, S13.
- [86] Hakenberg, J., Plake C., Royer L., Strobel H., Leser U., Schroeder M. (2008). Gene mention normalization and interaction extraction with context models and sentence motifs. *Genome Biology.*, **9**, Suppl 2, S14.
- [87] Huang, M., Ding, S., Wang, H. & Zhu, X. (2008). Mining physical protein-protein interactions from the literature. *Genome Biology.*, **9**, Suppl 2, S12.
- [88] Krallinger, M., Valencia, A. & Hirschman, L. (2008). Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biolg.*, **9**, Suppl 2, S8.

- [89] Zhou, W., Torvik, V. I. & Smalheiser, N. R. (2006). ADAM: another database of abbreviations in MEDLINE. *Bioinformatics.*, **22**, 2813-2818.
- [90] Hoffmann, R., Krallinger, M., Andres, E., Tamames, J., Blaschke, C. & Valencia, A. (2005). Text mining for metabolic pathways, signaling cascades, and protein networks. *Sci STKE.*, pe21.
- [91] Skusa, A., Ruegg, A. & Kohler, J. (2005). Extraction of biological interaction networks from scientific literature. *Briefings in Bioinformatics.*, **6**, 263-276.
- [92] Cohen, K. B. & Hunter, L. (2008). Getting started in text mining. *PLoS Computational Biology.*, **4**, e20.
- [93] Stapley, B. J. & Benoit, G. (2000). Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in medline abstracts. *Pacific Symposium on Biocomputing.*, **5**, 526-537.
- [94] Donaldson, I., Martin, J., de Bruijn, B., Wolting, C., Lay, V. & Tuekam, B. et al. (2003). PreBIND and Textomy--mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics.*, **4**, 11.
- [95] Hoffmann, R. & Valencia, A. (2004). A gene network for navigating the literature. *Nature Genetics.*, **36**, 664.
- [96] Stephens, M., aplakal, M., Mukhopadhyay, S., Raje, R. & Mostafa, J. (2001). Detecting gene relations from MEDLINE abstracts. *Pacific Symposium on Biocomputing.*, **6**, 483-496.
- [97] Cooper, J. W. & Kershenbaum, A. (2005). Discovery of protein-protein interactions using a combination of linguistic, statistical and graphical information. *BMC Bioinformatics.*, **6**, 143.
- [98] Ray, S. & Craven, M. (2005). Learning statistical models for annotating proteins with function information using biomedical text. *BMC Bioinformatics.*, **6**, Suppl 1, S18.
- [99] Jelier, R., Jenster, G., Dorssers, L. C., van der Eijk, C. C., van Mulligen, E. M. & Mons, B. et al. (2005). Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. *Bioinformatics.*, **21**, 2049-2058.
- [100] Jenssen, T. K., Laegreid, A., Komorowski, J. & Hovig, E. (2001). A literature network of human genes for high-throughput analysis of gene expression. *Nature Genetics.*, **28**, 21-28.
- [101] Alako, B. T., Veldhoven, A., van Baal, S., Jelier, R., Verhoeven, S. & Rullmann, T. et al. (2005). CoPub Mapper: mining MEDLINE based on search term co-publication. *BMC Bioinformatics.*, **6**, 51.
- [102] Becker, K. G., Hosack, D. A., Dennis, G., Jr., Lempicki, R. A., Bright, T. J. & Cheadle, C., et al. (2003). PubMatrix: a tool for multiplex literature mining. *BMC Bioinformatics.*, **4**, 61.
- [103] Fang, Y. C., Huang, H. C. & Juan, H. F. (2008). MeInfoText: associated gene methylation and cancer information from text mining. *BMC Bioinformatics.*, **9**, 22.
- [104] Han, J. & Kamber, M. (2006). *Data Mining: Concepts and Techniques.*, Morgan Kaufmann.
- [105] Yu, H., Hatzivassiloglou, V., Friedman, C., Rzhetsky, A. & Wilbur, W. J. (2002). Automatic extraction of gene and protein synonyms from MEDLINE and journal articles. *AMIA Symp.*, 2002., 919-923.

- [106] Huang, M., Zhu, X., Hao, Y., Payan, D. G., Qu, K. & Li, M. (2004). Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics.*, **20**, 3604-3612.
- [107] Yu, H. & Agichtein, E. (2003). Extracting synonymous gene and protein terms from biological literature. *Bioinformatics.*, **19**, Suppl 1, i340-349.
- [108] Florence, H., Anthony, L. L. & Fred, E. C. (2004). Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics.*, **20**, 557.
- [109] Divoli, A. & Attwood, T. K. (2005). BioIE: extracting informative sentences from the biomedical literature. *Bioinformatics.*, **21**, 2138-2139.
- [110] Marcus, M. P., Santorini, B. & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: the Penn Treebank. *Computational Linguistics.*, **19**, 313-330.
- [111] Dernatas, E. & Kokkinakis, G. (1995). Automatic stochastic tagging of natural language texts. *Computational Linguistics.*, **21**, 137-163.
- [112] Kupiec, J. (1992). Robust part-of-speech tagging using a Hidden Markov Model. *Computer Speech and Language.*, **6**.
- [113] Saric, J., Jensen, L. J., Ouzounova, R., Rojas, I. & Bork, P. (2005). Extraction of regulatory gene/protein networks from Medline. *Bioinformatics*.
- [114] Temkin, J. M. & Gilder, M. R. (2003). Extraction of protein interaction information from unstructured text using a context-free grammar. *Bioinformatics.*, **19**, 2046-2053.
- [115] Rzhetsky, A., Iossifov, I., Koike, T., Krauthammer, M., Kra, P. & Morris, M. et al. (2004). GeneWays: a system for extracting, analyzing, visualizing, and integrating molecular pathway data. *Journal of Biomedical Informatics.*, **37**, 43-53.
- [116] Li, Q. & Wu, Y. F. (2006). Identifying important concepts from medical documents. *J Biomed Inform.*, **39**, 668-679.
- [117] Crystal, D. (1997). *The Cambridge Encyclopedia of Languages (2nd ed.)*, Cambridge University Press, Cambridge.
- [118] Santos, C., Eggle, D. & States, D. J. (2005). Wnt pathway curation using automated natural language processing: combining statistical methods with partial and full parse for knowledge extraction. *Bioinformatics.*, **21**, 1653-1658.
- [119] Uramoto, N., Matsuzawa, H., Nagano, T., Murakami, A., Takeuchi, H. & Takeda, K. (2004). A text-mining system for knowledge discovery from biomedical documents. *IBM Systems Journal.*, **43**, 516-533.
- [120] Rindflesch, T. C., Rajan, J. & Lawrence Hunter, L. (2000). *Extracting molecular binding relationships from biomedical text*. 6th Applied Natural Language Processing Conference, 188-195.
- [121] Rinaldi, F., Schneider, G., Kaljurand, K., Hess, M., Andronis, C. & Konstandi, O. et al. (2007). Mining of relations between proteins over biomedical scientific literature using a deep-linguistic approach. *Artificial Intelligence in Medicine.*, **39**, 127-136.
- [122] Friedman, C. (2000). A Broad Coverage Natural Language Processing System. *American Medical Informatics Association Symposium*, 270 - 274.
- [123] Novichkova, S., Egorov, S. & Daraselia, N. (2003). MedScan, a natural language processing engine for MEDLINE abstracts. *Bioinformatics.*, **19**, 1699-1706.
- [124] Allen, J. (1994). *Natural language understanding*, Benjamin-Cummings Publishing Company, New York.

- [125] Sells, P. (1984). *Lectures on contemporary syntactic theories*, C S L I Publications.
- [126] Chiang, J. H. & Yu, H. C. (2003). MeKE: discovering the functions of gene products from biomedical literature via sentence alignment. *Bioinformatics.*, **19**, 1417-1422.
- [127] Kim, J. D., Ohta, T., Tateisi, Y. & Tsujii, J. (2003). GENIA corpus - a semantically annotated corpus for bio-textmining. *Bioinformatics*, **19**, i180-i182.
- [128] David, P. A. C., Bernard, F. B., William, B. L. & David, T. J. (2004). BioRAT: extracting biological information from full-length papers. *Bioinformatics*, **20**, 3206.
- [129] Cunningham, H. (2000). *Software Architecture for Language Engineering*. Department of Computer Science. pp. 244. University of Sheffield.
- [130] Chen, H. & Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics.*, **5**, 147.
- [131] Brants, T. (2000). *TnT - a statistical part-of-speech tagger*. 6th Applied Natural Language Processing Conference.
- [132] Chiang, J. H., Yu, H. C., Hsu, H. J. (2004). GIS: a biomedical text-mining system for gene information discovery. *Bioinformatics.*, **20**, 120.
- [133] Nasukawa, T. & Nagano, T. (2001). Text analysis and knowledge mining system. *IBM Systems Journal.*, **40**, 967-984.
- [134] Karopka, T., Scheel, T., Bansemer, S. & Glass, A. (2004). Automatic construction of gene relation networks using text mining and gene expression data. *Medical Informatics and the Internet in Medicine.*, **29**, 169-183.
- [135] Neff, M. S., Byrd, R. J. & Boguraev, B. K. (2004). The Talent system: TEXTRACT architecture and data model. *Natural Language Engineering.*, **10**, 307-326.
- [136] Cooper, J. & Byrd R. (1998). *Lexical navigation: visually prompted query refinement*. ACM Digital Libraries Conference.
- [137] Jang, H., Lim, J., Lim, J. H., Park, S. J., Lee, K. C. & Park, S. H. (2006). Finding the evidence for protein-protein interactions from PubMed abstracts. *Bioinformatics.*, **22**, e220-226.
- [138] Malik, R., Franke, L. & Siebes, A. (2006). Combination of text-mining algorithms increases the performance. *Bioinformatics.*, **22**, 2151-2157.
- [139] Cussens, J. & Nédellec, C. (eds) (2005). *Proceedings of the 4th Learning Language in Logic Workshop*, (LLL05), Bonn.
- [140] Mika, S. & Rost, B. (2004). NLProt: extracting protein names and sequences from papers. *Nucleic Acids Research.*, **32**, W634-637.
- [141] Horn, F., Lau, A. L. & Cohen, F. E. (2004). Automated extraction of mutation data from the literature: application of MuteXt to G protein-coupled receptors and nuclear hormone receptors. *Bioinformatics.*, **20**, 557-568.
- [142] Schneider, G., Rinaldi, F. & Dowdall, J. (2004). *Fast, deep-linguistic statistical dependency parsing*. 20th International Conference on Computational Linguistics. Association of Computational Linguistics, University of Geneva, Switzerland.
- [143] Reynar, J., Ratnaparkhi, A. (1997). *A maximum entropy approach to identifying sentence boundaries*. Fifth Conference on Applied Natural Language Processing, Washington, DC: University of Pennsylvania.
- [144] Ratnaparkhi, A. (1996). *A Maximum Entropy Model for Part-of-Speech Tagging*. Conference on Empirical Methods in Natural Language Processing., 133-142.
- [145] Minnen, G., Carroll, J. & Pearce, D. (2001). Applied morphological processing of English. *Natural Language Engineering.*, **7**, 207-223.

- [146] Mikheev, A. (1997). Automatics rule induction for unknown word guessing. *Computational Linguistics.*, **23**, 405-423.
- [147] Feng, C., Yamashita, F. & Hashida, M. (2007). Automated extraction of information from the literature on chemical-CYP3A4 interactions. *Journal of Chemical Information and Modeling.*, **47**, 2449-2455.
- [148] Ling, M. H., Lefevre, C., Nicholas, K. R. & Lin, F. (2007). *Re-construction of Protein-Protein Interaction Pathways by Mining Subject-Verb-Objects Intermediates.*, Second IAPR Workshop on Pattern Recognition in Bioinformatics (PRIB 2007). Springer-Verlag, Singapore.
- [149] Liu, H. & Lieberman, H. (2005). *Metafor: visualizing stories as code.* 10th International Conference on Intelligent User Interfaces.
- [150] Ling, M. H. (2006). An Anthological Review of Research Utilizing MontyLingua, a Python-Based End-to-End Text Processor. *The Python Papers*, **1**, 5-12.
- [151] Chen, L. (2006). *Automatic construction of domain-specific concept structures.* Technischen Universitat Darmstadt.
- [152] van Eck, N. J. & van den Berg, J. (2005). *A novel algorithm for visualizing concept associations.* 16th International Workshop on Database and Expert System Applications, (DEXA'05).
- [153] Lee, H., Yi, G. S. & Park, J. C. (2008). E3Miner: a text mining tool for ubiquitin-protein ligases. *Nucleic Acids Research.*, **36**, W416-422.
- [154] Kim, S., Shin, S. Y., Lee, I. H., Kim, S. J., Sriram, R. & Zhang, B. T. (2008). PIE: an online prediction system for protein-protein interactions from text. *Nucleic Acids Research*, **36**, W411-415.
- [155] Plake, C., Hakenberg, J. & Leser, U. Optimizing syntax patterns for discovering protein-protein interactions. *ACM Symposium on Applied Computing.*, 187-192. ACM Press (2005).
- [156] Barnickel, T., Weston, J., Collobert, R., Mewes, H. W. & Stumpflen, V. (2009) Large scale application of neural network based semantic role labeling for automated relation extraction from biomedical texts. *PLoS ONE.*, **4**, e6393.
- [157] Jiao, D. & Wild, D. J. (2009). Extraction of CYP chemical interactions from biomedical literature using natural language processing methods. *J Chem Inf Model*, **49**, 263-269.
- [158] National Library of Medicine. (2003). UMLS Knowledge Sources (14th ed.).
- [159] Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J. & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association.*, **1**, 161-174.
- [160] Sleator, D. & Temperley, D. (1991). *Parsing English with a Link Grammar.* Third International Workshop on Parsing Technologies.
- [161] Hao, Y., Zhu, X., Huang, M. & Li, M. (2005). Discovering patterns to extract protein-protein interactions from the literature: Part II. *Bioinformatics.*, **21**, 3294-3300.
- [162] von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M. & Foglerini, M. et al. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Research.*, **33**, D433-437.
- [163] Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A. & Mazo, I. (2004). Extracting human protein interactions from MEDLINE using a full-sentence parser. *Bioinformatics.*, **20**, 604-611.

- [164] Yakushiji, A., Tateisi, Y., Miyao, Y. & Tsujii, J. (2001). Event extraction from biomedical papers using a full parser. *Pacific Symposium on Biocomputing.*, **6**, 408-419.
- [165] Chen, E. S., Hripcsak, G., Xu, H., Markatou, M. & Friedman, C. (2008). Automated Acquisition of Disease Drug Knowledge from Biomedical and Clinical Documents: An Initial Study. *Journal of the American Medical Informatics Association.*, **15**, 87-98.
- [166] Osborne, J. D., Lin, S., Zhu, L. & Kibbe, W. A. (2007). Mining biomedical data using MetaMap Transfer (MMTx) and the Unified Medical Language System (UMLS). *Methods in Molecular Biology.*, **408**, 153-169.
- [167] Tiffin, N., Kelso, J. F., Powell, A. R., Pan, H., Bajic, V. B. & Hide, W. A. (2005). Integration of text- and data-mining using ontologies successfully selects disease gene candidates. *Nucleic Acids Research.*, **33**, 1544-1552.
- [168] Aerts, S., Haeussler, M., van Vooren, S., Griffith, O. L., Hulpiau, P. & Jones, S. J. et al. (2008). Text-mining assisted regulatory annotation. *Genome Biology.*, **9**, R31.
- [169] Hu, Z. Z., Narayanaswamy, M., Ravikumar, K. E., Vijay-Shanker, K. & Wu, C. H. (2005). Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics.*, **21**, 2759-2765.
- [170] Shah, P. K., Jensen, L. J., Boue, S. & Bork, P. (2005). Extraction of transcript diversity from scientific literature. *PLoS Computational Biology.*, **1**, e10.
- [171] Yang, H., Nenadic, G. & Keane, J. A. (2008). Identification of transcription factor contexts in literature using machine learning approaches. *BMC Bioinformatics.*, **9 Suppl 3**, S11.
- [172] Xu, H., Anderson, K., Grann, V. & Friedman, C. (2004). Facilitating cancer research using natural language processing of pathological reports. 11th World Congress on *Medical Informatics.*, 565-569.
- [173] Yuryev, A., Mulyukov, Z., Kotelnikova, E., Maslov, S., Egorov, S. & Nikitin, A. et al. (2006). Automatic pathway building in biological association networks. *BMC Bioinformatics.*, **7**, 171.
- [174] Matsuzawa, H. & Fukuda, T. (2000). *Mining structured association patterns from database*. 4th Pacific and Asia International Conference on Knowledge Discovery and Data Mining (PAKDD-2000)., 233-244.
- [175] Miyao, Y., Sagae, K., Saetre, R., Matsuzaki, T. & Tsujii, J. (2009). Evaluating contributions of natural language parsers to protein-protein interaction extraction. *Bioinformatics.*, **25**, 394-400.
- [176] Rodriguez-Esteban, R., Iossifov, I. & Rzhetsky, A. (2006). Imitating Manual Curation of Text-Mined Facts in Biomedicine. *PLoS Comput Biol.*, **2**.
- [177] Alex, B., Grover, C., Haddow, B., Kabadjov, M., Klein, E. & Matthews, M. et al. (2008). Assisted curation: does text mining really help? *Pac Symp Biocomput*, 556-567.
- [178] Swanson, D. R. (1990). Somatomedin C and arginine: implicit connections between mutually isolated literatures. *Perspectives in Biology and Medicine.*, **33**, 157-186.
- [179] Swanson, D. R. (1990). Medical literature as a potential source of new knowledge. *Bulletin of the Medical Library Association.*, **78**, 29-37.
- [180] Weeber, M., Kors, J. A. & Mons, B. (2005). Online tools to support literature-based discovery in the life sciences. *Briefings in Bioinformatics.*, **6**, 277-286.

- [181] Srinivasan, P., Libbus, B. & Sehgal, A. K. (2004). Mining MEDLINE: postulating a beneficial role for Curcumin Longa in retinal diseases. *BioLink Linking Biological Literature, Ontologies, and Databases.*, 33-40.
- [182] Bekhuis, T. (2006). Conceptual biology, hypothesis discovery, and text mining: Swanson's legacy. *Biomedical Digital Libraries*, **3**, 2.
- [183] Jelier, R., Schuemie, M. J., Veldhoven, A., Dorssers, L. C., Jenster, G. & Kors, J. A. (2008). Anni 2.0, a multipurpose text-mining tool for the life sciences. *Genome Biology*, **9**, R96.
- [184] Smalheiser, N. R., Torvik, V. I. & Zhou, W. (2009). Arrowsmith two-node search interface: A tutorial on finding meaningful links between two disparate sets of articles in MEDLINE. *Computer Methods and Programs in Biomedicine*, **94**, 190-197.
- [185] Sarkar, I. N. & Agrawal, A. (2006). Literature based discovery of gene clusters using phylogenetic methods. *AMIA Annu Symp Proc*: 689-693.
- [186] Hristovski, D., Peterlin, B., Mitchell, J. A. & Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *Int J Med Inform*, **74**, 289-298.
- [187] Yetisgen-Yildiz, M. & Pratt, W. (2006). Using statistical and knowledge-based approaches for literature-based discovery. *J Biomed Inform*, **39**, 600-611.
- [188] Colosimo, M. E., Morgan, A. A., Yeh, A. S., Colombe, J. B. & Hirschman, L. (2005). Data preparation and interannotator agreement: BioCreAtIVe task 1B. *BMC Bioinformatics*, **6**, Suppl 1, S12.
- [189] Wilbur, W. J., Rzhetsky, A. & Shatkay, H. (2006). New directions in biomedical text annotation: definitions, guidelines and corpus construction. *BMC Bioinformatics*, **7**, 356.
- [190] Tanabe, L., Xie, N., Thom, L. H., Matten, W. & Wilbur, W. J. (2005). GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC Bioinformatics*, **6**, Suppl 1, S3.
- [191] Pustejovsky, J., Castaño, J., Saurí, R., Rumshisky, A., Zhang, J. & Luo, W. (2002). *Medstract: Creating Large-scale Information Servers for Biomedical Libraries*. ACL 2002 Workshop on Natural Language Processing in the Biomedical Domain.
- [192] Collier, N., Park, H. S., Ogata, N., Tateishi, Y., Nobata, C. & Ohta, T. et al. (1999). *The GENIA project: corpus-based knowledge acquisition and information extraction from genome research papers*. Ninth Conference of the European Chapter of the Association for Computational Linguistics.
- [193] Ohta, T., Tateisi, Y., Mima, H. & Tsujii, J. (2002). *The GENIA corpus: an annotated research abstract corpus in molecular biology domain*. Human Language Technology Conference.
- [194] Cussens, J. & Dzeroski, S. (eds) (2000). *Learning Languages in Logic*. Springer, Berlin, Heidelberg, New York, Barcelona, Hong Kong, London, Milan, Paris, Singapore, Tokyo.
- [195] Vincze, V., Szarvas, G., Farkas, R., Mora, G. & Csirik, J. (2008). The BioScope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, **9**, Suppl 11, S9.
- [196] Pyysalo, S., Airola, A., Heimonen, J., Bjorne, J., Ginter, F. & Salakoski, T. (2008). Comparative analysis of five protein-protein interaction corpora. *BMC Bioinformatics*, **9**, Suppl 3, S6.

- [197] Zhang, Z., Tang, S. & Ng, S. K. (2005). Towards discovering disease-specific gene networks from online literature. *Advances in Bioinformatics and Computational Biology*. 3rd Asia-Pacific Bioinformatics Conference, 161-169.
- [198] Chatr-aryamontri, A., Kerrien, S., Khadake, J., Orchard, S., Ceol, A. & Licata, L. et al. (2008). MINT and IntAct contribute to the Second BioCreative challenge: serving the text-mining community with high quality molecular interaction data. *Genome Biology*, **9 Suppl 2**, S5.
- [199] Lee, L. C., Horn, F. & Cohen, F. E. (2007). Automatic Extraction of Protein Point Mutations Using a Graph Bigram Association. *PLoS Comput Biol*, **3**, e16.
- [200] Maguitman, A. G., Rechtsteiner, A., Verspoor, K., Strauss, C. E. & Rocha, L. M. (2006). Large-scale testing of biome informatics using Pfam protein families. *Pac Symp Biocomput*: 76-87.
- [201] Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U. & Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, **32**, D449-451.
- [202] Miotto, O., Tan, T. W. & Brusic, V. (2005). Supporting the curation of biological databases with reusable text mining. *Genome Inform*, **16**, 32-44.
- [203] Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M. & Bantoft, K. et al. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research*, **33**, D418-424.
- [204] Zhou, D. & He, Y. (2008). Extracting interactions between proteins from the literature. *Journal of Biomedical Informatics*, **41**, 393-407.
- [205] Scott, M., Lu, G., Hallett, M. & Thomas, D. Y. (2004). The Hera database and its use in the characterization of endoplasmic reticulum proteins. *Bioinformatics*, **20**, 937-944.
- [206] Basu, S., Bremer, E., Zhou, C. & Bogenhagen, D. F. (2006). MiGenes: a searchable interspecies database of mitochondrial proteins curated using gene ontology annotation. *Bioinformatics*, **22**, 485-492.
- [207] Martinez-Bueno, M., Molina-Henares, A. J., Pareja, E., Ramos, J. L. & Tobes, R. (2004). BacTRegulators: a database of transcriptional regulators in bacteria and archaea. *Bioinformatics*, **20**, 2787-2791.
- [208] Percudani, R. & Peracchi, A. (2009). The B6 database: a tool for the description and classification of vitamin B6-dependent enzymatic activities and of the corresponding protein families. *BMC Bioinformatics*, **10**, 273.
- [209] Blaineau, S. V. & Aouacheria, A. (2009). BCL2DB: moving 'helix-bundled' BCL-2 family members to their database. *Apoptosis*, **14**, 923-925.
- [210] Holliday, G. L., Bartlett, G. J., Almonacid, D. E., O'Boyle, N. M., Murray-Rust, P. & Thornton, J. M. et al. (2005). MACiE: a database of enzyme reaction mechanisms. *Bioinformatics*, **21**, 4315-4316.
- [211] Mao, C., Qiu, J., Wang, C., Charles, T. C. & Sobral, B. W. (2005). NodMutDB: a database for genes and mutants involved in symbiosis. *Bioinformatics*, **21**, 2927-2929.
- [212] Wood, D. L., Miljenovic, T., Cai, S., Raven, R. J., Kaas, Q. & Escoubas, P. et al. (2009). ArachnoServer: a database of protein toxins from spiders. *BMC Genomics*, **10**, 375.
- [213] Testa, O. D., Moutevelis, E. & Woolfson, D. N. (2009). CC+: a relational database of coiled-coil structures. *Nucleic Acids Res.*, **37**, D315-322.

- [214] Li, Y. & Chen, Z. (2008). RAPD: a database of recombinantly-produced antimicrobial peptides. *FEMS Microbiol Lett.*, **289**, 126-129.
- [215] Jacobs, G. H., Chen, A., Stevens, S. G., Stockwell, P. A., Black, M. A. & Tate, W. P. et al. (2009). Transterm: a database to aid the analysis of regulatory sequences in mRNAs. *Nucleic Acids Res.*, **37**, D72-76.
- [216] Jayakanthan, M., Muthukumaran, J., Chandrasekar, S., Chawla, K., Punetha, A. & Sundar, D. (2009). ZifBASE: a database of zinc finger proteins and associated resources. *BMC Genomics*, **10**, 421.
- [217] Kim, C., Kwon, S., Lee, G., Lee, H., Choi, J. & Kim, Y. et al. (2009). A database for allergenic proteins and tools for allergenicity prediction. *Bioinformation*, **3**, 344-345.
- [218] Gao, J., Agrawal, G. K., Thelen, J. J. & Xu, D. (2009). P3DB: a plant protein phosphorylation database. *Nucleic Acids Res.*, **37**, D960-962.
- [219] Encinar, J. A., Fernandez-Ballester, G., Sanchez, I. E., Hurtado-Gomez, E., Stricher, F. & Beltrao, P. et al. (2009). ADAN: a database for prediction of protein-protein interaction of modular domains mediated by linear motifs. *Bioinformatics*, **25**, 2418-2424.
- [220] Magkrioti, C. K., Spyropoulos, I. C., Iconomidou, V. A., Willis, J. H. & Hamodrakas, S. J. (2004). cuticleDB: a relational database of Arthropod cuticular proteins. *BMC Bioinformatics*, **5**, 138.
- [221] George, R. A., Spriggs, R. V., Thornton, J. M., Al-Lazikani, B. & Swindells, M. B. (2004). SCOPEC: a database of protein catalytic domains. *Bioinformatics*, **20**, Suppl 1, I130-I136.
- [222] Li, B. & Gallin, W. J. (2004). VKCDB: voltage-gated potassium channel database. *BMC Bioinformatics*, **5**, 3.
- [223] Vucetic, S., Obradovic, Z., Vacic, V., Radivojac, P., Peng, K., Iakoucheva, L. M., et al. (2005). DisProt: a database of protein disorder. *Bioinformatics*, **21**, 137-140.
- [224] Guo, A., He, K., Liu, D., Bai, S., Gu, X. & Wei, L. et al. (2005). DATF: a database of Arabidopsis transcription factors. *Bioinformatics*, **21**, 2568-2569.
- [225] Gao, G., Zhong, Y., Guo, A., Zhu, Q., Tang, W. & Zheng, W. et al. (2006). DRTF: a database of rice transcription factors. *Bioinformatics*, **22**, 1286-1287.
- [226] Tung, M. & Gallagher, D. T. (2009). The Biomolecular Crystallization Database Version 4, expanded content and new features. *Acta Crystallogr D Biol Crystallogr.*, **65**, 18-23.
- [227] Sun, Q., Zyballov, B., Majeran, W., Friso, G., Olinares, P. D. & van Wijk, K. J. (2009). PPDB, the Plant Proteomics Database at Cornell. *Nucleic Acids Res.*, **37**, D969-974.
- [228] Chandra, N. R., Kumar, N., Jeyakani, J., Singh, D. D., Gowda, S. B. , & Prathima, M. N. (2006). Lectindb: a plant lectin database. *Glycobiology*, **16**, 938-946.
- [229] Nogales-Cadenas, R., Abascal, F., Diez-Perez, J., Carazo, J. M. & Pascual-Montano, A. (2009). CentrosomeDB: a human centrosomal proteins database. *Nucleic Acids Res.*, **37**, D175-180.
- [230] Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M. & Hannay, T. et al. (2009). PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674-679.
- [231] McDowall, M. D., Scott, M. S. & Barton, G. J. (2009). PIPs: human protein-protein interaction prediction database. *Nucleic Acids Res.*, **37**, D651-656.
- [232] Zhao, X. M., Zhang, X. W., Tang, W. H. & Chen, L. (2009). FPPI: Fusarium graminearum Protein-Protein Interaction Database. *J Proteome Res.*

- [233] Chatr-aryamontri, A., Ceol, A., Peluso, D., Nardozza, A., Panni, S. & Sacco, F. et al. (2009). VirusMINT: a viral protein interaction database. *Nucleic Acids Res.*, **37**, D669-673.
- [234] Chen, J. Y., Mamidipalli, S. & Huan, T. (2009). HAPPI: an online database of comprehensive human annotated and predicted protein interactions. *BMC Genomics.*, **10** Suppl 1, S16.
- [235] Andres Leon, E., Ezkurdia, I., Garcia, B., Valencia, A. & Juan, D. (2009). EcID. A database for the inference of functional interactions in *E. coli*. *Nucleic Acids Res.*, **37**, D629-635.
- [236] Lin, C. Y., Chen, C. L., Cho, C. S., Wang, L. M., Chang, C. M. & Chen, P. Y. et al. (2005). hp-DPI: Helicobacter pylori database of protein interactomes--embracing experimental and inferred interactions. *Bioinformatics.*, **21**, 1288-1290.
- [237] Goll, J., Rajagopala, S. V., Shiu, S. C., Wu, H., Lamb, B. T. & Uetz, P. (2008). MPIDB: the microbial protein interaction database. *Bioinformatics.*, **24**, 1743-1744.
- [238] Pawlicki, S., Le Bechec, A. & Delamarche, C. (2008). AMYPdb: a database dedicated to amyloid precursor proteins. *BMC Bioinformatics.*, **9**, 273.
- [239] Theodoropoulou, M. C., Bagos, P. G., Spyropoulos, I. C., Hamodrakas, S. J. (2008). gpDB: a database of GPCRs., G-proteins, effectors and their interactions. *Bioinformatics.*, **24**, 1471-1472.
- [240] Chuan Tong, J., Meng Song, C., Thiam Joo Tan, P. & Chee Ren, E. A AS. (2008). BEID: Database for sequence-structure-function information on antigen-antibody interactions. *Bioinformation.*, **3**, 58-60.
- [241] Yimeng, D., Pierre-FranÁois, B., Gianluca, P., Yann, P., James, N. & Pierre, B. (2004). ICBS: a database of interactions between protein chains mediated by \leq -sheet formation. *Bioinformatics.*, **20**, 2767.
- [242] Beuming, T., Skrabaneck, L., Niv, M. Y., Mukherjee, P. & Weinstein, H. (2005). PDZBase: a protein-protein interaction database for PDZ-domains. *Bioinformatics.*, **21**, 827-828.
- [243] Dou, Y., Baisnee, P. F., Pollastri, G., Pecout, Y., Nowick, J. & Baldi, P. (2004). ICBS: a database of interactions between protein chains mediated by beta-sheet formation. *Bioinformatics.*, **20**, 2767-2777.
- [244] Yang, C. Y., Chang, C. H., Yu, Y. L., Lin, T. C., Lee, S. A. & Yen, C. C., et al. (2008). PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database. *Bioinformatics.*, **24**, i14-20.
- [245] Song, S., Huang, Y., Wang, X., Wei, G., Qu, H., Wang, W. & et al. (2009). HRGD: a database for mining potential heterosis-related genes in plants. *Plant Mol Biol.*, **69**, 255-260.
- [246] Richardson, C. J., Gao, Q., Mitsopoulous, C., Zvelebil, M., Pearl, L. H. & Pearl, F. M. (2009). MoKCa database--mutations of kinases in cancer. *Nucleic Acids Res.*, **37**, D824-831.
- [247] Sagar, S., Kaur, M., Dawe, A., Seshadri, S. V., Christoffels, A. & Schaefer, U. et al. (2008). DDESC: Dragon database for exploration of sodium channels in human. *BMC Genomics.*, **9**, 622.
- [248] Miranda-Saavedra, D., De, S., Trotter, M. W., Teichmann, S. A. & Gottgens, B. (2009). BloodExpress: a database of gene expression in mouse haematopoiesis. *Nucleic Acids Res.*, **37**, D873-879.

- [249] Mao, F., Dam, P., Chou, J., Olman, V. & Xu, Y. (2009). DOOR: a database for prokaryotic operons. *Nucleic Acids Res.*, **37**, D459-463.
- [250] Liu, B. & Pop, M. (2009). ARDB--Antibiotic Resistance Genes Database. *Nucleic Acids Res.*, **37**, D443-447.
- [251] Dinger, M. E., Pang, K. C., Mercer, T. R., Crowe, M. L., Grimmond, S. M. & Mattick, J. S. (2009). NRED: a database of long noncoding RNA expression. *Nucleic Acids Res.*, **37**, D122-126.
- [252] Essack, M., Radovanovic, A., Schaefer, U., Schmeier, S., Seshadri, S. V. & Christoffels, A. et al. (2009). DDEC: Dragon database of genes implicated in esophageal cancer. *BMC Cancer.*, **9**, 219.
- [253] Kim, C. K., Kim, J. S., Lee, G. S., Park, B. S. & Hahn, J. H. (2008). PlantGM: a database for genetic markers in rice (*Oryza sativa*) and Chinese cabbage (*Brassica rapa*). *Bioinformation.*, **3**, 61-62.
- [254] Ding, G., Lorenz, P., Kreutzer, M., Li, Y. & Thiesen, H. J. (2009). SysZNF: the C2H2 zinc finger gene database. *Nucleic Acids Res.*, **37**, D267-273.
- [255] Boby, T., Patch, A. M. & Aves, S. J. (2005). TRbase: a database relating tandem repeats to disease genes for the human genome. *Bioinformatics.*, **21**, 811-816.
- [256] Sakharkar, M. K. & Kangueane, P. (2004). Genome SEGE: a database for 'intronless' genes in eukaryotic genomes. *BMC Bioinformatics.*, **5**, 67.
- [257] Brockman, J. M., Singh, P., Liu, D., Quinlan, S., Salisbury, J. & Gruber, J. H. (2005). PACdb: PolyA Cleavage Site and 3'-UTR Database. *Bioinformatics.*, **21**, 3691-3693.
- [258] Shimada, M. K., Matsumoto, R., Hayakawa, Y., Sanbonmatsu, R., Gough, C. & Yamaguchi-Kabata, Y. et al. (2009). VarySysDB: a human genetic polymorphism database based on all H-InvDB transcripts. *Nucleic Acids Res.*, **37**, D810-815.
- [259] Shionyu, M., Yamaguchi, A., Shinoda, K., Takahashi, K. & Go, M. (2009). AS-ALPS: a database for analyzing the effects of alternative splicing on protein structure, interaction and network in human and mouse. *Nucleic Acids Res.*, **37**, D305-309.
- [260] Koscielny, G., Le Texier, V., Gopalakrishnan, C., Kumanduri, V., Riethoven, J. J. & Nardone, F. et al. (2009). ASTD: The Alternative Splicing and Transcript Diversity database. *Genomics.*, **93**, 213-220.
- [261] Duan, S., Zhang, W., Cox, N. J. & Dolan, M. E. (2008). FstSNP-HapMap3, a database of SNPs with high population differentiation for HapMap3. *Bioinformation.*, **3**, 139-141.
- [262] Ackermann, A. A., Carmona, S. J. & Aguero, F. (2009). TcSNP: a database of genetic variation in *Trypanosoma cruzi*. *Nucleic Acids Res.*, **37**, D544-549.
- [263] Palaniswamy, S. K., Jin, V. X., Sun, H. & Davuluri, R. V. (2005). OMGProm: a database of orthologous mammalian gene promoters. *Bioinformatics.*, **21**, 835-836.
- [264] Kim, J., Seo, J., Lee, Y. S. & Kim, S. (2005). TFExplorer: integrated analysis database for predicted transcription regulatory elements. *Bioinformatics.*, **21**, 548-550.
- [265] Gallo, S. M., Li, L., Hu, Z. & Halfon, M. S. (2006). REDfly: a Regulatory Element Database for *Drosophila*. *Bioinformatics.*, **22**, 381-383.
- [266] Morris, R. T., O'Connor, T. R. & Wyrick, J. J. (2008). Osiris: an integrated promoter database for *Oryza sativa* L. *Bioinformatics.*, **24**, 2915-2917.
- [267] Rushton, P. J., Bokowiec, M. T., Laudeman, T. W., Brannock, J. F., Chen, X., Timko, M. P. (2008). TOBFAC: the database of tobacco transcription factors. *BMC Bioinformatics.*, **9**, 53.

- [268] Kaas, Q., Westermann, J. C., Halai, R., Wang, C. K., Craik, D. J. (2008). ConoServer, a database for conopeptide sequences and structures. *Bioinformatics.*, **24**, 445-446.
- [269] O'Brien, E. A., Zhang, Y., Wang, E., Marie, V., Badejoko, W., Lang, B. F. et al. (2009). GOBASE: an organelle genome database. *Nucleic Acids Res.*, **37**, D946-950.
- [270] Maselli, V., Di Bernardo, D. & Banfi, S. (2008). CoGemiR: a comparative genomics microRNA database. *BMC Genomics.*, **9**, 457.
- [271] Lu, T., Huang, X., Zhu, C., Huang, T., Zhao, Q. & Xie, K. et al. (2008). RICD: a rice indica cDNA database resource for rice functional genomics. *BMC Plant Biol.*, **8**, 118.
- [272] Lim, D., Cho, Y.M ., Lee, K. T., Kang, Y., Sung, S. & Nam, J. et al. (2009). The Pig Genome Database (PiGenome): an integrated database for pig genome research. *Mamm Genome.*, **20**, 60-66.
- [273] Gauthier, J. P., Legeai, F., Zasadzinski, A., Rispe, C. & Tagu, D. (2007). AphidBase: a database for aphid genomic resources. *Bioinformatics.*, **23**, 783-784.
- [274] Cameron, R. A., Samanta, M., Yuan, A., He, D. & Davidson, E. (2009). SpBase: the sea urchin genome database and web site. *Nucleic Acids Res.*, **37**, D750-754.
- [275] Nystrom, J., Fierlbeck, W., Granqvist, A., Kulak, S. C. & Ballermann, B. J. (2009). A human glomerular SAGE transcriptome database. *BMC Nephrol.*, **10**, 13.
- [276] Lee, B. & Shin, G. (2009). CleanEST: a database of cleansed EST libraries. *Nucleic Acids Res.*, **37**, D686-689.
- [277] Beldade, P., Rudd, S., Gruber, J. D. & Long, A. D. (2006). A wing expressed sequence tag resource for *Bicyclus anynana* butterflies., an evo-devo model. *BMC Genomics.*, **7**, 130.
- [278] Schlamp, K., Weinmann, A., Krupp, M., Maass, T., Galle, P. & Teufel, A. (2008). BlotBase: a northern blot database. *Gene.*, **427**, 47-50.
- [279] Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J. C. & Dwight, S. S. et al. (2001). The Stanford microarray database. *Nucleic Acid Research.*, **29**, 152-155.
- [280] Markus, R., Srinivas, V., Samuel, A., Johan, S. & Jari, H. k. (2004). ACID: a database for microarray clone information. *Bioinformatics.*, **20**, 2305.
- [281] Singh, M. K., Srivastava, S., Raghava, G. P. & Varshney, G. C. (2006). HaptensDB: a comprehensive database of haptens, carrier proteins and anti-hapten antibodies. *Bioinformatics.*, **22**, 253-255.
- [282] Lomize, M. A., Lomize, A. L., Pogozheva, I. D. & Mosberg, H. I. (2006). OPM: orientations of proteins in membranes database. *Bioinformatics.*, **22**, 623-625.
- [283] Zhang, S., Xia, X., Shen, J., Zhou, Y. & Sun, Z. (2008). DBMLoc: a Database of proteins with multiple subcellular localizations. *BMC Bioinformatics.*, **9**, 127.
- [284] Zheng, C. J., Zhou, H., Xie, B., Han, L. Y., Yap, C. W. & Chen, Y. Z. (2004). TRMP: a database of therapeutically relevant multiple pathways. *Bioinformatics.*, **20**, 2236-2241.
- [285] Barrett, T. & Edgar, R. (2006). Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Methods Enzymol.*, **411**, 352-369.
- [286] Lakshmanan, L. V. S., Sadri, F. & Subramanian, S. N. (2001). SchemaSQL - An extention to SQL for multidatabase interoperability. *ACM Transactions on Database Systems*, **26**, 476-519.
- [287] Wyss, C. M. & Robertson, E. L. (2005). Relational languages for metadata integration. *ACM Transactions on Database Systems*, **30**, 624-660.

- [288] Fristensky, B. (2007). BIRCH: a user-oriented, locally-customizable, bioinformatics system. *BMC Bioinformatics*, **8**, 54.
- [289] Garcia Castro, A., Chen, Y. P. & Ragan, M. A. (2005). Information integration in molecular bioscience. *Appl Bioinformatics*, **4**, 157-173.
- [290] Zhou, W., Smalheiser, N. R. & Yu, C. (2006). A tutorial on information retrieval: basic terms and concepts. *J Biomed Discov Collab.*, **1**, 2.
- [291] Hirschman, L., Park, J. C., Tsujii, J., Wong, L. & Wu, C. H. (2002). Accomplishments and challenges in literature data mining for biology. *Bioinformatics*, **18**, 1553-1561.
- [292] Biagini, R. E., Krieg, E. F., Pinkerton, L. E. & Hamilton, R. G. (2001). Receiver operating characteristics analyses of Food and Drug Administration-cleared serological assays for natural rubber latex-specific immunoglobulin E antibody. *Clinical and Diagnostic Laboratory Immunology*, **8**, 1145-1149.
- [293] Gjengsto, P., Paus, E., Halvorsen, O. J., Eide, J., Akslen, L. A. & Wentzel-Larsen, T. et al. (2005). Predictors of prostate cancer evaluated by receiver operating characteristics partial area index: a prospective institutional study. *Journal of Urology*, **173**, 425-428.
- [294] Margolis, D. J., Bilker, W., Boston, R., Localio, R., Berlin, J. A. (2002). Statistical characteristics of area under the receiver operating characteristic curve for a simple prognostic model using traditional and bootstrapped approaches. *Journal of Clinical Epidemiology*, **55**, 518-524.
- [295] Rosman, A. S. & Korsten, M. A. (2007). Application of summary receiver operating characteristics (sROC) analysis to diagnostic clinical testing. *Advances in Medical Science*, **52**, 76-82.
- [296] Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285-1293.
- [297] Hersh, W. (2005). Evaluation of biomedical text-mining systems: lessons learned from information retrieval. *Briefings in Bioinformatics*, **6**, 344-356.
- [298] Wheeler, D. L., Barrett, T., Benson, D. A., Bryant, S. H., Canese, K. & Chetvernin, V. et al. (2006). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **34**, D173-180.
- [299] Stolovitzky, G. A., Kundaje, A., Held, G. A., Duggar, K. H., Haudenschild, C. D. & Zhou, D. et al. (2005). Statistical analysis of MPSS measurements: application to the study of LPS-activated macrophage gene expression. *Proceedings of the National Academy of Science, U S A*, **102**, 1402-1407.
- [300] Shi, L. & Campagne, F. (2005). Building a protein name dictionary from full text: a machine learning term extraction approach. *BMC Bioinformatics*, **6**, 88.
- [301] Rzhetsky, A., Zheng, T. & Weinreb, C. (2006). Self-correcting maps of molecular pathways. *PLoS ONE*, **1**, e61.
- [302] Wren, J. D. & Garner, H. R. (2004). Shared relationship analysis: ranking set cohesion and commonalities within a literature-derived relationship network. *Bioinformatics*, **20**, 191-198.
- [303] Ling, M. H. T., Leferve, C. & Nicholas, K. R. (2008). Filtering microarray correlations by statistical literature analysis yields potential hypotheses for lactation research. *The Python Papers*, **3**, 4.
- [304] Grover, C., Klein, E., Lascarides, A. & Lapata, M. (2002). *XML-based NLP Tools for Analysing and Annotating Medical Language*. Second International Workshop on NLP and XML (NLPXML-2002).

- [305] Grover, C., Lapata, M. & Lascarides, A. (2003). A comparison of parsing technologies for the biomedical domain. *Natural Language Engineering.*, **1**, 1-38.
- [306] Ling, M. H. T., Leferve, C. & Nicholas, K. R. (2008). A Case Study where Parts-of-Speech Tagging Error Does Not Adversely Affect Extraction of Protein-Protein Interactions from Text. *The Python Papers.*, **3**, 65-80.
- [307] Voorhees, E. & Buckland, L. P. (eds) (2005). *The Fourteen Text REtrieval Conference Proceedings. National Institute of Standards and Technology (NIST).*, the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA), Gaithersburg, Maryland.
- [308] Cano, C., Monaghan, T., Blanco, A., Wall, D. P. & Peshkin, L. (2009). Collaborative text-annotation resource for disease-centered relation extraction from biomedical text. *Journal of Biomedical Informatics.*
- [309] Newman, D., Hettich, S., Blake, C., Merz, C. (1998). *UCI Repository of machine learning databases.* University of California, Department of Information and Computer Science, Irvine., CA.
- [310] Kano, Y., Nguyen, N., Saetre, R., Yoshida, K., Miyao, Y. & Tsuruoka, Y. et al. (2008). Filling the gaps between tools and users: a tool comparator, using protein-protein interaction as an example. *Pacific Symposium on Biocomputing*: 616-627.
- [311] Lourenco, A., Carreira, R., Carneiro, S., Maia, P., Glez-Pena, D. & Fdez-Riverola, F. et al. (2009). @Note: a workbench for biomedical text mining. *J Biomed Inform.*, **42**, 710-720.
- [312] Altman, R. B., Bergman, C. M., Blake, J., Blaschke, C., Cohen, A. & Gannon, F. et al. (2008). Text mining for biology--the way forward: opinions from leading scientists. *Genome Biology.*, **9**, Suppl 2, S7.
- [313] Muller, M., Marko, K., Daumke, P., Paetzold, J., Roesner, A. & Klar, R. (2007). Biomedical data mining in clinical routine: expanding the impact of hospital information systems. *Medinfo.*, **12**, 340-344.
- [314] Caporaso, J. G., Deshpande, N., Fink, J. L., Bourne, P. E., Cohen, K. B. & Hunter, L. (2008). Intrinsic evaluation of text mining tools may not predict performance on realistic tasks. *Pacific Symposium on Biocomputing*, 640-651.
- [315] Roberts, P. M. & Hayes, W. S. (2008). Information needs and the role of text mining in drug development. *Pacific Symposium on Biocomputing*, 592-603.
- [316] Kabiljo, R., Clegg, A. B. & Shepherd, A. J. (2009). A realistic assessment of methods for extracting gene/protein interactions from free text. *BMC Bioinformatics.*, **10**, 233.
- [317] Roberts, P. M. (2006). Mining literature for systems biology. *Briefings in Bioinformatics.*, **7**, 399-406.
- [318] Couto, F. M., Silva, M. J., Lee, V., Dimmer, E., Camon, E. & Apweiler, R. et al. (2006). GOAnnotator: linking protein GO annotations to evidence text. *J Biomed Discov Collab.*, **1**, 19.
- [319] Lussier, Y., Borlawsky, T., Rappaport, D., Liu, Y. & Friedman, C. (2006). PhenoGO: assigning phenotypic context to gene ontology annotations with natural language processing. *Pac Symp Biocomput*, 64-75.
- [320] Natarajan, J. & Ganapathy, J. (2007). Functional gene clustering via gene annotation sentences., MeSH and GO keywords from biomedical literature. *Bioinformation.*, **2**, 185-193.

Chapter 13

INTERACTION DESIGN PROCESS FOR AMBIENT INFORMATION IN UBIQUITOUS SPACE

Takuya Yamauchi

Graduate School of Media and Governance, Keio University,
SFC, Fugisawa, Japan

The purpose of this chapter is to describe constructing methodology for a distribution system. Ubiquitous system enabled to detect environmental information and human motion from places due to an improvement of sensing technology and the distribution system. Previous ubiquitous technology focused on system configuration for building distribution system related to software engineering. However current technology for the ubiquitous computing is required to suit to various architectures such as home and public spaces and the system expected to be used for art and design. Thus a design process for the system should be flexible for places and usages from the beginning of architecture. This chapter explains the methodology for a middleware in the distributed system.

Free servers that are developed and maintained by open community are available for wide operating systems, and developers enabled to use the free server easily. As a result, various distribution systems were operated in a internet, and a social network such as e-learning and e-commerce were widespread. Features of the distribution system in the internet were flexible cooperation between the servers for each purpose. The cooperation system between the servers via internet was simple middleware in order to share necessary information for various purposes between the servers.

Sensor network and interaction design have developed in ubiquitous computing, and a connection with electric devices by the sensor network in small area will be widespread like WAN. Detecting data by the sensors in the electric devices is shared in the distribution system, and the shared information is used to control the electric appliances corporatively. One of the roles for the middleware is how the distribution systems unify and control the electric devices, and a novel method or an analysis for the middleware should be considered to create ubiquitous space. Figure 1 shows usage for design pattern in real world, digital world and ubiquitous computing. Program code that is composed of a software design pattern

is embedded in the distributed system. The design pattern that is efficient program pattern is used for object oriented language in the digital world. The design pattern in program code derived from pattern language for architectural design in the real world by Christopher Alexander, and structures of architecture were categorized by patterns language for each usage.

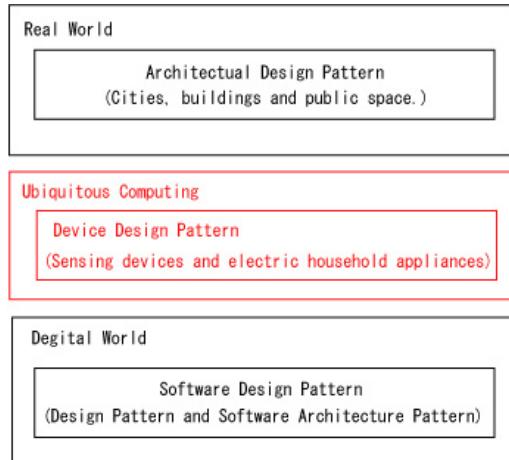


Figure 1. Design Pattern.

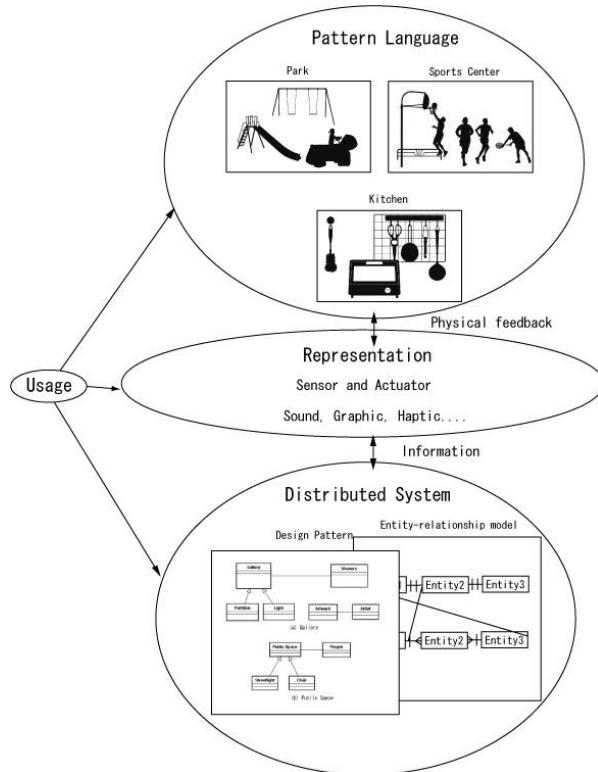


Figure 2. Representation for Ubiquitous Space.

A concept of the pattern language was used for software engineering. In the case of ubiquitous computing, electric appliances or sensors are connected by sensor network in the architecture or suburban structures, and the distribution system for ubiquitous computing is structured by the middleware. In order to design the architectures and suburban structures connecting with the sensors and the electric appliances by the middleware, novel design patterns for interface between architectures in the real world and the distribution system in the digital world should be considered (Figure 2) For instance, setup information related to the sensors and the electric appliances in particular architectures or building are patterned, and how to represent information from the distribution system or usages for the electric appliances are recorded for each spaces.

Analysis for Middleware

We have considered the distributed system, the interfaces for devices and a motion of users. On the other hand, pattern language is structured method of design, and architectures are made by each pattern. Similarly in the case of the distribution system for ubiquitous environment, how to process information from sensor inputs is different for each room.(Figure 3) The process such a representation for the information in rooms should be categorized as patterns for each room in order to reuse configuration record for building the distribution system. The patterns include an architecture design and dynamic information design for the representation in the ubiquitous space, and middleware for constructing ubiquitous space is required to consider the static architecture design including shape and surface in a building and dynamic design for sensing information from sensors. Ambient intelligence, context awareness and affordable device are considered to embed functionalities in the distributed system. The distributed system that is composed of design pattern reflects concept of pattern language related to the architecture design, and the architecture of the distributed system is made by design methodology based on a clear concept. Thus, in order to embed a clear vision in the distributed system, the clear vision is clarified by using a methodology such as waterfall, iterative and incremental development and agile development for interaction design or media art. The middleware should be implemented by new design methodology, and setting configuration, concept and experience are recorded to reuse similar implementation process.

Implementation Process

Architecture design and distribution system are constructed by many types of implementation process. To implement a middleware for ubiquitous space, a house and the distribution system are designed by novel design process. The novel processes are required to imagine whole concept including architecture concept and context awareness for the distribution system, and a representation by the middleware system is also important factor for distribution system in ubiquitous space. The representation is how responsive environments as the distribution systems support users in a room. For instance, information representation is different for a purpose of a usage in the room, and the distribution system

should be structured by the purpose of the usage based on both room architecture and functionality of the distribution system.

Several methodologies exist to implement the design process, and the distribution system for ubiquitous space is constructed by efficient methods of software engineering such as waterfall, iterative and incremental development and agile development. The methods that consist of analysis, implementation and test are used for each phase.

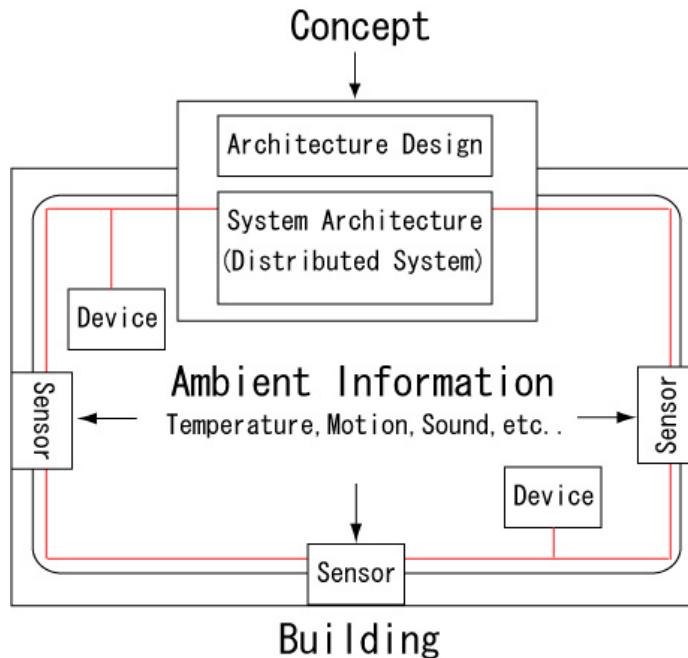


Figure 3. Ambient Information.

Idea and Analysis

Construction of a distribution system needs to consider supporting users in ubiquitous room. Thus, proposals of clients, customers and artists should be investigated in the beginning of analysis phase, and analysis needs to be checked whether several points in the proposals are possible or not. If the several points in the proposal are possible to be implemented in the distributed system, the proposal is summarized in the specification. The distribution system is implemented based on summarizing specification, and functionalities in the distribution system are checked in test phase. Necessary phases are analysis and design phase in a cycle, and how designers specify ideas in design phase is important factor in the cycle. For instance, if the designers find new idea in test phase, important to return previous phase as analysis phase, and an iterative cycle is required to make the distribution system. The cycle should be recorded to reuse other cases as patterns.(Figure 4)

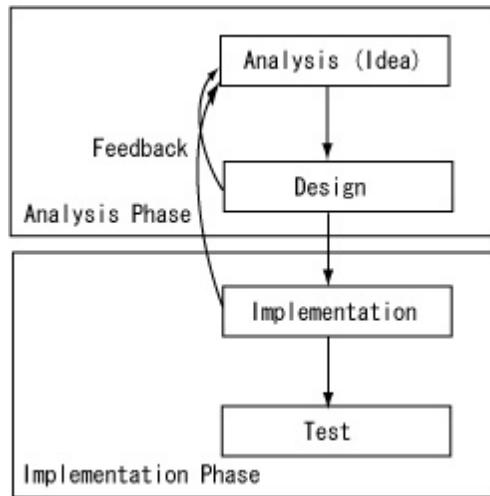


Figure 4. Design Process.

Conclusion

This chapter described the design methodology to build the distribution system in ubiquitous space. Due to improvement of sensing systems or distribution systems, environmental systems enable detection of motion of human or ambient information. Current distribution systems are only implemented by knowledge of information technology related to software engineering, and however, an implementation of new environmental system is required to consider surrounded information in the distribution system such as building, room, usage in the room and so on. The chapter proposed categorization as patterns for architecture design and distribution system. The patterns are composed to records related to analysis, implementation and test. The pattern should be reused as patterns for efficient construction.

Chapter 14

CHRONOLOGIZING WEB PAGES FOR EFFECTIVE SEARCH

Sunil Kumar Kopparapu * and Arijit De†

TCS Innovation Lab - Mumbai

Tata Consultancy Services

Yantra Park, Subhashnagar,

Thane (West), Maharashtra, India

Abstract

Search engines have become a part of our e-life. The simplest way to get near to the information that one is looking for is invariably fueled by a search engine. Due to large amount of on-line data, invariably there are multiple pages that satisfy the search criteria and hence ranking is inevitable. Most of the search engines today use some mechanism to rank the result pages and use this to display which search result page goes first. The ranking is based on information, meta or otherwise, that is readily available or easily derivable from the web pages. An important component that the search engine today does not exploit is the "age of the web page" because this temporal information is not available via the web page readily except probably for news type of information which comes usually with a date tag in the text. The "age of the page" dimension can be effectively used by search engines to rank the search results in a chronological order. For example, a search like "Monsoon in Mumbai" in the monsoon period might signify that the user is looking for information on the "current" monsoon situation rather than the highly ranked page, using some criteria, discussing about monsoon. Access to a chronologically ordered display of search results will find definite use. The reason search engines can not provide the chronological rank order is because of the absence of "age of the page" information. In the proposed paper we will elaborate on the need for dimension which helps in ranking web pages in chronological order. We investigate and discuss existing and new techniques based on natural language processing which can help in chronologizing web pages.

*E-mail address: SunilKumar.Kopparapu@TCS.Com

†E-mail address: Arijit6.D@TCS.Com

1. Introduction

Internet has replaced the television and the radio as the principal source of information and communication for businesses, individuals, media in all walks of life ever since it was commercialized about two decades ago. In the era of Web 2.0, the World Wide Web (WWW) had not only grown in size, but its contents have grown to become multimedia and dynamic. Today data created on the web is not merely text, but also images and videos recorded by digital cameras or web cams, streaming audio and videos extending from podcasts, net casts, music, music videos, chat transcripts, messages exchanged in discussion forums; opinions in blogs make up for the traffic and the content on the Internet. Today data gets created, updated and outdated with more speed than ever before and there is no sign of slowing down. The enormity of data created is so great that some researchers estimate that today's daily rate of content creation is about the same as the content created in three months a couple of years back. With a tremendous data explosion on the web, there is obviously a need to search through this vast volume of data to seek relevant information. This is where a Search Engine (SE) steps in. SE's have become a part of our e-life. SE's are ever popular tools to search and navigate through this ocean of information created. With burgeoning size of multimedia content on the WWW, SE's are increasingly applying more content based multi-media retrieval techniques to sort through all this information. The simplest way to get near the information that one is looking for is invariably fueled by a SE. The widespread use of SE can be captured by the fact that unlike a couple of years ago, today, we do not bookmark or remember the URLs¹ of any web page, we search for it as and when we are in need of it. The reason is (a) with ever changing landscape of the data on the web, we are likely to find different and hopefully more useful source of data than what was available when we accessed the same information earlier and (b) availability of a plethora of SE's which are able to get you to the information you seek.

Search Engines have a tough job at hand especially with the huge amount of data on the Internet that they have to sieve through to dig out the suitable web pages that match the query of the user. Too much data created asynchronously by different sources makes the task of SE's being able to produce a single output web page in response to a query very hard. This paves way for a search engine to give multiple web page output, albeit ranked, in response to a query.

The popularity of SE's resonates from the usable high ranked results that they produce and what distinguishes one SE from another is the way the search output pages are ranked. Different SE's adopt different mechanisms to associate a score against each output web page and then use the score to rank the order in which they are displayed. One of the popular information that the SE's use is to give more importance to a genuine news sites like CNN rather than to a social networking site or blog or a web based bulletin board. There are several mechanisms adopted for SE's to rank order the web pages, most of these are directed towards trying to give the most relevant answer set in response to the query by understanding the intent of the query.

It is to be noted that as the web captures and records data, it implicitly creates a time sequenced snapshot of happenings or changes around us. This adds a new dimensionality: time reference to data on the WWW. This dimensionality has been mildly exploited by

¹More popularly called the HTTP address

SE's in the news articles domain where there is an explicit correspondence between the news article and the date and time. This information is used by the SE's to not only display news in a chronological sequence but also allows user to seek latest or current news. But the same chronological ordering, unfortunately, has not been used to display the results by any of the well known SE's for non-news articles or articles where there is no explicit mention of the time corresponding to the article. When time information is not explicitly mentioned in an article it requires analysis of the content or information in a web page.

While content based retrieval techniques can now be used to search for multi-media content, it has yet to develop techniques that can actually look at an information source, be it text or multimedia, and determines the time-stamp of the information content. In this article, we discuss the need for chronolization of web pages to capture the transient and temporal nature of information evolving on the web and discuss how SE's can use this information to sort and rank web-pages in chronological order.

The rest of the article is organized as follows. In Section 2. we introduce the functioning of the SE's and show how they have been evolving since they first appeared. We discuss in Section 3., the need for SE rankings and discuss how SE's rank results they retrieve, discussing some well know ranking algorithms briefly. In Section 4. we discuss several mechanisms of combining SE results to come up with a rank of the ranked results. In Section 5. we discuss the temporal nature of the web and describe how one can capture aspects of time of information on the web which can be used for chronological analysis of web pages. it can not capture the full temporal dimension of information content and we summarize in Section 6..

2. Search Engine Overview

Typically SE's are web based tools for searching information on the World Wide Web. A SE provides a web interface where a short query can be entered. The retrieval system within the SE responds with a list of URLs or web pages relevant to the query in some sense. Often a system generated, relevance score is given along with each document, indicating the degree of relevance of each document to the query. The list of web pages is displayed in the decreasing order of relevance. Architecturally, a SE can be classified into two distinct components, namely, (a) storage and (b) retrieval.

The storage components not only help discover new web pages but also update previously discovered web pages on the web. They index and archive web pages discovered so that they can be presented to a user. The most important storage component is the web crawler which travels from a root page of a web page through links from it, to other pages, recursively navigating through the whole WWW, in a depth or breadth first crawl, in an effort to discover new and changed web pages. These pages are indexed and stored in an index database, where record of each document that the web crawler has crawled is maintained and updated.

User queries to a SE are accepted mostly (but not necessarily) through a web based user interface. Minimally, keywords from the query are used to search the index database, though essentially most SE's use some sort of query processing, which might also include cross language translation, query intent identification, user profiling, or collaborative filtering before the actual search against the index database.

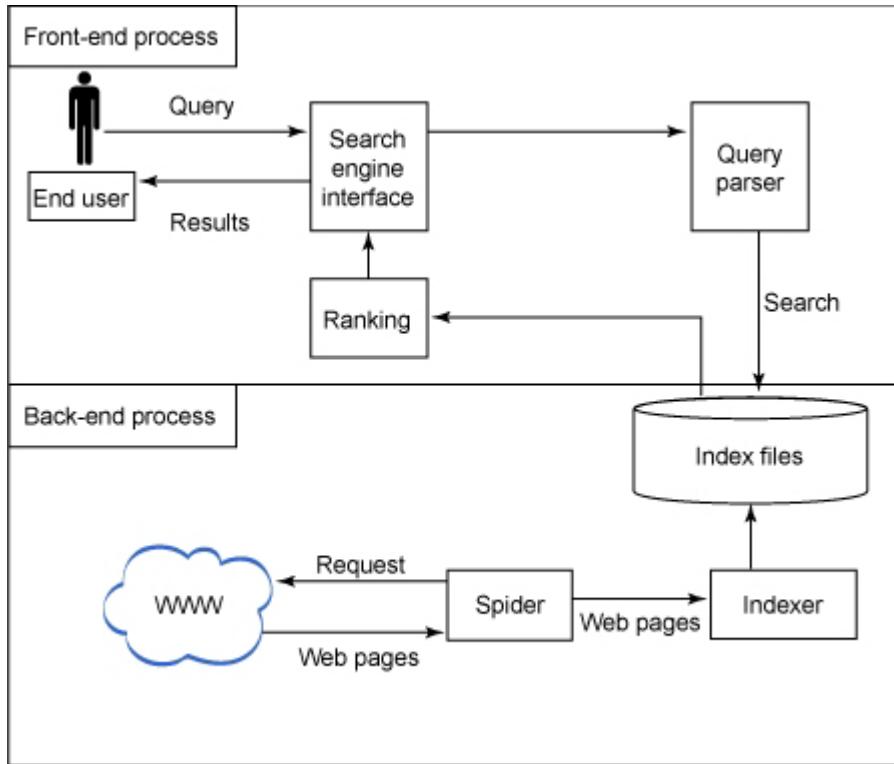


Figure 1. Typical Architecture of a Search Engine.

The retrieval components of a SE return a set of web pages which the SE considers *relevant* to the user query. However, relevance is neither a Boolean, meaning a web page might be classified anywhere between relevant (1) and non relevant (0). This varying degree of relevance for pages makes it necessary rank them in descending order of their relevance. A SE's ranking algorithm evaluates a web page relevance score and ranks the web pages in decreasing order of their relevance to the query. While the frequency of the web crawling keeps the SE updated and current the use of complex algorithms and heuristics to rank the relevant web pages determine the ability of a SE to display 'desired' information. Structurally all the SE's are same and they differ (a) in the process of how frequently they crawl the web and updated their index database and (b) the ranking mechanism used to display the list of web pages.

Zhou [1] description of a SE is illustrated in Figure 1. Zhou refers to retrieval components as front-end processes and refers to storage components as back-end processes.

3. Ranking Algorithms Overview

We start this section by questioning the need for ranking search engine results. When a user passes a query to a search engine, he is looking for information relevant to a query. However, what is relevant to one person, might be useless or non relevant to another. As search engines have no method of detecting query intent to a specific level, it is hard to

determine what relevant information for a particular user is and what is not. Also queries to a search engine are in natural language, not in a precise, machine interpretable language. Sometimes queries are malformed and might not even give accurate information about what the user wants. There is, thus, a need for returning more than one relevant web page in a ranked order of system computed relevance, which can be accurate to some degree.

Ranking based on system computed scores for retrieved web pages is a well explored area. System Relevance scores are computed by matching query terms with words or sequence of words within a web page. Traditional measures include Term Frequency (TF), Inverse Document Frequency (IDF), TF-IDF (Term Frequency-Inverse Document Frequency) as well as other more complex measures such as Okapi BM 25 [2] and LMR [3].

Ranking based on system computed scores for retrieved web pages is a well explored area. According to simple document ranking techniques typically view the meta tags and the text content of a web page and determine its relevance to a search query by using a variety of information such as keyword meta tags, URL information, title of the website, frequency of a query keyword, keywords in section headers, graphics, overall size of a page and proximity of keywords defines how closely correlated are the keywords.

A good example is the proximity search technique borrowed from text processing. Proximity search technique involves looking for multiple query terms that are found in a document and appear within a certain distance from each other. Sometimes proximity, searches can also involve looking for bi-grams and tri-gram sequences that match for a query and a document.

Link analysis algorithms use the structure of Internet hyperlinks pointing to a page as an effective indicator of the relevance and importance of a web page. If a web page is cited by other web pages it is considered popular. Notable link analyzing algorithms are the PageRank algorithm [4], and HITS algorithm [5]. Teoma [6] owned by Ask.com uses the HITS algorithm while Google.com [7] uses PageRank algorithm. The diagram below outlines ranking algorithms and functions.

3.1. PageRank Algorithm

PageRank extends the idea of what has been used in academic citation literature for a long time to determine the importance or quality of a page by largely counting the number of citations or back links to a given page. PageRank extends this idea by (a) not counting links from all pages equally and by (b) normalizing by the number of links on a page [4]. In case of a Topic-Based PageRank, teleporting takes place within a specific topic. So when there is a page which does not link to any other pages within a specific domain, teleporting allows a jump to another page within the same domain.

3.2. TrustRank Algorithm

Many changes in the ranking algorithm have been seen to counter spamming. The issue of web spamming to fool SE's has been one of the major test of ranking technology. Spamming pages are typically commercially driven and aspire to boost their rankings in SE's result lists returned. Use of link farms, what provide a huge number of links to spam pages is a common approach takes by spammers. This typically beats link analysis algorithms.

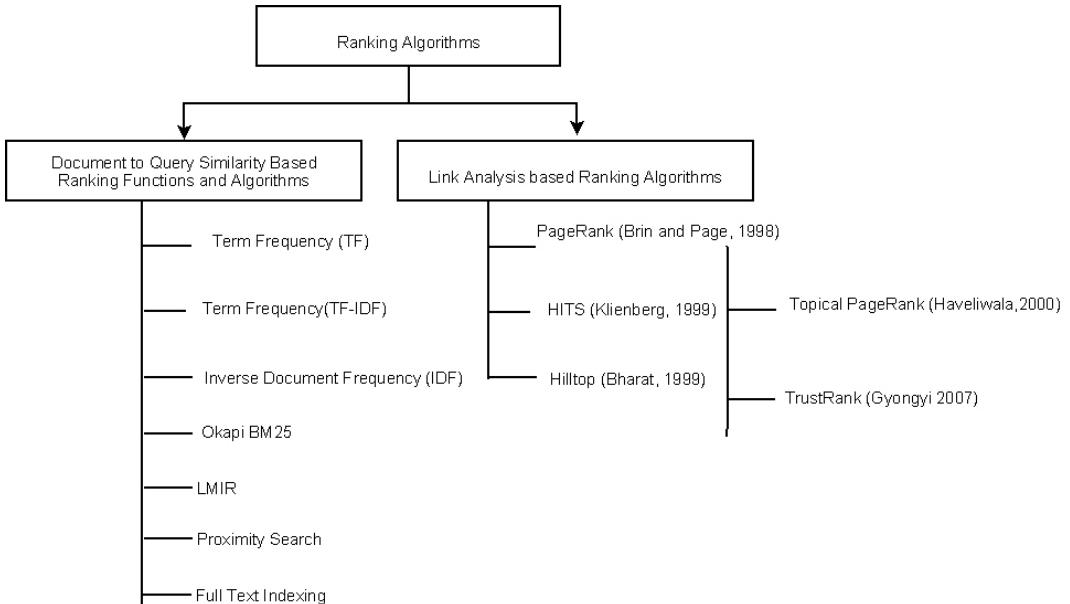


Figure 2. Summary of Ranking Algorithms.

Gyongyi et. al. [8] proposed TrustRank for combating spam. TrustRank is a link analysis algorithm that starts with a small set of pre evaluated expert pages. These pages are the starting points for an outward crawl to seek out similarly trustworthy pages. As the crawl moves outward, the level of trust goes down. The trust score is factored in while ranking web pages.

3.3. Hilltop Algorithm

Hilltop algorithm [9] is somewhat similar to TrustRank as it employs *expert pages* and hyperlinks emanating from these pages. Expert pages are pages about specific topics that are trusted in their content. There are links from these pages to other non-affiliated pages. Some pages that are cited by a lot of expert pages are considered authorities and obtain a higher rank. The key is defining affiliation. The original algorithm states that two pages are affiliated if the first three octets of the Internet Protocol (IP) address are the same or the rightmost non generic token in the host name is the same. For example, www.bbc.com and www.bbc.co.uk would be affiliates. The most relevant expert pages and the relative distance of pages retrieved from them are used to rank the latter.

3.4. Hyperlink-Induced Topic Search (HITS)

Hyperlink-Induced Topic Search or HITS [5], is another web PageRanking algorithm that employs link analysis. The algorithm calculates two values for a page: (a) its authority, which estimates the value of the content of the page, and (b) its hub value, which estimates the value of its links to other pages. It is also known as the Hubs and Authority algorithm. The hub and authority determination is restricted to the results retrieved in response to a

search query. The algorithm proceeds by recursively determine authority and hub values with an authority value being calculated as the sum of the scaled hub values that point to that page. In turn, hub values are calculated as the sum of the scaled authority values of the pages it points to. The relevance of linked pages is employed in calculating the hub and authority values.

4. Metasearching

A metasearch engine is a tool to search the web using multiple SE's in parallel. Metasearch engines have become increasingly useful with the expansion of the Web and the limited coverage of any one SE. Metasearch engines can additionally combine results from specialized SE's. A metasearch engine passes a query to multiple SE's, retrieves results from each of them in the form of a result list, and then combines these results into a single result list. Two key components of a metasearch engine are (a) the query dispatcher which decides to which SE's to send the query and (b) the result merger which decided how to merges the results from multiple SE's.

There are two types of metasearching: (a) external metasearching which involves merging results from autonomous SE's which work independent of each other. There may be an overlap in the web pages included in the search results. Typically, this type of metasearching happens when only result lists are available without any relevance score information; (b) internal metasearch, on the other hand, multiple sub SE focus on different information sources, within the same database. They return results in the form of result lists. These lists are then merged using a result merging model. In an internal metasearch engine, relevance scores are generally available from each sub SE.

4.1. Ranking Metasearch Results

Ranking of results returned by multiple SE's can be looked upon as a multi-criteria decision making problem. Each SE returns its own ranked list of web pages. The task of ranking involves merging these ranked lists into a single ranked list, thereby assigning ranks to web pages based on some combination of ranks determined by different SE's.

Result merging for metasearch is the application of data fusion techniques to search results. Data fusion techniques have been applied to develop result merging models in the past. Early research in this field includes the Logistic Regression Model [10], and the Linear Combination Model [11], [12]. Aslam and Montague proposed two models [13], the Borda-Fuse and Weighted Borda-Fuse. Diaz [14, 15] came up with a comprehensive linguistic quantifier guided fuzzy result merging model based on the Ordered Weighted Average (OWA) operator.

The Borda-Fuse model was proposed by Aslam and Montague [13] and is based on Borda-Count [16]. Each SE ranks a given set of documents. For each SE, the top ranked candidate is given d points (called Borda points). The next document receives $d-1$ points and so on. The documents are ordered according to the total number of points, gained due to their position on multiple SE's. The document receiving the most points is ranked at the top. Weighted Borda Fuse model [13] determines document ranks as sum of the product of the weight attached to a particular search engine result list and the number of points

accumulated by the document in that search engine results. Weights can be based on an overall assessment of the performance of the SE such as its average precision.

Diaz [14, 15] developed a fuzzy result merging model based on Yager's [17] Ordered Weighted Average (OWA) operator. The OWA operator uses a multi-criteria decision making approach where a decision function F is constructed by means of which one can combine several criteria and evaluate the degree to which an alternative satisfies the criteria. Diaz [14, 15] applies the OWA operator in result merging by determining the ranks as the degree to which the web pages (alternatives) satisfy the SE (criteria). The OWA model was further extended for metasearch by De et. al. [18]. Experimental Results [14, 15, 18] have shown improved result merging performance when using fuzzy models over Borda Fuse model.

5. Temporal Ranking of Web Pages

While most of the research in the area of search engine result ranking has been on the aspect of trying to identify the most popular or most relevant set of pages so as to enable SE's to display the results in the decreasing order of relevance. One aspect that has not been addressed is the aspect of chronological ranking of the web pages. Today SE's provide this aspect of ordering/ranking on news articles. It should be noted that there is a strong correlation between the date of publication of the news article and the contents of the news article. This correlation allows SE's to rank the articles based on the currency of the news article, thus allowing users to search for news articles which were created in the last one hour, last one week and so on. This chronological ordering or ability of a SE to determine the age of the page is possible only because there is an explicit reference to date and time in the news article. Though not explicitly mentioned in technical articles, a form of chronolization can be observed in scientific articles where one can determine the age of the article in terms of the *not published before date* by looking at the list of references (mandatory in scientific articles). The date of publication of the most recent cited reference in the list of references is the date *after* which the article has been published.

Chronolization has two aspects, (a) the date of creation or modification of the article (popularly called time stamp of the article); this is researched under the topic of Web Archiving² and (b) the reference in time to which the content of the article refers to. Though not mandatory, in many cases, as is evident in news articles, both the time-stamp and the reference in time of the information content might be same. But in general this is not true, for example, an article describing an incident in World War II could be written in 2009 – in this case the time stamp of the article is 2009 while the time reference of the article is 1939-45³. In another scenario, there can be an article which describes a scenario of the future (Sci-Fi article) written in 2009 about the nature of environment in 2020. Here the time stamp is 2009 but the information content is timed to the year 2020 (into future). Clearly there is a need to distinction between the time stamp and the time associated with the information in a page. In this article we refer to chronolization with respect to the time of the information content which is important for user searching for information.

²A popular web site archiving is hosted at achieve.org

³1939-1945 refers to the actual time when the World War II was fought

5.1. What Is Web Page Chronologizing?

Chronologization should not be confused with web archiving which takes snapshots at different times of a web page and store them in an ordered fashion based on the time of the snapshot. Chronologization is the ranking of web pages in descending order of currency of the content. If a page has information on a topic which is current then it should be ranked at the top. Chronologization helps us determine and hence rank on top the most *recent* information on a topic in response to a query and not the web page that has been most recently altered. The last updated web page on the topic might not contain the most recent information. There is a need for ranking of web pages based on chronological ordering. For example, consider a query *Monsoon in Mumbai*; in this case, using chronological ordering, a page about the current monsoon situation in Mumbai would be ranked higher against the most popular page which refers to one of the worst monsoon in Mumbai⁴. In another example, look at a student working on a school project or a research project; (s)he would want to be able to see the most recent articles (which is likely to contain recent findings) first and the older articles (whose content or theory be obsolete) later. Ability to determine the age of the content is non-trivial and requires deep language processing, especially when there is no explicit reference to time on the page. One could in theory build a chronological WWW where each article is related to every other article in relation to time but this is hardly necessary. One could create a chronological ordering of the web pages returned by a SE (or a set of SE's) in response to a query. In this case, we need to time order a finite and a small set of web pages; which makes the chronologization process tractable and less ambiguous. Ranking based on chronologization would benefit; today none of the SE's gives an option to visualize the results based on chronological order.

5.2. How To Achieve Chronologization?

Web archiving can record snapshots of pages with time stamps attached to them. Using these snapshots it is easy to track changes to the page and create a time line in which information appears and is removed from the page. However, it difficult to say if the current contents of the page contain the most recent information on the topic. It is also difficult to rank pages on the currency of information on a topic. A trivial way to do this would be to select and rank pages in the order of their last modified times (time stamp) obtained from web archives.

Consider there are N result pages that the SE⁵ returns in response to a user query. Chronologization would mean to rank the N result pages in an order that would place the page with the current information at top and rank it high while the page containing later or older information would be ranked lower. Let P_1, P_2, \dots, P_N be the N pages returned by a SE in response to a query. The idea is to analyze the content of the page and determine the *PublishDate* and *ContentDate*. The *PublishDate* would be a reference to the date on which the page was published; this can be determined by using several cues like (a) using the information in archive.org, (b) looking at the metadata for the publishing

⁴Today almost invariably all the SE's pop up the page corresponding to the monsoon flooding of Mumbai on July 26, 2005

⁵could also be from multiple SE's

Table 1. Event-Time Knowledge Base

Event	Time
World War II	1939-1945
Christmas	Dec 25
Gandhi birthday	02-10-1869
:	:
Children's Day	14 November

date, (c) look at the HTTP address, (d) look for a time stamp⁶ in the text, (e) look for an explicit *Last Updated* like string. If there are several *PublishDate*'s, we choose the one that is closest to the current date and time⁷. Determining *ContentDate* which is useful to rank the pages in the chronological order requires language processing on the contents of the page and an access to some kind of a event-time knowledge base. A typical event-time knowledge base is shown in Table 1. One could use a minimal parsing system [19] to identify the *ContentDate* of the page. Typically, a single article could have multiple *ContentDate*'s especially if the content refers to information over a period of time⁸. All the dates associated with the *ContentDate* are meaningful and are used to rank the pages in chronological order. A simple way is to assign multiple ranks with the page having multiple *ContentDate*'s. For example, if there are two pages P_1 and P_2 ; let the *ContentDate* of P_1 be 15 Aug 1947 and 26 Jan 1950 and the *ContentDate* of P_2 be Dec 1949. Then we would rank the pages as P_1 (due to *ContentDate* being 26 Jan 1950), P_2 , P_1 (due to *ContentDate* being 15 Aug 1947). In the rest of this section we give a heuristics based approach to order the pages in a chronological order.

Natural language processing (NLP) can be used in ranking retrieved web pages based on the information currency. Following this, NLP can be used to read through (parse) the content of web pages and segment it into sentences. The grammatical tense (present, past and future) of each sentence can be determined using corresponding rules for verb conjugation and sentence structure using NLP [20]. If a majority of the sentences in the page are written in a specific tense then we can say that the *tone of writing* of the page is in that particular tense. If the tone of the page is in present tense and the time stamp on the page is close to the current time, then it is highly likely that the page is current should be ranked high. Table 2 below simple heuristics to rank a page into three categories based on possible currency of information. A brief description of Table 2 is given below.

Heuristic 1 (Present Tense; Date approximately Current Date) This implies that most likely the web page was written recently and since it has been written in the present tense it indicates that it is describing a recent even or happening. Hence its ranking should be chronologically higher.

⁶could be in different formats

⁷system date and time

⁸a page related to history

Table 2. Heuristics to determine chronological ranking of web pages

Heuristic	Tone of Web Page	Time Stamp	Chronological Ranking
1	Present	Date \approx Current	High
2	Present	Date older than Current	Low
3	Past	Date \approx Current	Low
4	Past	Date older than Current	Low
5	Future	Date \approx Current	Medium
6	Future	Date older than current	Medium

Heuristic 2 (Present Tense; Date older than Current Date) This implies that most likely the web page was written in the past and even though it was written in the present tense, at this time it is referring to a past even or happening. Hence its ranking should be low in the search engine list.

Heuristic 3 (Past Tense; Date approximately Current Date) This implies that most likely the web page has been written recently but recounts a past even or happening. Hence its ranking should be low in the search engine list.

Heuristic 4 (Past Tense; Date older than Current Date) This implies that the document was written in the past and recounted a historic event then. Hence its ranking should be low.

Heuristic 5 (Future Tense; Date approximately Current Date) This implies that the document was written recently but seems to recount a future even or happening. Hence it is highly likely it is reporting on a current or future event and the information content is the latest. Hence its ranking should be high

Heuristic 6 (Future Tense; Date older than Current Date) This implies that the document was written in the past but seem to recount a future even or happening. The even might have happened in the recent past or might happen in future. There is a higher likelihood of its happening in the present time. Hence the ranking should be medium.

Figure 3 and Table 2 show a scheme for chronological ranking of SE results into three categories high, medium and low based on how current the information contained in the page is. By identifying the *tone* or most prevalent tense of the textual content of the web page and determining last modified time-stamp of the web page we can rank web pages into three categories high, medium and low. When displaying results in a chronological order, web pages classified as high are displayed first, web pages classified as medium are ranked after that and web pages ranked low are classified at the bottom. Standard ranking algorithms such as PageRank can then be used to rank web pages within a category.

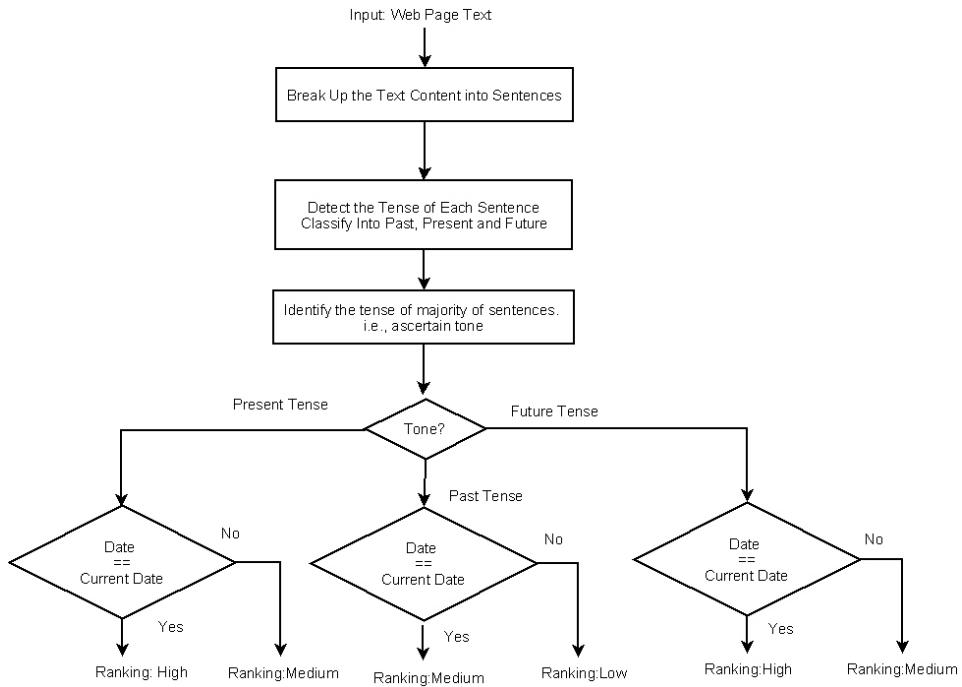


Figure 3. Ranking Web Pages by Temporal value of information content.

6. Summary

In this chapter, we have discussed the fundamental architecture of a search engine (SE). A search engine explores the web through a crawler; indexes, stores and archives web pages. A SE parses the query and in response retrieves a set of results relevant, in some way, to the query and displays them in an ranked sequence. Ranking is very crucial and the popularity of a SE rests in the way it ranks the search results. In this chapter we provided an overview of some of the most widely used ranking algorithms used by most SE's. We also discussed the temporal nature of the web and demonstrated the need for ranking results in a chronological order. Most chronological rankings are employed use time-stamp information and reference well established web archives. However, these web archiving based information, use the creation time of a page instead of the actual information content of a page. We proceed to outline a NL-based approach to determine the age of actual information content and an approach to rank web pages based on this approach.

References

- [1] Deng Peng Zhou, “Beef up web search applications with lucene: Improve searches with a more robust app from the apache jakarta project,” in *IBM Developer Works*, 2006.
- [2] Robertson S. E., “Overview of the okapi projects,” in *Journal of Documentation*, 1997, vol. 53, pp. 3–7.

- [3] Zhai C. and Lafferty J. A, “Study of smoothing methods for language models applied to ad hoc information retrieval,” in *Proceedings of SIGIR*, 2001, pp. 334–342.
- [4] Sergey Brin and Lawrence Page, “The anatomy of a large-scale hypertextual web search engine,” in *Computer Networks and ISDN Systems*, 1998, pp. 107–117.
- [5] Jon Kleinberg, “Authoritative sources in a hyperlinked environment,” in *Journal of the ACM*, 1999, vol. 46, pp. 604–632.
- [6] Chris Sherman, “Teoma vs. google, round two,” in *Search Engine Watch*, April 2002.
- [7] Lawrence Page, “Method for node ranking in a linked database,” US Patent 6285999 for Google Inc, 2001.
- [8] Zoltn Gyngyi, Hector Garcia-Molina, and Jan Pedersen, “Combating web spam with trustrank,” in *Proceedings of the International Conference on Very Large Data Bases*, 2004, vol. 30, p. 576.
- [9] Krishna Bharat and George A. Mihaila, “Hilltop: A search engine based on expert documents,” in *WWW9 Conference*, May 15-19 2000.
- [10] D.A. Hull, J. O. Pedersen, and H. Schtze, “Method combination for document filtering,” in *Proceedings of the 19th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, August 1996, pp. 279–287.
- [11] P. Thompson, “A combination of expert opinion approach to probabilistic information retrieval, part 1: The conceptual model,” in *Information Processing and Management*, Nov. 1990, vol. 26, pp. 371–382.
- [12] P. Thompson, “A combination of expert opinion approach to probabilistic information retrieval, part 2: mathematical treatment of ceo model,” in *Information Processing and Management*, Nov. 1990, vol. 26, pp. 383–394.
- [13] J. Aslam and M. Montague, “Models for metasearch,” in *Proceedings of the 24th annual international ACM SIGIR Conference on Research and Development in Information Retrieval*, 2001, pp. 276–284.
- [14] A. De E. D. Diaz and V.V. Raghavan, “A comprehensive owa-based framework for result merging in metasearch,” in *Rough Sets, Fuzzy Sets, Data Mining, and Granular-Soft Computing*, Sept. 2005, pp. 193–201.
- [15] E. D. Diaz, *Selective Merging of Retrieval Results for Metasearch Environments*, Ph.D. thesis, University of Louisiana, Lafayette, LA, May. 2004.
- [16] J. C. Borda, “Memoire sur les elections au scrutin,” in *Histoire de l'Academie Royale des Sciences*, 1781.
- [17] R. R. Yager, “On ordered weighted averaging aggregation operators in multi-criteria decision making,” in *Fuzzy Sets and Systems*, July 1983, vol. 10, pp. 243–260.

- [18] Arijit De, Elizabeth E. Diaz, and Vijay V. Raghavan, “On fuzzy result merging for metasearch,” in *FUZZ-IEEE 2007*, 23-26 July 2007, pp. 1–6.
- [19] Sunil Kumar Kopparapu, Akhlesh Srivastava, and P. V. S. Rao, “Minimal parsing key concept based question answering system,” in *HCI (3)*, Julie A. Jacko, Ed. 2007, vol. 4552 of *Lecture Notes in Computer Science*, pp. 104–113, Springer.
- [20] Qiu Gui Su, “Determining time frames,” in <http://mandarin.about.com/od/grammar/a/aspect.htm>, 2009.

Chapter 15

TOWARDS ON DEMAND IT SERVICE DEPLOYMENT

Jai Dayal², Casey Rathbone², Lizhe Wang¹ and Gregor von Laszewski^{1,*}

¹Pervasive Technology Institute, Indiana University

2729 E 10th St., Bloomington, IN 47408, U.S.A.

²Service Oriented Cyberinfrastructure Lab, Rochester Institute of Technology
Bldg 74, Lomb Memorial Drive, Rochester, NY 14623-5608

Abstract

Complex IT systems allow users to create, organize, and share the users' services and computing resources. As these IT systems become more complex, the more difficult the application deployment process becomes. Deployment, the process of making the application or service available for use, often requires the installation, customization, and configuration of many inter-operable heterogeneous system components. The application developer must understand the many dependencies and configuration parameters required to enable interoperability between the components.

In this chapter, we will present the deployment solution for a live complex IT system, the Emergency Services Directory (ESD). Many different technologies exist that attempt automate the application deployment process by allowing a user to describe the dependencies, provide the configuration parameters, and specify the application's required technologies, such as the operating system, database or Web server technology.

ESD's deployment solution takes advantage of the benefits provided by virtualisation, and virtual appliances in particular. Each of ESD's components are wrapped and contained within a virtual appliance image. To automatically and on-demand deploy the virtual appliance image, we use the Cyberaide Creative tool, which a tool in the on-going Cyberaide project.

1. Introduction

Typically, an IT service or application consists of several components inter-operating to perform a set of functions or tasks. Service deployment can be considered as the process of acquisition and execution of the service, i.e., making the service ready for use [1, 2]. A

*E-mail address: laszewski@gmail.com

deployment process typically consists of several operations and configuration parameters and there often exists several dependencies between these operations and parameters. Deployment typically consists of the release of the software, the configuration of the software, and the installation of the software [1].

Depending on the application, the deployment process can be quite complex requiring a user or developer to have a detailed understanding of the applications individual components. Additionally, an IDC survey states that out of the \$95 billion dollars spent on application operating costs, 19% can be attributed to the cost of deployment alone [3].

With today's advanced IT infrastructures, such as compute Clouds, many users, who typically can benefit from such infrastructures, avoid using these technologies as the users find them too complex and having too steep of a learning curve. The complexity of these systems is largely due to the vast heterogeneity of the components contained with distributed infrastructures. Applications require specific operating system drivers and configurations as well as software libraries and packages [2]. In heterogeneous environments, access to resources matching the application's specific requirements can not always be guaranteed. In such an environment, a user may be required to provide several version of the application to match the infrastructure's different resources [4].

To help assuage some of the mentioned problems with heterogeneous environments, users turn to virtual machine (VM) technology. VMs offer users the following benefits [5]:

- Customized OS: The user has great flexibility to customize the operating system to meet the application's run-time requirements. These customizations can include, but are certainly not limited kernel level customizations to memory and storage customizations.
- Ease of Management: Virtual Machines can easily be shutdown or restarted, easing the system reconfiguration process. Additionally, VMs can easily be migrated to different physical machines, thus allowing the application to easily operate during a physical machine's downtime.
- System Security: VM users can define the operating system's privilege without affecting the privileges of the underlying operating system. For example, root access is often required to install operating system modules. Installation of these modules only effect the VM. Additionally, if a configuration causes the system to crash, only the VM is affected, while the underlying machine and operating system remain operational.

While VMs provide us with a flexible and easy to deploy platform, VMs alone fall short of automatic and on-demand service deployment. For example, each time the VM is shutdown, the service will have to be re-deployed.

Virtual appliances take VMs a step further by containing both the platform and the service. Launching and executing virtual appliances, however, still requires the user to understand many technical details. For example, a user must understand which basic software packages or components are needed for the application. The user must also understand the functional dependencies between these packages. For example [2] stats that IBM DB2 has approximately 40 configuration parameters that must be resolved during the deployment process.

Our contribution uses the Cyberaide Creative tool to allow a user to select the appropriate virtual appliance image, and handles the resource selection, resolves the dependencies between components and packages, and automates the deployment, shielding the user from the process. Using ESD as a case study, we evaluate Cyberaide Creative's success.

The rest of this chapter is organized as follows. Section II presents some background information and compares our solution to the existing solutions in the field. Section III discusses and formulates the deployment process. Cyberaide and Cyberaide Creative are presented in Section IV. Section V describes the ESD project, our case study, and Section VI presents our deployment solution. Section VII evaluates our solution and our conclusions are presented in Section VIII.

2. Background and Related Work

Automatic and on-demand service deployment is not a new idea, and in fact, there currently exist several tools to facilitate the deployment process. A very common one is the package managers found in various Linux distributions such as Fedora and Ubuntu. These tools consist of a base repository containing a number of packages. These tools are fully capable of resolving dependencies amongst packages and fully automate the installation process. However, these tools are platform dependent and are not used in highly heterogeneous distributed environments.

Dearle [1] discusses the general concept of automatic on-demand service deployment, and states that the future of this paradigm will rely heavily on virtualisation. Dearle also examines six current technologies that facilitate the automatic deployment of services. This discussion, however, does not present a architecture or framework, but suggests possible technological solutions.

Kecskemeti et al [4] presents a detailed framework to deploy a service in a heterogeneous environment, but the work seems to present the automatic deployment of computing platforms, such as the operating system and database technology. The work also does not provide a specific application and an evaluation to determine the effectiveness of their proposed framework.

In [2], Sun et al provide a good discussion and formulation of the deployment process. They also provide two typical service platforms, such as a LAMP stack on a single node, which closely relates to the ESD project, and a Web service architecture in a distributed environment. To facilitate the deployment process, they use virtual appliances. Their work, however, only provides the service platforms and does not discuss the steps needed to migrate a full application or service to the platform. Also, this work does not propose a design or framework for a tool to facilitate the deployment of the virtual appliance, which they appear to do manually.

Our work differs from the above studies by providing a framework that delivers two parts, the first is the automatic on-demand deployment of the service platform using virtual appliances, and the second is the on-demand automatic deployment of the service to the platform. We show how this can be achieved using Cyberaide Creative to deploy the virtual appliance, and a set of scripts to migrate the service codes and data.

3. Service Deployment Process

In this section, we formulate the general process of deployment and present a simple model to formulate and discuss the complexity of a deployment process.

3.1. Deployment Process Overview

Deployment takes place at the end of the software life-cycle, hence it is a post-production activity [1]. There are many deployment guidelines, strategies, and tools, but typically share the same general deployment process [2]. These guidelines also tend to vary based on the type of deployed service. Services needing only a single machine typically have a more straight forward approach than do distributed services. Based on [2, 1, 6], we can generalize the deployment process:

- Determine the dependencies that exist in the deployment process. For example, if a component requires a database, the database installation and configuration should come first.
- Understand the communication and relationships between the different machines needed for the application.
- Resolve the dependencies at the platform layer, for example, the installation of a component might require a specific compiler.
- Install the application. This requires moving the source files, binaries, and etc. to the targeted environment. This step also requires the user to follow the dependencies required by the service.
- Activate the service and monitor the service's operation to identify any malfunctions.

3.2. Deployment Model and Problem Definition

A deployment process consists of several identifiable elements, a service, a number of operations, and a set of dependencies which specify the order in which the operations must occur.

$$\text{Deployment} = \{\text{Service}, \text{Operations}, \text{Dependencies}\} \quad (1)$$

A software service is composed of any number of components, where a component is the minimum entity in the service. Formally, a service *service* is modeled as:

$$\text{Service} = \{\text{component}_i \mid 1 \leq i \leq I\} \quad (2)$$

where I is the number of individual components in the service. The set of operations *operations* can be formulated as:

$$\text{Operations} = \{\text{op}_j \mid 1 \leq j \leq J\} \quad (3)$$

where J is the total number of operations in the deployment process.

Typically, an operation op_j requires several configuration parameters *parameters*:

$$\text{parameters} = \{\text{param}_k \mid 0 \leq k \leq K\} \quad (4)$$

where K represents the number of configuration parameters for operation op_j .

The dependencies of a deployment process represent the order in which the operations must be performed. We model the dependencies as a directed graph, *Dependencies*:

$$\text{Dependencies} = \{\text{Operations}, D\} \quad (5)$$

where *Operations* is the set of vertices and D is a set of edges, or dependencies, between operations:

$$D \subseteq \text{Operation} \times \text{Operation} \quad (6)$$

For example, a dependency (op_1, op_2) means that op_1 depends on op_2 , i.e., op_2 must happen before op_1 . [7, 8] further discuss the modeling of software systems and dependencies. Figure 1 depicts a simple dependency graph. Deployment technologies such as RedHat's RPM use dependency graphs to determine and resolve dependencies between software packages.

As we can see, as the number of components and operations grow, the more labor intensive the deployment process becomes for the user, especially in regards to the dependencies between operations. Our goal is not to reduce the number of dependencies or operations in a deployment process, but rather to *hide* the deployment complexity from the user.

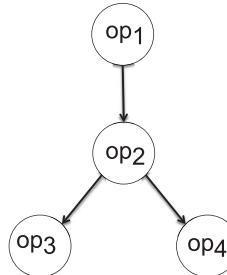


Figure 1. Simple Dependency Graph.

4. The Cyberaide Project and Cyberaide Creative

This section provides an overview the Cyberaide project and Cyberaide Creative, a tool within the Cyberaide project. Cyberaide Creative is the tool we use to automatically and on-demand deploy our ESD service.

4.1. The Cyberaide Project

As mentioned in the introduction, users can benefit from complex IT infrastructures in a number of ways, but users often have a hesitant attitude towards using these technologies due to the amount of technical knowledge these technologies require. Cyberaide provides a possible solution to this problem. Several tools have been integrated into the Cyberaide project, such as the Cyberaide Toolkit and the Cyberaide Shell [9].

Cyberaide enjoys the following essential features [9]:

- *Ease of use*: make the JavaScript based API and interfaces useful for Grid and Web developers.
- *Low installation footprint*: support fast downloads as well as an easy maintenance through a small manageable code base.
- *Security*: gain access to Grid resources in order to avoid compromising the system. This is especially important due to known limitations of JavaScript.
- *Basic Grid functionality*: is provided for developers to create Grid-based client applications.
- *Advanced functionality*: is offered as many developers do not want to replicate functionality provided by other Grid middleware and upperware.

The framework is designed in layers and comprised of different components. (see also Fig. 2). A web client that provides access to Grid functionality and components that can be deployed in a web server are provided. A service called “*mediator service*” mediates tasks to the Grid and basically is a secure server that provides most of the functionalities in regard to the Grid.

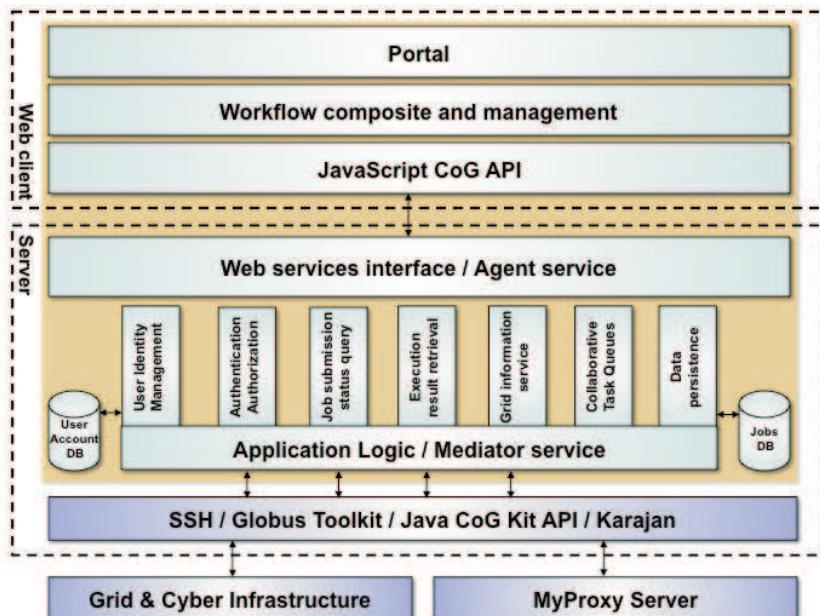


Figure 2. System Architecture.

- Web client: provides elementary functionality to access the Grid through a portal user interface.
- Server: contains two logical parts:

- Agent service: is the intermediate service between Web client and mediator service; works as proxy for users to interact with the mediator service.
- Mediator Service: is the bridge between the Grid and the client library. The mediator service offers different functionalities and contains the application logic.

Because of the separation between the service and the client the development of Cyberaide shell was possible. this is a system shell that facilitates the use of cyberinfrastructures. It contains four high level design components: object management, cyberinfrastructure backends, command line interpreters, and services (see Fig. 3) [9].

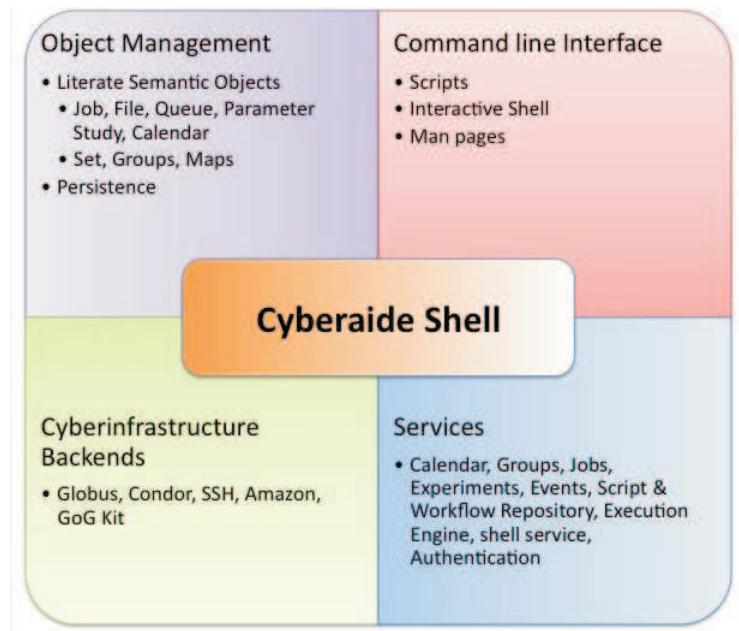


Figure 3. High level design of Cyberaide Shell.

4.2. Cyberaide Creative Paradigm

The paradigm presented using Cyberaide Creative is an on-demand resource allocation system. This creates an environment where resources can be optimally utilized as demands request them. The Cyberaide Creative system is used as a tool for acquisition of the production grid running either Condor or Globus ToolKit and Cyberaide Gridshell is the operating interface to the grid. The benefits of this paradigm are the ability to outsource resources for less cost than to maintain a complete internal system for peak resource consumption.

Figure 4 shows the computing paradigm of Cyberaide Creative:

1. Users send requirement to Cyberaide creative to demand cyberinfrastructures from Clouds, for example, a condor cluster, or a computational Grid with Globus Toolkit as a middleware.
2. Cyberaide creative then construct a cyberinfrastructure for users, which is pre-installed some Grid middleware, like condor, Globus and Cyberaide shell.

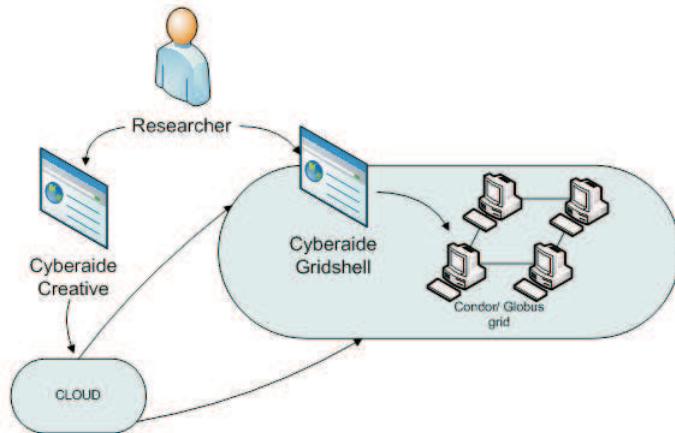


Figure 4. Cyberaide Creative Paradigm [10].

3. Users then on demand access the cyberinfrastructure with aides of cyberaide shell.

[10] presents Cyberaide Creative's use cases. For our deployment, we are using the scenario where a user requests a single VM workstation. This work extends this scenario by deploying a fully configured and ready to use virtual appliance. Cyberaide Creative uses VM Ware ESXi technology to create and store the virtual appliance images. For a more in depth discussion on Cyberaide and Cyberaide Creative, see [9, 10].

5. Emergency Services Directory

In this section, we will describe the ESD service and we will formulate the operations and parameters required to deploy this service.

5.1. Overview

The Emergency Services Directory (ESD) is a nation wide directory containing information about emergency service resources, such as fire departments, ambulances, and law enforcement agencies. The directory aims to meet the needs of several different classes of users from the general public, to state and national level government officials. Each different class of user requires different levels of access to the data, and also needs to perform role specific tasks. For example, a regional fire 5 chief may need to send an alert all near by fire stations, while ambulance providers may need to know what capabilities neighboring ambulance providers offer.

The directory also has an on-demand printing service that allows designated users to export information from the database into a customize format, such as a directory, booklet or pamphlet. The Society for Total Emergency Programs (STEP) council currently distributes a printed version of the directory to 911 dispatchers, ambulances, and other emergency service providers in Rochester, NY area. Additionally, the printing service must be scalable so multiple users can print on-demand customized documents simultaneously. The printing

service consists of a set of Perl and shell scripts and can be operated independently from the Web site.

This real-life health-care application is categorized as a complex IT system as it relies on harmonious operation between independent technologies. Build on top of a LAMP (Linux, Apache, MySQL, and PHP), the on-line directory uses the Drupal content management system [6] to handle user requests, and to render the sites web pages. For added functionality, a developer can create custom PHP modules with in Drupal, or the developer can download and install pre-made modules supported by Drupal. ESD uses both methods.

Fine grained access to the site, and the underlying MySQL database, is controlled via role-based access control (RBAC). Drupal provides role-based access control out of the box, but this access control is limited only to content contained with in the Drupal system, such as the various Web pages and modules. Figure 5 shows ESD's design.

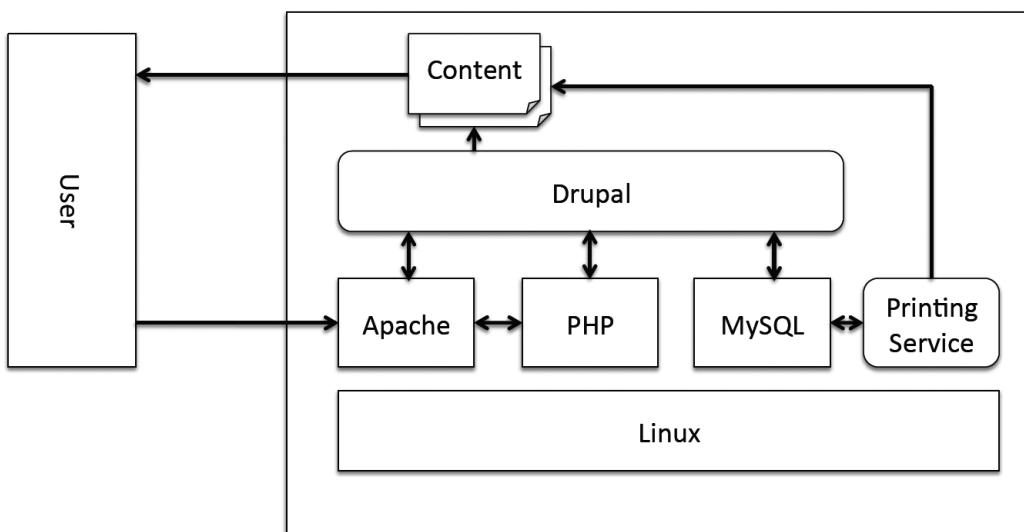


Figure 5. ESD Design.

ESD was chosen as a case study for this project because it is a real live example of a complex IT system, and has very detailed and well defined technical requirements. The deploy as a service paradigm must be able to accurately deploy a project with such detailed specifications.

5.2. Deployment Strategy Overview

The first step in the deployment process is to determine the dependencies between the components and operations in the deployment process. First, we have our base operating system, which is Linux. For our distribution, we use Canonical's Ubuntu operating system. Next, the Drupal content management system requires a MySQL database, an Apache Web server with PHP capabilities. It does not matter which order we install MySQL and Apache, but PHP's installation and configuration strongly depends on Apache. The printing service requires two tools, Perl and Latex. There are no dependencies between these tools, but naturally, they must be completed after the OS configuration. After we understand the

dependencies, we can begin resolving the dependencies by performing the operations in order.

After the LAMP stack is in place, we can install and configure Drupal. Unfortunately, there are several steps in Drupal's configuration process that can not be automated, so the user will have to be present during the deployment to select some simple configuration options. Since Drupal is modular, however, it is possible to migrate any custom modules, codes, or themes by placing them in the appropriate location in Drupal's directory structure. Since the directory structure is well defined, the migration of the modules is included in our deployment scripts.

After the MySQL database has been installed, we can begin to load the database with the appropriate information. It may seem there are no dependencies between populating the MySQL database and installing Drupal, but Drupal maintains its user and site information in the database. The data containing information about ESD's emergency services is also stored in the MySQL database. While the ESD content is not related to Drupal, the process is simplified by having only one database migration step. If the ESD information and the Drupal information were stored in separate databases, then two steps may be required.

In summary, we can define the steps required for the ESD deployment process as follows:

1. Select the appropriate operating system or VM image
2. Install and configure Apache Server
3. Install and configure PHP support for Apache Server
4. Install and configure MySQL
5. Install Perl (usually not needed)
6. Install Latex for printing service
7. Install Printing Service source files
8. Install Drupal CMS
9. Configure Drupal CMS during installation
10. Migrate database contents for Drupal and ESD

Figure 6 displays the dependency chart for ESD's deployment process.

In the next section, we show how we can contain steps 1 - 6 within the virtual appliance. By adding a few installation scripts to the virtual appliance image, we can finish steps 7 - 9 automatically.

6. Deployment Solution

In this section, we describe how we create the virtual appliance on demand, and how we then deploy the application automatically. Both steps are implemented using the Cyberaide Creative tool.

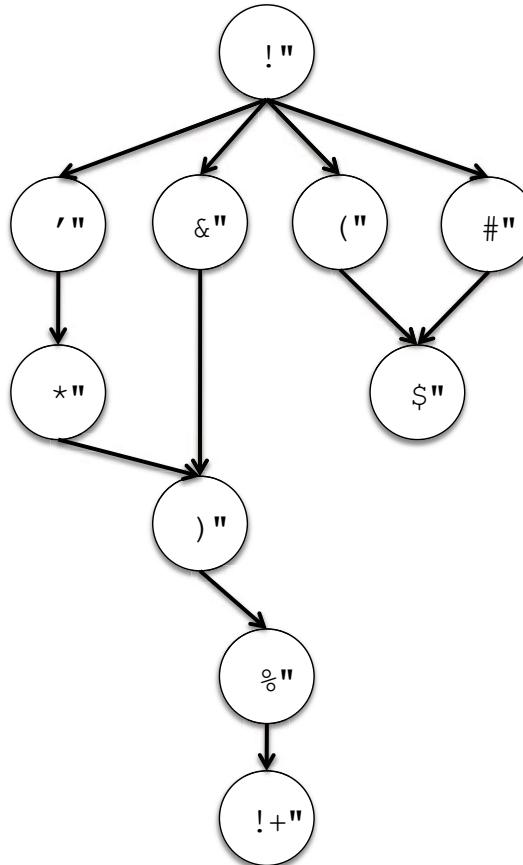


Figure 6. ESD Dependency Graph.

According to VMWare, the creation of a virtual appliance can be generalized into three steps [11]: 1) Install the guest operating system; 2) Install the specific software, such as Apache and MySQL, needed by the service; 3) Provide an interface so the user can oversee and participate in the virtual appliance creation process.

As [11] points out, it is important that the user participates in the appliance creation process as there are several configuration parameters that require a user's input, for example the network configuration. Cyberaide Creative provides us with such an interface via a Web service.

Using Cyberaide Creative [10], here are the steps needed to deploy the virtual appliance and to move the service contents to the platform:

1. The user logs into the Cyberaide Creative Web service and selects the type of VM image required.
2. The Web service sends these parameters to the ESXi Server, which then provides the required VM image.
3. The user then specifies the packages needed and the dependencies between these packages. VMWare requires that the user specifies the dependencies.

4. The user provides the Web service with any custom installation or data migration scripts.
5. The Web service then forwards these requirements and scripts to the ESXi Server, which then installs the required packages.
6. The ESXi Server then launches the appliance to a host machine, runs the scripts and returns the address and login information to the Web service.
7. The Web service forwards the address and login information to the user. The user can now directly login to the instantiated appliance.

In our implementation, scripts were written to handle the download and installation of the Drupal CMS, the creation of the database tables needed for the ESD and Drupal applications, and then for the migration of the database contents.

It should be noted that the installation of Drupal is not completed until the user logs in to the virtual appliance and agrees to certain licenses and specifies certain parameters. It should also be noted, that the VMware hypervisor only installs the appliance (OS and basic packages). The rest of the installation is done on the appliance via the provided scripts.

7. Solution Evaluation

Our goal is not reduce dependencies or operations in a deployment process, but it is to simply shield the user from the details of the deployment process. Furthermore, if the deployment process is not automated, the user will have to repeat the deployment steps each time the service needs to be re-deployed. Using Cyberaide Creative, we can create a virtual appliance image, and store and retrieve the image as needed.

In the previous section, we discussed the operations required to configure and create the appliance on top of the VM image. This process only has to be performed once, as the fully configured and created image is then stored by VMWare ESXi. From this point on, the user only has to re-launch the appliance image, while all configurations, packages, and dependencies are retained.

An important metric is how long the on-demand automated deployment process takes. This metric is highly variable as it depends on the network connection, the virtual appliance size, and the amount of data to be transferred during the migration process. In our scenario, the basic virtual image and packages is around 600MB. The amount of data in the database was less than 20MB. Additionally, the appliance images and ESD data were stored at a local repository.

Another important metric deals with the overhead caused by using a virtual machine. In some cases, added overhead may not be acceptable. Many studies [5] have been conducted to evaluate overhead caused by virtual machines. Figure 7 displays the performance of virtual machines using Linpack to measure performance.

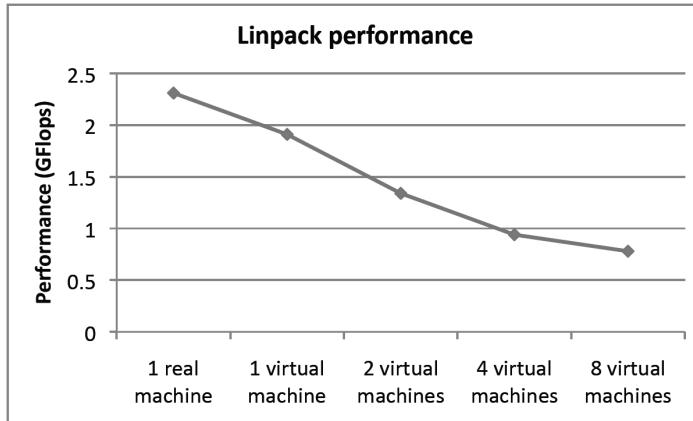


Figure 7. VM Overhead [10].

8. Future Work and Conclusion

From the above discussion, we can see that there are still several steps in the deployment process that are not yet automated, such as the configuration of Drupal. Such scenarios are a result of the application itself, and possible cannot be handled by middleware.

There are several other steps that can be dealt with, such as the configuration of the virtual appliance to operate with the host machine's network. In the future, deployment projects should be extended to include a more general deployment description language. Several studies have investigated this process [12, 13]. These tools languages are typically XML based, but this is not required. [2] discusses a deployment description language, but only in the context of their framework. In the future, Cyberaide Creative plans to make use of a deployment description language.

In summary, we have shown that using the Cyberaide Creative tool, we can simplify a service's deployment process, from the user's perspective. The user's service or application and a virtual machine image are combined into a virtual appliance. The virtual appliance is then deployed using Cyberaide Creative. Cyberaide Creative works to shield the user from the details of both the virtual appliance creation and deployment.

References

- [1] Alan Dearle. Software deployment, past, present and future. In *FOSE '07: 2007 Future of Software Engineering*, pages 269–284, Washington, DC, USA, 2007. IEEE Computer Society.
- [2] Changhua Sun, Le He, Qingbo Wang, and Ruth Willenborg. Simplifying service deployment with virtual appliances. In *SCC '08: Proceedings of the 2008 IEEE International Conference on Services Computing*, pages 265–272, Washington, DC, USA, 2008. IEEE Computer Society.
- [3] Julie Smith David, David Schuff, and Robert St. Louis. Managing your total it cost of ownership. *Commun. ACM*, 45(1):101–106, 2002.

-
- [4] Gabor Kecskemeti, Peter Kacsuk, Gabor Terstyanszky, Tamas Kiss, and Thierry Delaire. Automatic service deployment using virtualisation. *Parallel, Distributed, and Network-Based Processing, Euromicro Conference on*, 0:628–635, 2008.
 - [5] Wei Huang, Jiuxing Liu, Bulent Abali, and Dhabaleswar K. Panda. A case for high performance computing with virtual machines. In *ICS '06: Proceedings of the 20th annual international conference on Supercomputing*, pages 125–134, New York, NY, USA, 2006. ACM.
 - [6] Ashraf Aboulnaga, Kenneth Salem, Ahmed A. Soror, Umar Farooq Minhas, Peter Kokosielis, and Sunil Kamath. Deploying database appliances in the cloud. *IEEE Data Eng. Bull.*, 32(1):13–20, 2009.
 - [7] Christian Pich, Lev Nachmanson, and George G. Robertson. Visual analysis of importance and grouping in software dependency graphs. In *SoftVis '08: Proceedings of the 4th ACM symposium on Software visualization*, pages 29–32, New York, NY, USA, 2008. ACM.
 - [8] Pietro Abate, Jaap Boender, Roberto Di Cosmo, and Stefano Zacchiroli. Strong dependencies between software components, 2009.
 - [9] Gregor von Laszewski, Fugang Wang, Andrew Younge, Xi He, Zhenhua Guo, and Marlon Pierce. Cyberaide JavaScript: A JavaScript Commodity Grid Kit. In *GCE08 at SC'08*, Austin, TX, Nov. 16 2008. IEEE.
 - [10] Casey Rathbone, Lizhe Wang, and Gregor von Laszewski. Cyberaide creative: Provision grid infrastructures in clouds.
 - [11] VMWare. Best practices for building virtual appliances. Technical report.
 - [12] S. Lacour, C. Perez, and T. Priol. Generic application description model: Toward automatic deployment of applications on computational grids. In *GRID '05: Proceedings of the 6th IEEE/ACM International Workshop on Grid Computing*, pages 284–287, Washington, DC, USA, 2005. IEEE Computer Society.
 - [13] Wojtek Goscinski and David Abramson. Distributed ant: A system to support application deployment in the grid. In *GRID '04: Proceedings of the 5th IEEE/ACM International Workshop on Grid Computing*, pages 436–443, Washington, DC, USA, 2004. IEEE Computer Society.

Chapter 16

A COMPLETE PHYSICAL LAYER ARCHITECTURE FOR ROBUST HIGH RATE SPEECH WATERMARKING ON ANALOG CHANNELS AND IP NETWORKS

Simone Menci*

Dipartimento di Elettronica e Telecomunicazioni
Università degli Studi di Firenze

Abstract

High Rate Speech Watermarking is a simple yet powerful technology, proposed in [1] by K. Hofbauer and G. Kubin, for embedding digital data in speech signals. Its basic working principle is related to two well-known properties of voice signals, usually exploited by vocoders for rate compression and artificial speech synthesis, represented by *linear prediction* and *voicing state*. After the signal undergoes an LPC (Linear Predictive Coding) analysis, the obtained residual (also called voice excitation) is split up into voiced and unvoiced segments by means of a pitch detection algorithm; while the voiced segments cross the system unmodified, the unvoiced ones, thanks to their white noise spectral properties, are easily modified by watermark signal by means of a carefully chosen embedding strategy. The decoding is performed through a very similar scheme, which is advantageous for implementation. Despite this structural simplicity, the system permits us to achieve far higher data rates than previous speech watermarking systems with very limited perceptual impact on voice quality, making possible the implementation of new data channels hidden in voice transporting conversation-related data services without additional cost for bandwidth. This technology is also interesting for other applications of voice storage and transmission, both analogue and digital.

In [2] we showed its efficiency and feasibility for the case study of aircraft authentication in Air Traffic Control (ATC) communications. In this chapter we propose a complete architecture for high rate speech watermarking, with simple and efficient solutions to address three main issues, not yet addressed in the first work by original authors, with a satisfying trade-off between performance and complexity: channel coding, synchronization and residual equalization. Still using the ATC scenario as a case study, we highlight the advantages of the proposed algorithms and how they can

*E-mail address: simone.menci@unifi.it

make feasible the implementation of this efficient watermarking principle. The system is also a viable solution for the implementation of added value, analogue communication, high speed side data services or simple transmissions of, e.g., specific information as aircraft position and speed, sensors telemetry, or navigation parameters. In this chapter we also show an application example where it is used in conjunction with digital voice encoding and IP-based networks connectivity (Voice over IP) for purposes of authentication or added value services.

Keywords: Digital Watermarking, Speech Processing, LPC, Voice Activity Detection, Data Hiding, Voice over IP

1. Introduction

Digital Watermarking is a versatile technology for digital data embedding in multimedia signals [3]. Watermark was developed and applied on paper for the first time in the 13th century for purposes of document authentication. The digital format nowadays used for transmission and storage of data has allowed a modern reworking of the concept of watermarking for multimedia (images, videos, audio and voice) sources, with similar goals (mainly copyright and intellectual property rights (IPR) protection). Digital watermarking is not only targeted for secure multimedia storage, but can also have an active role in communications systems, particularly for voice and video real-time transmissions for bandwidth multiplication, data hiding, authentication and security, added value services. As an example, Speech Watermarking (SpW) techniques may provide hidden data channels for parties authentication, security or signaling purposes, or simply for additional conversation-related data services without additional cost for bandwidth, depending on achievable bit rate.

High Rate Speech Watermarking is a simple yet powerful solution, proposed in [1] by K. Hofbauer and G. Kubin, to achieve data rates of a few thousands of bits/s in a common 3.6 kHz voice host signal sampled at 8 kHz. Its basic working principle is related to two well-known properties of voice signals, usually exploited by vocoders for rate compression and artificial speech synthesis, represented by *linear prediction* and *voicing state*. After the signal undergoes an LPC (Linear Predictive Coding) analysis, the obtained residual (also called voice excitation) is split up into voiced and unvoiced segments by means of a pitch detection algorithm; while the voiced segments cross the system unmodified, the unvoiced ones, thanks to their white noise spectral properties, are easily modified by watermark signal by means of a carefully chosen embedding strategy. The decoding is performed through a very similar scheme, which is advantageous for implementation. Despite this structural simplicity, the system permits us to achieve far higher data rates than previous speech watermarking systems with very limited perceptual impact on voice quality, making possible the implementation of new data channels hidden in voice transporting conversation-related data services without additional cost for bandwidth. This technology is also interesting for other applications of voice storage and transmission, both analogue and digital.

In [2] we showed its efficiency and feasibility for the case study of aircraft authentication in Air Traffic Control (ATC) communications. In this chapter we propose a complete architecture for high rate speech watermarking, with simple and efficient solutions to address three main issues, not yet addressed in the first work by original authors, with a satisfying trade-off between performance and complexity: channel coding, synchronization

and residual equalization. Still using the ATC scenario as a case study, we highlight the advantages of the proposed algorithms and how they can make feasible the implementation of this efficient watermarking principle.

The system is also a viable solution for the implementation of added value, analogue communication, high speed side data services or simple transmissions of, e.g., specific information such as aircraft position and speed, sensors telemetry, or navigation parameters. Treating the hidden data channel as a packet-based digital link, we can multiplex various services with different priority classes. In the chapter we also show the versatility of the system through an application example where it is used in conjunction with digital voice encoding and IP-based networks connectivity (Voice over IP) for purposes of authentication or added value services. The high rate system is tested in conjunction with three popular VoIP codecs; their coexistence is evaluated in terms of specific QoS requirements, such as received voice quality and watermark extraction latency and reliability, showing the feasibility of authentication service for most evaluated conditions and the possibility of additional data services in some usage scenarios.

The chapter is organized as follows. Section 2. presents a brief review of the main concepts of digital watermarking. Section 3. presents the current state of the art for Speech Watermarking, including a brief explanation of the LPC principle and of the traditional LPC-based watermarking systems. Section 4. describes the main functional blocks of the original high rate watermarking system, as proposed in [1], and its main flaws which we addressed. Sections 5., 6., 7. present, respectively, the proposed solutions for channel coding, synchronization and residual equalization. Section 8. presents the example ATC scenario. Section 9. shows simulation results for each of the three addressed problems, together with considerations of data and speech quality. In Section 10. an example of application scenario for high rate watermarking is presented in the form of integration with VoIP codecs for the augmentation of an ATC terrestrial network. It details the problems that a voice transmission must face on an IP network and the motivations for an alternative secure data channel hidden in voice signal, shows the example scenario used for the evaluation of watermarked VoIP performance, and presents simulation results in terms of data and speech quality. Finally, some conclusions and perspectives for future work are given in Section 11..

2. Digital Watermarking: A Brief Review

Digital watermarking is based on the embedding of digital data (“watermark”) into video, image or audio signals (“host signals”) without degrading their perceptual quality. Since its beginning, this research topic has gained popularity for its wide applicability to many fields such as authentication, security, archiving, copyright and intellectual property rights (IPR) protection, steganography, broadcast monitoring, copy prevention, traitor tracking, legacy system enhancement and, generally speaking, transmission/storage of additional hidden data in multimedia signals [3]. In addition, it can be used to resolve rightful ownership and it can provide evidence of copyright infringements after the copyright violation has occurred. Watermark information is a special kind of bit pattern that behaves as an intentional noise inserted into the digital signal to transport the information useful to the

above mentioned applications: the data transmitted are usually inseparably bound to the host signal.

Depending on the specific type of host signal (video, image, audio, voice) and due to their specific characteristics (bandwidth and spectral density, time continuity and correlation, level of spatial and temporal sparsity or redundancy, etc.), different schemes emerged as the most efficient or robust solutions. Clearly *Speech Watermarking* (SpW) must be based on the voice features to be robust and efficient [4][5][6]. Digital watermarking is not only targeted for secure multimedia storage, but can also have an *active role in communications systems*, particularly for voice and video real-time transmissions for bandwidth multiplication, data hiding, authentication and security, added value services. As an example, SpW techniques may provide hidden data channels for parties authentication, security or signaling purposes, or simply for additional conversation-related data services without additional cost for bandwidth, depending on achievable bit rate. This mode of operation implies *blind watermarking*, because the decoder necessarily does not know the original host signal.

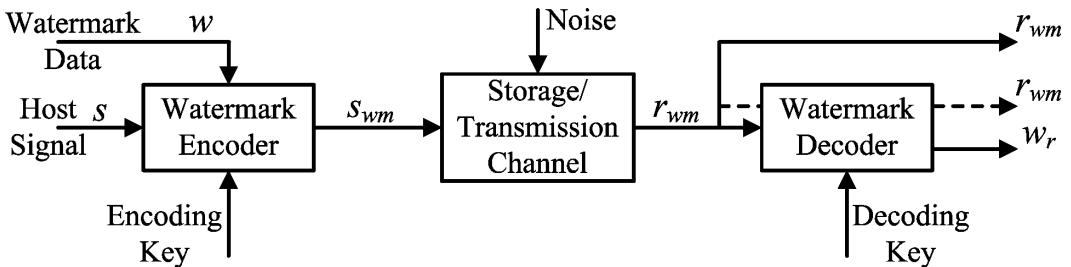


Figure 1. Generic architecture for a digital watermarking system [3].

Figure 1 shows the generic structure for a digital watermarking system, where the key elements are the watermark *encoder* (or embedder), responsible for the integration between *host signal* s and *watermark signal* w in the *watermarked signal* s_{wm} , and the *decoder*, which extracts (or retrieves) watermark from the watermarked signal. A private key may be needed in some cases to allow secure encoding and decoding operations. A storage/transmission channel may be interposed between encoder and decoder; this channel may add noise and distortions to the watermarked signal, making decoding more difficult if this operation is not appropriately modified. The dotted line at the decoder indicates that in some cases, depending on channel and chosen embedding technique, it will be possible to cancel watermark signal from watermarked signal, reverting partially or totally the effects of watermarking to recover the original host signal.

A good watermarking algorithm must satisfy at least two constraints: *imperceptibility* and *robustness*. Imperceptibility refers to the ability of a watermarking technique to embed the watermark with maximum energy (hence with better detectability) together with minimal perceptual effects on the host signal; clearly this definition often implies a subjective evaluation of the host signal quality. Robustness refers to the ability to withstand attempts of removal or alteration of inserted watermarks; in particular the watermarking algorithm is said *robust* if it is able to withstand channel impairments, noise and normal signal process-

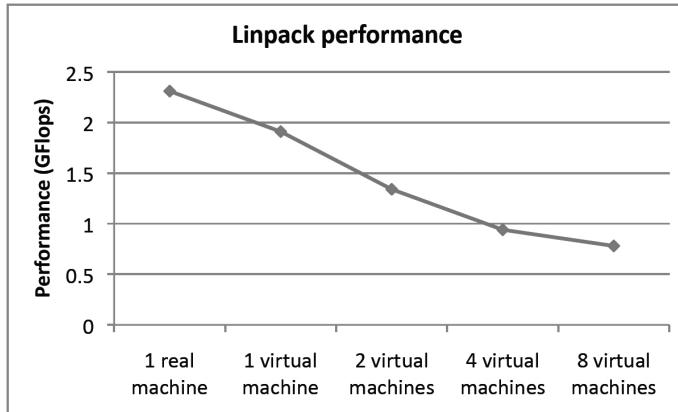


Figure 7. VM Overhead [10].

8. Future Work and Conclusion

From the above discussion, we can see that there are still several steps in the deployment process that are not yet automated, such as the configuration of Drupal. Such scenarios are a result of the application itself, and possible cannot be handled by middleware.

There are several other steps that can be dealt with, such as the configuration of the virtual appliance to operate with the host machine's network. In the future, deployment projects should be extended to include a more general deployment description language. Several studies have investigated this process [12, 13]. These tools languages are typically XML based, but this is not required. [2] discusses a deployment description language, but only in the context of their framework. In the future, Cyberaide Creative plans to make use of a deployment description language.

In summary, we have shown that using the Cyberaide Creative tool, we can simplify a service's deployment process, from the user's perspective. The user's service or application and a virtual machine image are combined into a virtual appliance. The virtual appliance is then deployed using Cyberaide Creative. Cyberaide Creative works to shield the user from the details of both the virtual appliance creation and deployment.

References

- [1] Alan Dearle. Software deployment, past, present and future. In *FOSE '07: 2007 Future of Software Engineering*, pages 269–284, Washington, DC, USA, 2007. IEEE Computer Society.
- [2] Changhua Sun, Le He, Qingbo Wang, and Ruth Willenborg. Simplifying service deployment with virtual appliances. In *SCC '08: Proceedings of the 2008 IEEE International Conference on Services Computing*, pages 265–272, Washington, DC, USA, 2008. IEEE Computer Society.
- [3] Julie Smith David, David Schuff, and Robert St. Louis. Managing your total it cost of ownership. *Commun. ACM*, 45(1):101–106, 2002.

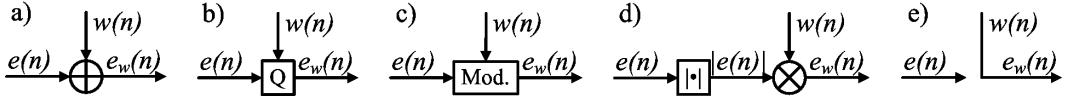


Figure 4. Examples of embedding techniques for voice residual: a) addition, b) quantization, c) modulation, d) sign multiplication, e) substitution or replacement [1].

features of this class in the case of a voice signal will be explained in subsection 3.1.. However, theoretical considerations highlighted an intrinsic limit on the achievable trade-off between data rate, speech quality and robustness due to the interference between host signal and watermark. As a consequence, the watermark has to be kept, for perceptual reasons, usually 16 – 20 dB below the speech signal, making the watermark detection a difficult task. As showed in the related state of the art [8][4][5], this class of speech watermarking systems can achieve bit rates in the range 30 – 200 bits/s.

Figure 3 shows that the embedding can be also performed in a transformed domain represented by the (usually adaptive) transform function $T(z)$. The watermark signal $w(n)$ is then applied on the transformed signal $e(n)$ by mean of an appropriate embedding technique and then the watermarked transformed signal $e_w(n)$ is retransformed to the time domain to obtain the watermarked signal. This mode of operation is not very common for voice, but is popular for other watermarking applications, mainly image processing, where a transformation through DCT (Discrete Cosine Transform) or DWT (Discrete Wavelet Transform) is used to represent the image in a more convenient domain for coding or watermarking [3][7]. In general this operation is performed for watermarking to obtain a better embedding efficiency in terms of perceived host signal quality, data rate and data robustness.

Figure 4 shows various examples of embedding techniques for the transformed signal proposed in the literature for various types of signals. The original transformed signal may be simply added to watermark (Figure 4a) [7], quantized depending on data (Figure 4b), modulated by watermark in terms of modulus or sign (Figures 4c and 4d), or completely replaced by watermark (Figure 4e). A notable example of these techniques is QIM (*Quantization Index Modulation*) [10] which, in principle, relied on requantising the signal with different codebooks depending on the watermark message, then the embedding of data in the least significant bits (LSBs) followed this principle, with proposed enhancements with bit rates in the range from 3 to 300 bits/s, but generally speaking these techniques have low robustness to noise.

3. Speech Watermarking

Figure 5 shows the high-level structure for most SpW systems. Watermark data are usually encoded for error protection (needed for channel/voice interferences), they are modulated to form the watermark signal, and then embedded in voice signal with addition or some other embedding strategy. At the receiver, the watermarked signal may be sent directly to the user (e.g. for reasons of legacy equipment enhancement) or processed together with watermark (for example to cancel data from speech); the time-frequency data embedding is inverted with an equalization operation followed by necessary operations on watermark

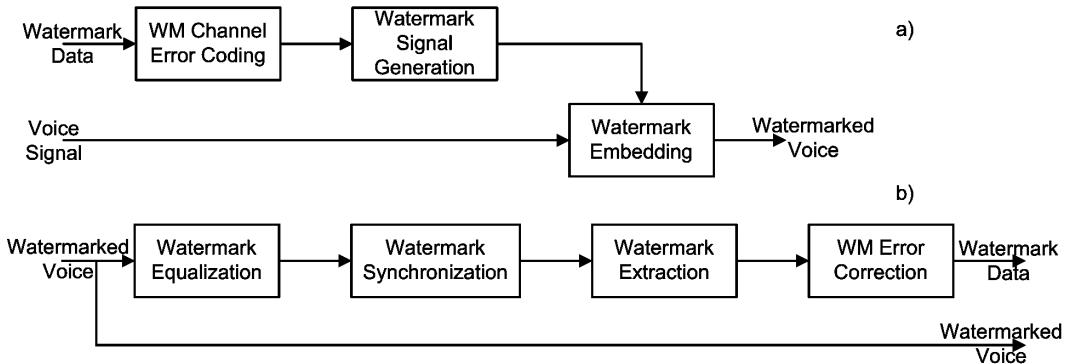


Figure 5. Functional structure of generic Speech Watermarking encoder (a) and decoder (b) [11].

signal such as synchronization, demodulation and decoding.

Concerning embedding strategies, many algorithms have been proposed in the recent decades [1]. Algorithms based on the QIM of line spectrum pair (LSP) parameters (Hatada et al., [12]), pitch period (Celik et al., [13]) and LPC residual (Geiser et al., [14]), with bit rates in the range from 3 to 300 bit/s, have been proposed. Methods which generalize QIM modulating the signal or one of its parameters have been also proposed, achieving bit rates in the range of 400 bits/s. The system presented in [15] by Sakaguchi et al. estimates and inverts the polarity of speech segments and embeds one watermark-bit per syllable. By contrast, Girin et al. in [16] modulates the frequency of selected partials in a sinusoidal speech model and achieves bit rates in the range of 400 bit/s.

From this brief review of speech watermarking it can be concluded that about 400 bits/s is the theoretical limit on bit rate that we can expect from the reported methods. However, in our experience [5], data reliability considerations make difficult to overcome the threshold of 100 bits/s in practical applications where impairments are present due, for example, to channel fading or bandwidth constraints; hence this rate value will be considered, throughout this chapter, as *the practical maximum data rate for actual SpW systems based on traditional approaches*. This is mostly due, generally speaking, to the fact that great care is taken to maintain the original shape or properties of the speech signal in order to minimize perceptual distortion; this is a limit due to the time-domain embedding approach. High Rate SpW uses a different but more efficient approach, explained in section 4..

3.1. Linear Predictive Coding-based Model

The most popular traditional method for speech watermarking is the *LPC-based model*, which uses the *improved spread-spectrum* principle [9], that is a time-frequency shaping, adaptive with respect to voice, of the spreaded watermark signal by mean of Linear Predictive Coding analysis-synthesis (which will be explained in subsection 3.2.); the embedding is in time domain. Two design parameters are important for data-voice quality trade-off: SWR (Speech to Watermark Ratio) that is the intended power ratio between voice and watermark in every transmitted frame, and WF (Watermark Floor) that is the minimum level

of watermark power in speech signal to guarantee data transmission in silence periods. The detailed scheme is described in many references [8][4][5]; in this chapter we consider useful to remember only the following synthetic main features [5]:

- Data rate: 100 bits/s thanks to QPSK modulation (100 bit is also the data frame size);
- Good quality of watermarked voice (MOS=3.6 – 4.4);
- Negligible encoding latency, decoding latency limited to about 100 ms;
- Design dependent on speech bandwidth.

Such a system enables, both for analogue and digital channels, only authentication service, which needs, in most cases, only a few tens of bits/s to encompass a sufficient number of digital signatures, both for public and private networks. Nonetheless, it is a robust and reliable watermarking system.

3.2. Linear Predictive Coding Analysis and Synthesis

An efficient and well known method for representing the characteristics and modeling the spectrum of human voice is *linear prediction analysis/synthesis*, which finds numerous applications in voice processing and also in other fields of signal processing [17]. It is often used in watermarking to obtain a limited frequency masking [8]. The method is based on the assumption that the signal's sample at time n can be written as a linear combination of the p immediately previous samples (that is “predicted” basing on them); this case is referred to as linear prediction of order p . The method is hence of autoregressive (AR) type of order p ; in the transformed z domain the corresponding transfer function has only poles and all zeros in the origin of the complex plane.

On the other hand, voice is the result of the modulation yielded from the human vocal tract (throat, mouth, teeth, lips) of the sound emission which comes from the vocal chords, and that is called *excitation*. So, the human vocal tract may be represented as a time-varying filter which acts on the phonemes, characterized by a transfer function $H(z)$ with only poles, in correspondence of which the frequency response presents peaks (*formants*) where, in the nearest neighboring spectral regions, the greater part of the signal energy is accumulated. The linear prediction is hence very suitable to model, both in time and frequency, the vocal signals.

With the AR model, the p parameters a_k of the linear prediction (called *LP coefficients*) may be efficiently determined solving a linear system of equations [18]. This may be shown starting from the linear prediction model of voice signal $s(n)$:

$$s(n) = \sum_{k=1}^p a_k s(n-k) + e(n) \quad (1)$$

where $e(n)$ represents the *prediction error*. Generally speaking, for analysis purposes, it is not normally available, so the equation reduces to:

$$\hat{s}(n) = \sum_{k=1}^p a_k s(n-k) \quad (2)$$

where $\hat{s}(n)$ is the *predicted sample* at time step n . Consequently the prediction error is:

$$e(n) = s(n) - \hat{s}(n) = s(n) - \sum_{k=1}^p a_k s(n-k) \quad (3)$$

In the (normalized) frequency domain, the transfer function between $e(n)$ and $s(n)$ corresponding to the (1) is:

$$H_{IIR}(F) = \frac{1}{1 - \sum_{k=1}^p a_k e^{-j2\pi F}} = \frac{1}{H_{FIR}(F)} \quad (4)$$

which is useful for *speech synthesis*, whereas $H_{FIR}(F)$ (more useful for *speech analysis*) is given by:

$$H_{FIR}(F) = 1 - \sum_{k=1}^p a_k e^{-j2\pi F} \quad (5)$$

Since the voice signal is a *small scale stationary signal* (that is stationary on small intervals of a maximum duration of 10 – 20 ms) the LPC analysis is often performed on frames of the duration of 20 ms but overlapped of 50% with the purpose to obtain a more refined analysis. The signal is successively windowed using a window function, e.g. a Hamming window²:

$$s_m(n) = s(mN + n) w(n) \quad (6)$$

where

$$w(n) = \begin{cases} 0.54 - 0.46 \cos(2\pi n/N) & \text{for } 0 \leq n \leq N \\ 0 & \text{otherwise} \end{cases}$$

and where the subscript m represents the frames index. The prediction coefficients are determined minimizing the mean square error (MSE) between predicted and real samples for a frame of the voice signal [18]:

$$E_m = \sum_{n=0}^{N-1} e_m^2(n) = \sum_{n=0}^{N-1} (s_m(n) - \hat{s}_m(n))^2 = \sum_{n=0}^{N-1} \left(s_m(n) - \sum_{k=1}^p a_k s_m(n-k) \right)^2 \quad (7)$$

The error E_m is minimized imposing:

$$\frac{\partial E_m}{\partial a_i} = 0, \quad 1 \leq i \leq p \quad (8)$$

From the (8) and (7), it may be obtained:

$$\sum_{k=1}^p a_k \sum_{n=0}^{N-1} s_m(n-k) s_m(n-i) = - \sum_{n=0}^{N-1} s_m(n) s_m(n-i) \quad (9)$$

A set of p equations, known as *Yule-Walker equations*, is obtained, which must be solved as a function of the parameters a_k [17]:

$$\begin{bmatrix} r(0) & r(1) & \dots & r(p-1) \\ r(1) & r(0) & \dots & r(p-2) \\ \vdots & \vdots & \ddots & \vdots \\ r(p-1) & r(p-2) & \dots & r(0) \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} r(1) \\ r(2) \\ \vdots \\ r(p) \end{bmatrix} \quad (10)$$

²These hypotheses are also assumed in the present work.

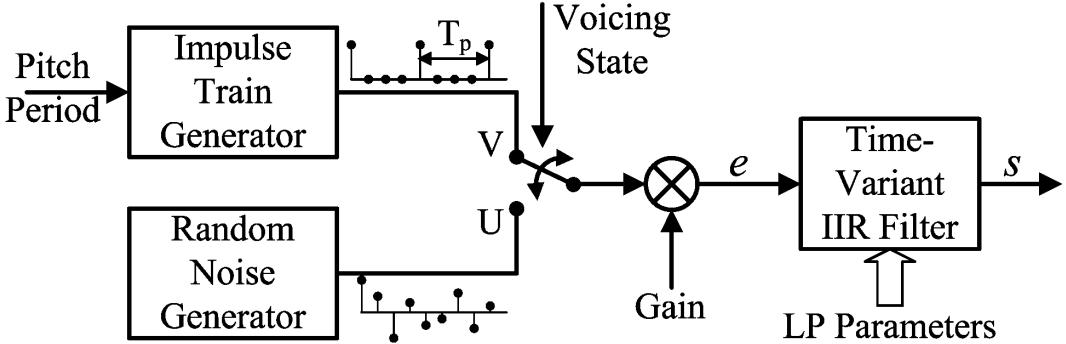


Figure 6. Speech synthesis model based on LPC model [20].

This system may be rewritten in a compact notation as:

$$R a = r \quad (11)$$

where the autocorrelation sequence $r(k)$ is defined as:

$$r(k) = \sum_{n=0}^{p-1+k} s_m(n)s_m(n+k) \quad (12)$$

The solution for a_k (assuming that correlation matrix R is invertible) may be written as:

$$a = R^{-1}r \quad (13)$$

In its matrix representation, the system is individuated by a Toeplitz matrix [19]: in this case, in order to avoid the direct inversion of the matrix, the *Levinson-Durbin algorithm* may be used, which has a computational complexity relatively low (of order p^2), and which provides a very efficient method to compute the LP coefficients. For a detailed description of the algorithm, see e.g. [19] and [17]. The optimal parameters determined imposing (8) make also the prediction error $e(n)$ orthogonal to the p previous predicted samples (*MMSE orthogonality principle*):

$$\sum_{n=0}^{N-1} e_m(n)s_m(n-i) = 0, \quad 1 \leq i \leq p \quad (14)$$

Usually for voice the order of prediction p is fixed to a value ranging from 8 to 12; in this work p has been fixed to 7 to keep low the computational complexity. This value will be further justified in subsection 9.6..

4. High Rate Speech Watermarking

The approach proposed in [1] by Hofbauer and Kubin is based on the transform domain concept as shown in Figure 3. The idea is to exploit the well known mechanism of

production of voice in the human vocal tract, as depicted in Figure 6, often called *source filter model of speech production* [21]. The voice activity state can be roughly divided in two classes, *voiced* and *unvoiced*³ [17]. In the voiced segments (usually responsible of 70-75% of the total voice time) the excitation e impressed by the vocal chords on the air coming from the lungs is approximately a periodic impulse train with a so called *pitch period* T_p . In the unvoiced segments, the excitation is well modeled with a random white noise signal with a suitable power. This scheme is the base for most well known low rate LPC vocoders [22]. Switching between these two states on the base of the info estimated by a suitable pitch detection algorithm, it is possible to scale excitation e with a time-variant gain factor and then synthesize voice signal s thanks to an LPC synthesis filter (a parametric time-varying IIR all-poles filter) which models the response of vocal tract in the current segment. The source filter model provides a very accurate representation for voice signals, particularly for unvoiced segments, where extended listening tests [23] have shown that speech resynthesized using pure white noise as excitation in the unvoiced segments led to MOS⁴ values almost identical to those of original speech. The conclusion is that *the speech quality does not degrade when the LPC residual in unvoiced segments is replaced by white noise of equal power*.

The same principle used for speech coding may be easily exploited for watermarking if we keep in mind that watermark data are usually pseudonoise random data. LPC analysis (which has been briefly detailed in subsection 3.2.) has been chosen in High Rate SpW as transform function $T(z)$: given a speech signal, the coefficients of the vocal tract filter are obtained by a linear prediction based on previous samples. An adaptive FIR filtering of the speech signal results in the so-called prediction *residual* (or voice *excitation*) e , which acts as the watermarking domain. Unfortunately, although the excitation in the voiced segments is correctly modeled by source filter model, its structure is not suitable to sustain watermarking without incurring in a sensible speech quality degradation. However, unvoiced segments have an occurring ratio of 25-30%, so a rate sensibly higher than traditional SpW systems (like the LPC-based system shown in subsection 3.1.) is achievable with negligible voice quality degradation.

The basic structure of the resulting watermarking system is shown in Figures 7 (encoder) and 8 (decoder), where blue-shaded blocks will be introduced as part of the new proposed solutions, and concurrently dotted lines and blocks from original scheme in [1] will be dropped since not needed anymore. The following discussion will be initially related to the basic system, while the new blocks will be detailed in Sections 5., 6. and 7..

4.1. Watermark Encoding

Let us assume as reference a voice signal sampled at 8 kHz and PCM (Pulse Code Modulation) encoded with a resolution of 16 bits. Watermark encoding is performed through the

³This is a simplification, because in normal speech there are also cases of hybrid voiced-unvoiced or voiceless segments.

⁴Mean Opinion Score (MOS) is an evaluation scale commonly adopted for the subjective measurement of voice quality [24][25]; the scale goes from 1 (bad quality) to 5 (best quality), and is evaluated through extensive campaigns involving a sufficiently high number of listeners, each one giving his or her opinion on the audio quality.

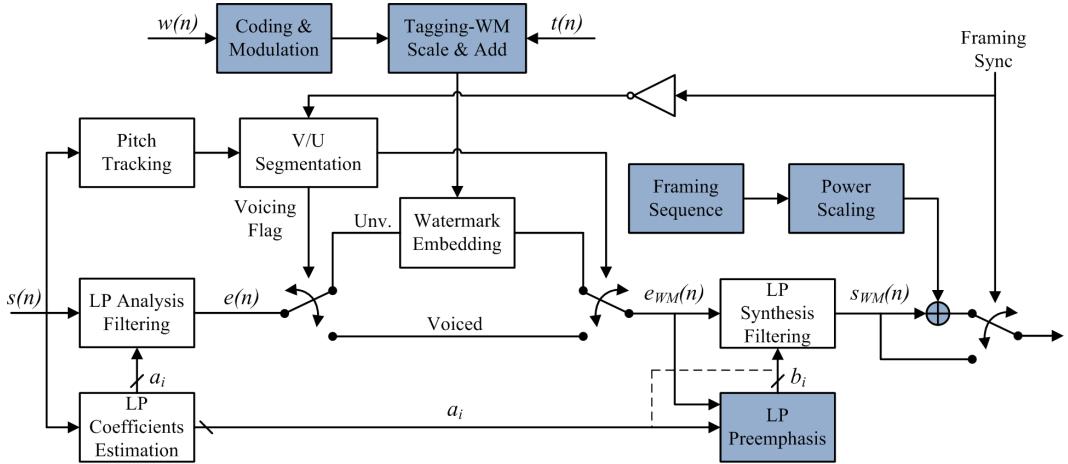


Figure 7. Block diagram of the High Rate Speech Watermarking encoder with proposed solutions for channel coding, synchronization and equalization.

LPC approach above discussed. The same LP filter coefficients are used for the analysis and for the inverse synthesis filter; they are estimated from the speech signal and periodically updated on a frame (or packet) basis, whose length is 30 bits: a 7th-order LP analysis is updated every frame using the Levinson-Durbin algorithm [26] on a window length of 160 samples. The choice of a such low order will be justified in the subsection 9.6.. After the watermark embedding has modified the residual, the speech signal is resynthesised using the LPC parameters set.

The V/U segmentation⁵ is performed thanks to a pitch detection algorithm. For reasons of homogeneity, we resorted to the same autocorrelation-based pitch tracking algorithm used in [1] and implemented in PRAAT program [27]. This algorithm has five important parameters that may affect the segmentation: *Number of Candidates*, *Silence Threshold*, *Minimum Pitch*, *Ceiling* (Maximum Pitch) and *Voicing Threshold*. As we explained in [2], the first two parameters do not affect significantly the segmentation, while the others are relevant for the system optimization, since a better voice segmentation can lead to a higher quality for watermarked voice and a better robustness of segmentation to channel noise. After careful optimization [2], we found the optimal parameter values, with *Minimum Pitch*=250 Hz, *Ceiling*=950 Hz and a *Voicing Threshold* of 0.35.

We assume that watermark binary sequence undergoes a preliminary bipolar encoding, essentially a BPSK modulation, with values ± 1 . In [1], three embedding techniques were investigated, multiplication, replacement and replacement + spreading. Differently from that, in this chapter we resorted only to the *replacement* method due to superior performance with respect to multiplication and the excessive throughput loss caused by spreading, which may be replaced by channel coding. With replacement the residual is replaced by the binary watermark signal (Figure 4e). For each frame, the power of the watermark signal is matched with the power of the original residual; mean power matching has given better results than instantaneous power matching.

⁵In the following sections of this chapter voiced and unvoiced will be often abbreviated as V and U.

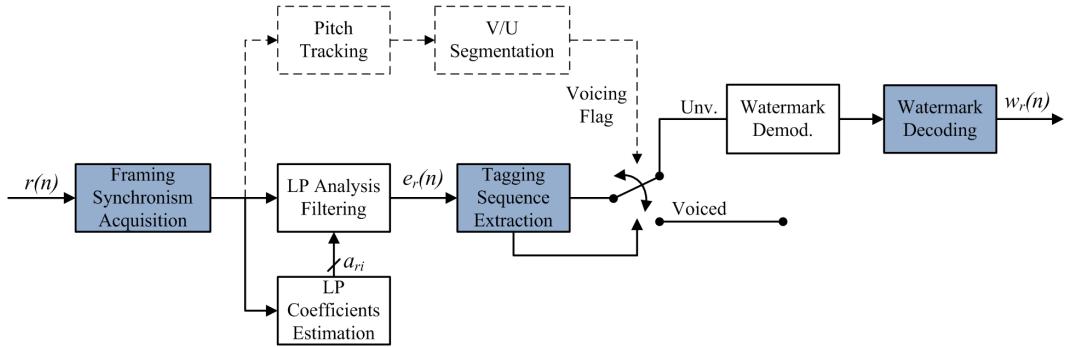


Figure 8. Block diagram of the High Rate Speech Watermarking decoder with proposed solutions for channel coding, synchronization and equalization.

4.2. Watermark Decoding

The decoder must be able to extract, from the received signal, the watermark code. Watermarked voice signal is sent directly to the output; this is possible since watermarking, thanks to the V/U-based integration process, has not significantly altered the voice quality. We assume here, as a coexistence constraint, that the receiver cannot perform a cancellation of data from received voice, since we assume that voice must be received without problems also by possible legacy receivers (from which the *imperceptibility constraint* arises).

The watermark decoder obeys the same structure and processing steps as the encoder (Figure 8). After the LPC analysis, the recovered residual undergoes a BPSK demodulation; the watermark bits are recovered from the estimation of the sign of residual.

The almost perfect symmetry between encoder and decoder is an advantage as it is possible to realize an hardware or software modular implementation, with only a slight customization, decreasing design and deployment costs.

4.3. Scheme Advantages, Flaws and Open Issues

The High Rate watermarking method has the following synthetic features, already studied in deep in [1] and [2]:

- Data rate: 2000–2400 bits/s, possibly expandable with use of higher modulation orders and/or more aggressive profiles of V/U segmentation;
- Very good quality of watermarked voice ($MOS \approx 4.4$);
- Low sensitivity to noise channel;
- Negligible encoding latency, decoding latency limited to about 20 ms;
- Substantial design independence from speech bandwidth.

In [2] it has been also shown its efficiency and feasibility for the case study of aircrafts authentication in ATC communications. Such a system has greater potential performance than traditional SpW systems, like the LPC-based system shown in subsection 3.1.: it can

carry not only authentication data, but has also the capability to enable new data services related to the conversation, like e.g. instant messaging, or even carry IP network hidden signaling data for a greater security level (the low decoding latency is a very interesting feature for VoIP applications). However, great care must be taken to make this performance feasible in a real scenario.

Theoretical and practical considerations, in conjunction with computer simulations [2] have highlighted flaws and open issues which must be solved. We propose three major enhancements in terms of channel coding, synchronization and residual equalization, approximately ordered in terms of growing complexity. These algorithms are developed keeping an eye both on performance and implementation complexity, and are detailed in the next three Sections.

1. *Channel Coding.* The hidden data channel requires an adequate protection against channel noise and other distortions to give good performance in terms of BER (Bit Error Rate) and PER (Packet Error Rate) for desired applications. Uncoded transmission can not sufficiently satisfy performance target levels, and data spreading is inefficient in terms of bit rate. A suitable Forward Error Correction (FEC) method must be chosen carefully keeping count of the particular features of this system. The proposed solution is showed in Sec. 5..
2. *Decoder Synchronization.* Watermark synchronization is never a simple task, particularly in speech systems, due to the need of deep hiding of watermark in host signal for perceptual reasons. This makes harder its reliable extraction, and this is even more true for this system, where data transmission is finely packetized and transmitted without continuity. The problem is divided in *frame synchronism* (the detection of packet limits at the decoder) and the *detection of packet nature* (Voiced/Unvoiced). This problem is hence strictly tied to segmentation; the solution proposed (Sec. 6.) must have superior performance and lower complexity with respect to segmentation repeated at the decoder.
3. *Residual Equalization.* The insertion of watermark data yields the difficulty to correctly recover LPC coefficients at the decoder. In Sec. 7. it will be showed that watermark insertion makes LPC coefficients recovered by the decoder different from that computed by encoder, and this phenomenon is not dependent by channel noise level; this causes a distortion, on recovered watermarked residual, similar to that related to a band-limited channel. So, the decoded residual must be equalized in order to make system performance as closer as possible to the ideal case, without the need of coefficients transmission (impossible for lack of available bandwidth). The solution proposed is based on *encoding preemphasis*.

5. Channel Coding

Original system proposal does not comprehend channel coding, causing unsatisfying performance in terms of BER and PER [2]. On the other side, data spreading is inefficient in terms of bit rate. The choice of an appropriate channel coding solution is not easy, due to the impulsive nature of data packets transmission; moreover, packets are small (only 30

bits). Due to this, advanced channel coding techniques (like turbo codes, LDPC, etc.) are not very effective (e.g. turbo codes have too small interleaving gain) [28]. On the other side, grouping packets for coding is not recommended, due to consequent high decoding latency ([2], system feasibility analysis).

So, we resorted to simple convolutional coding [26] operated on single packets; after careful optimization, both theoretical and through computer simulations, we chose the maximum free distance code with $r = 1/2$, $K = 3$ and $g = [5\ 7]_8$ [29]. A such low memory has been demonstrated useful in maximizing performance with a such low packet size, since low memory means small packets of error when the decoding fails. However, to fully exploit channel coding gain, trellis termination would be needed by mean of 2 tail bits insertion, and this entails a throughput loss of (at least) 13%, considered unacceptable. We eventually chose *tail-biting* coding to overcome this problem [30]: the idea is to force, in both encoding and decoding phases, an equivalence between initial and final trellis state instead to forcing them to zero state. In this way we are able to avoid tail bits and throughput loss, accepting an increase in decoding complexity. Optimal decoding can be carried out with a modified Viterbi algorithm [26] iterating on each possible initial and final state: it gives optimal performance but has maximum complexity. Suboptimal decoding may be an alternative solution, usually based on the repetition of received codeword to simulate trellis circularity, as depicted in Figure 9; in this second case we resorted to the algorithm described in [30]. Both solutions will be tested for performance in subsection 9.2..

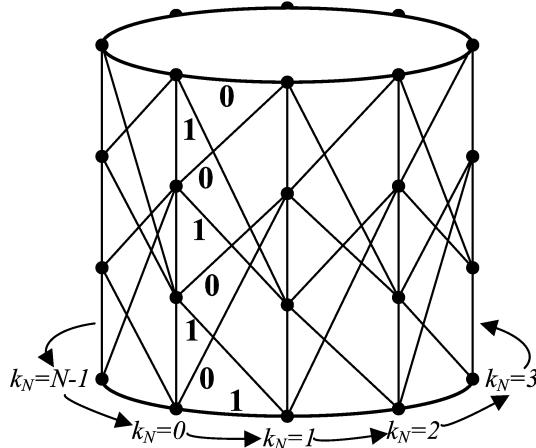


Figure 9. The principle of circular trellis for ML decoding of tail biting codes: an example of convolutional code with 4 states and decoding depth $N = 8$, where $k_N = k \bmod N$.

6. Decoder Synchronization

Watermark synchronization is always a difficult task, particularly for this system due to deep hiding of watermark in host signal for perceptual reasons and not continuous packet transmission. This makes harder its reliable extraction. We divided this operation in *Frame*

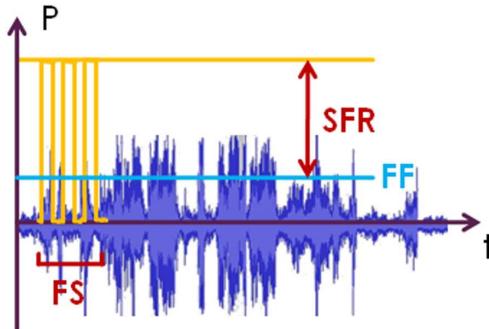


Figure 10. Frame synchronization parameters with respect to speech signal.

Synchronism (the detection of packet limits at the decoder) and the detection of packet nature (V/U) through *Packet Tagging*. This problem is hence strictly tied to V/U segmentation. In [2] it has been shown that although segmentation operated at the receiver is robust to channel noise, it gives unsatisfying performance in terms of correctness of voicing state recovery for each received packet. Our aim is to design a more effective and simple solution than segmentation operated at the decoder.

6.1. Frame Synchronization

Frame synchronization refers to the correct detection of packet limits at the decoder, needed due to channel and/or resampling delay. This is done through the addition, on the very initial speech transmission samples, of a *framing sequence* which the decoder will search for at the start of its operation on the received signal⁶. The sequence is made of pseudo-noise random data, and is defined by the parameters FS (its length), SFR (Speech to Framing sync Ratio), sequence's power w.r.t. speech signal, and FF (Framing sync. Floor), the minimum sequence power level to deal with silence periods (probable at start of transmission). The overall situation is depicted in Figure 10, where the proportions do not respect real values chosen for parameters. During FS transmission, watermark integration is obviously deactivated, with a negligible impact on system throughput. The decoder needs only a simple cross-correlator to achieve frame synchronism. The solution is simple and suitable not only for our example scenario, but also for all applications. Performance results are given in subsection 9.3..

6.2. Packet Tagging

Packet tagging is the solution proposed to send to decoder information on the nature (V/U) of each data packet instead of recovering it at the decoder itself through segmentation. A *tagging sequence* TS (pseudo-noise random sequence) $t(n)$ is added to the watermark $w(n)$ in unvoiced tracts, after suitable scaling for both of them (Figure 11) in order to maintain total residual power. The decoder check if TS is present through a cross-

⁶This implies that watermark encoder is able to synchronize voice and data; this is an hypotheses valid in most applications, but its fairness is very application specific.

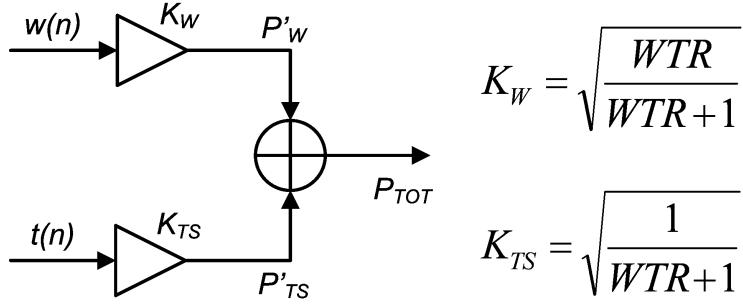


Figure 11. Functional diagram for block Tagging-WM Scaling & Add.

correlation followed by threshold comparison and, if present, it cancels it from residual to recover watermark data. Scaling factors depend on WTR (Watermark to Tagging sequence Ratio) parameter, which must be carefully chosen, along with threshold S, for the right performance trade-off between watermark and tagging. Performance results (subsection 9.3.) will be evaluated in terms of probabilities of right detection P_{rs} (Right Sync. Probability), P_{md} (Missed Detection Probability), P_{fa} (False Alarm Probability).

7. Residual Equalization

Here we show that *watermark insertion makes LPC coefficients recovered by the decoder different from that computed by encoder*, and this phenomenon is not (only) dependent by channel noise level; this causes, on recovered watermarked residual, a distortion similar to that related to a *band-limited channel*. So, the *decoded residual must be equalized* in order to make system performance as closer as possible to the ideal case, without the need of coefficients transmission (impossible for lack of available bandwidth). The solution proposed is based on *encoding preemphasis*. Referring to Figures 7 and 8, we can extend the LPC analysis made in subsection 3.2. to the entire watermarking system considering the presence of additive noise channel $n(n)$ and not ideal received LP parameters, to write the relation between decoded and transmitted residual, namely the *hidden data channel input-output relation*:

$$E_r(z) = \frac{1 - \sum_{i=1}^p a_{ri}z^{-i}}{1 - \sum_{i=1}^p a_i z^{-i}} \cdot E(z) + \left[1 - \sum_{i=1}^p a_{ri}z^{-i} \right] \cdot N(z) \quad (15)$$

which highlights that there are two factors of degradation, *colored channel noise* and *hidden channel transfer function* $H(z) = \frac{1 - \sum_{i=1}^p a_{ri}z^{-i}}{1 - \sum_{i=1}^p a_i z^{-i}} \neq 1$, both due to LP parameters mis-identification ($a_{ri} \neq a_i$) and channel noise.

The first effect is due only to channel noise. If we define the channel SNR (Signal to Noise Ratio) as

$$SNR = \frac{P_{s_{wm}}}{P_n}$$

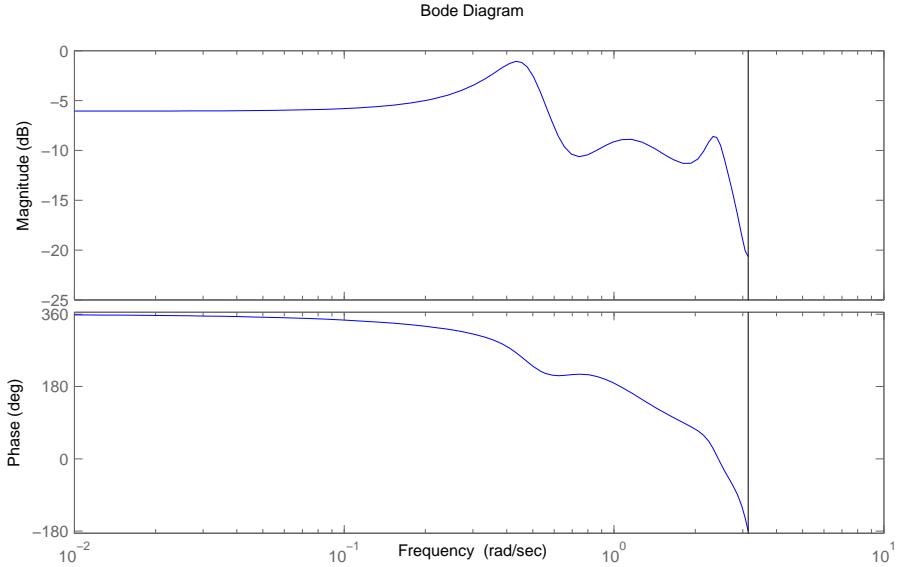


Figure 12. Example of distortion function $H(F)$.

where $P_{s_{wm}}$ is the power of the marked transmitted signal and P_n is the power of the channel noise, it is possible to write:

$$P_r = P_{s_{wm}} + P_n = P_{s_{wm}} + \frac{P_{s_{wm}}}{SNR} = P_{s_{wm}} \cdot \left(\frac{SNR + 1}{SNR} \right) \quad (16)$$

where P_r is the power of the received signal. If the SNR experienced by the signal is high (like, e.g., that of an ATC channel, about 20–50 dB) the term $\frac{SNR+1}{SNR}$ is approximated to unity, and, although colored, the noise has a negligible effect. Moreover, if the noise has the same bandwidth of the signal, the receive analysis filter has no effect, because the SNR is the same before and after the filter. However the colored nature of the noise is to be considered to make a conservative design of the channel code in order to guarantee acceptable performance.

The watermark insertion is responsible of the difference among received parameters a_{ri} and original parameters a_i , from which the distortion function $H(z)$ raises. Figure 12 shows an example of amplitude and phase distortion; this function has an irregular low-pass filter characteristics (due to errors occurring mostly on highest order parameters), representative of the most part of the real cases. This phenomenon is dominant on colored noise, and is originated from *loss of orthogonality* between residual (modified by watermark) and voice prediction, as depicted in Figure 13. As explained by the linear prediction theory (subsection 3.2.), $s(n)$, a_i and $e(n)$ form a triad related in terms of the orthogonality principle of the linear MMSE theory [17]. After the alteration of residual $e(n)$ in $e_{WM}(n)$ and reusing the a_i to resynthesize $s_{WM}(n)$, the new triad $\{s_{WM}(n), a_{ri}, e_r(n)\}$ is not anymore related according to that principle, at least in the unvoiced periods. The decoder, which is based on the LPC analysis for the recovery of the marked residual $e_r(n)$, will be forced to determine new parameters a_{ri} such that $\{s_{WM}(n), a_{ri}, e_r(n)\}$ is a triad related according

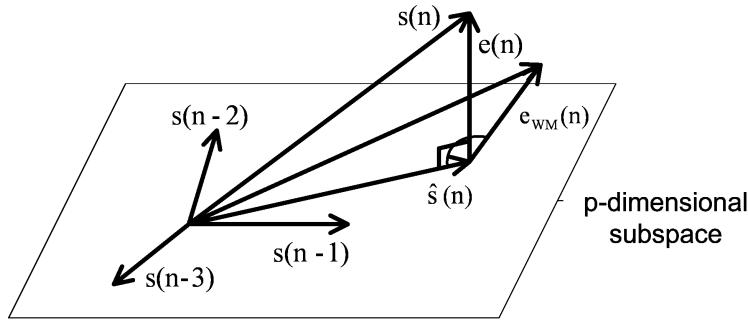


Figure 13. Linear Prediction orthogonality principle and residual modification.

to the orthogonality principle, hence causing a distortion of $e_r(n)$ with respect to $e_{WM}(n)$, quantitatively represented by the transfer function $H(z)$ (even in absence of channel noise). In order to reduce or eliminate such distorting effect, there are two ways:

- changing the elaboration performed by the receiver in something different from the simple LPC analysis used in transmission, e.g. a Least Squares estimator;
- modify the parameters which we use to resynthesize the voice at the exit of the encoder to make the new triad $\{s_{WM}(n), b_i, e_{WM}(n)\}$ being orthogonal. In this way the receiver, operating always with a symmetrical structure based on LPC analysis, will be always able to recover $e_{WM}(n)$ without distortion (except that from noise channel). This intentional distortion of the LPC parameters (called transmission *preequalization* or *preemphasis*) is used to guarantee the absence of distortions at the receiver side of the data channel, although in spite of the voice (at least in the unvoiced periods): for this reason it will be important to choose the new parameters b_i in such a way that minimizes the perceptual impact on the voice signal.

We developed this second solution, believed more interesting in terms of containment of receiver complexity. Let us restart from the expression

$$e(n) = s(n) - \sum_{i=1}^p a_i s(n-i), \quad 0 \leq n \leq (N-1) \quad (17)$$

where N is the analyzed frame (packet) size. This expression, already shown in subsection 3.2. for the LPC analysis, provides the error with minimal energy in the case of usage of optimal a_i parameters. For perceptual reasons $e(n)$ and $e_{WM}(n)$ must have the same mean energy:

$$E_m = \sum_{n=0}^{N-1} e^2(n) = \sum_{n=0}^{N-1} e_{WM}^2(n), \quad 0 \leq n \leq (N-1). \quad (18)$$

In the original system it is possible to write the relation

$$e_{WM}(n) = s_{WM}(n) - \sum_{i=1}^p a_i s_{WM}(n-i), \quad 0 \leq n \leq (N-1) \quad (19)$$

that is a relation of synthesis because this time e_{WM} , a_i and p are the known inputs and $s_{WM}(n)$ is the result. However $\{s_{WM}(n), a_i, e_{WM}(n)\}$ is not anymore an orthogonal triad, therefore new parameters b_i are searched, such that $\{s_{WM}(n), b_i, e_{WM}(n)\}$ is an orthogonal triad (note that the synthesized signal $s_{WM}(n)$ will not be the same of before):

$$e_{WM}(n) = s_{WM}(n) - \sum_{i=1}^p b_i s_{WM}(n-i), \quad 0 \leq n \leq (N-1). \quad (20)$$

The (20) are N equations in $(N+p)$ unknowns, so the problem is underdetermined. This means that ∞^p orthogonal triads exist in the $(N+p)$ -dimensional signal space, all with $e_{WM}(n)$ as residual and hence all valid for the correct LPC analysis at the decoder. However, these are not equivalent in terms of generated voice, which must be as near as possible to the original one in perceptual terms. Therefore the choice criterion of the b_i must be the production of the lowest vocal distortion; since the parameters alteration is required only in the unvoiced periods where the voice phonation is not present (that is perceptually less relevant) and concerning only 25 – 30% of the voice itself, it is possible that a moderate distortion of the LPC parameters is sufficient to maintain a high DMOS (Differential Mean Opinion Score).

Therefore we have the problem of analytically expressing in an effective way the concept of low perceptual impact of the new parameters on the voice quality. Let us consider the following approach. The transfer function of the synthesis filter is (in origin):

$$H(z) = \frac{1}{1 - \sum_{k=1}^p a_k z^{-k}}$$

with frequency response

$$H(F) = \frac{1}{1 - \sum_{k=1}^p a_k e^{-j2\pi F k}}$$

Replacing the parameters a_k with the new parameters b_k , the transfer function becomes

$$H'(z) = \frac{1}{1 - \sum_{k=1}^p b_k z^{-k}}$$

with frequency response

$$H'(F) = \frac{1}{1 - \sum_{k=1}^p b_k e^{-j2\pi F k}}.$$

Considering a pseudonoise excitation input to the filter, its frequency response becomes fundamental to shape the spectral characteristics of the generated voice. Therefore we impose the minimization criterion of the perceptual impact as the *maximization of the “spectral likelihood”* between the old and the new filter, that is:

$$b_k, \quad 1 \leq k \leq p : \quad \min_{b_k} \|H'(F) - H(F)\|_2^2, \quad \forall F \in [0, 1]$$

The optimal parameters b_k must minimize the energy of the error on the spectrum of the resynthesized voice (called ΔH):

$$\begin{aligned}\Delta H &= \|H'(F) - H(F)\|_2^2 = \\ &= \int_0^1 \left| \frac{1}{1 - \sum_{k=1}^p b_k e^{-j2\pi F k}} - \frac{1}{1 - \sum_{k=1}^p a_k e^{-j2\pi F k}} \right|^2 dF\end{aligned}\quad (21)$$

Note that the obvious solution would be $b_k = a_k, \forall k = 1, \dots, p$, but this solution would not give an orthogonal triad, as already said, therefore these two different requisites will have to be integrated. It is clear that the considered problem configures itself like a *constrained nonlinear optimization problem*. First of all, the function to be minimized may be elaborated as follows:

$$\begin{aligned}\Delta H &= \int_0^1 \left| \frac{1 - \sum_{k=1}^p a_k e^{-j2\pi F k} - (1 - \sum_{k=1}^p b_k e^{-j2\pi F k})}{(1 - \sum_{k=1}^p b_k e^{-j2\pi F k}) \cdot (1 - \sum_{k=1}^p a_k e^{-j2\pi F k})} \right|^2 dF = \\ &= \int_0^1 \left| \frac{\sum_{k=1}^p (b_k - a_k) e^{-j2\pi F k}}{(1 - \sum_{k=1}^p b_k e^{-j2\pi F k}) \cdot (1 - \sum_{k=1}^p a_k e^{-j2\pi F k})} \right|^2 dF\end{aligned}\quad (22)$$

The numerator of the integrand function is the DTFT (Discrete Time Frequency Transform) of the difference sequence among new and old parameters:

$$\Delta D(F) = \sum_{k=1}^p (b_k - a_k) e^{-j2\pi F k}\quad (23)$$

On the other hand, if the expected spectral distortion is small, minimizing ΔH equals to minimize ΔD .

Lemma: Let it be

$$\begin{aligned}\delta_f &\triangleq \frac{1}{y} - \frac{1}{x} \\ \delta_D &\triangleq y - x\end{aligned}$$

with $x \neq 0, y \neq 0$.

$$\frac{1}{y} = \frac{1}{x} + \delta_f = \frac{1}{x + \delta_D}$$

from which:

$$\begin{cases} \delta_f &= \frac{1}{x + \delta_D} - \frac{1}{x} = -\frac{\delta_D}{x + \delta_D} \\ \delta_d &= \frac{x}{\delta_f + 1} - x = -\frac{x^2 \delta_f}{x \delta_f + 1} \end{cases}$$

Remembering that $x \neq 0$

$$\begin{aligned}\lim_{\delta_f \rightarrow 0} \delta_D &= 0 && \text{(necessary condition)} \\ \lim_{\delta_D \rightarrow 0} \delta_f &= 0 && \text{(sufficient condition)}\end{aligned}$$

which means that in an interval sufficiently small around x , it is equivalent to minimize δ_f or δ_D . \square

So, basing on this lemma, let us consider the functional derived taking only the numerator of (22):

$$\Delta D = \int_0^1 \left| \sum_{k=1}^p (b_k - a_k) e^{-j2\pi F k} \right|^2 dF \quad (24)$$

ΔD is the energy of the error sequence $(b - a)$, computed in the frequency domain. For the Parseval theorem [26] we can transport it in the time domain:

$$\begin{aligned} \Delta D &= \int_0^1 \left| \sum_{k=1}^p (b_k - a_k) e^{-j2\pi F k} \right|^2 dF = \\ &= \sum_{k=1}^p (b_k - a_k)^2 \stackrel{\Delta}{=} E_{\Delta LP} \end{aligned} \quad (25)$$

$E_{\Delta LP}$ (energy of the difference among old and new LP parameters) is the function to be minimized. The unknowns of the problem are the p parameters b_k and the N samples of the marked voice s_{WM} :

$$E_{\Delta LP}(b_1, \dots, b_p, s_{WM}(0), \dots, s_{WM}(N-1))$$

where there have been also inserted the N samples of resynthesized-marked voice as fictitious variables; the a_k must be instead treated like known constants. We solve the system through the *Lagrange multipliers' method*. Assumed:

$$\begin{array}{ll} f(x) = E_{\Delta LP}(x) & \text{function to be minimized} \\ g_i(x) = c_i & \text{constraint functions, } i = 1, \dots, N_v \end{array}$$

where $x = (b_1, \dots, b_p, s_{WM}(0), \dots, s_{WM}(N-1))$ is the unknowns vector, and N_v the number of constraint functions, we can write the *Lagrangian function*:

$$\Lambda(x) = f(x) + \sum_{i=1}^{N_v} \lambda_i (g_i(x) - c_i), \quad \lambda_i \neq 0, \quad i = 1, \dots, N_v \quad (26)$$

The constraint functions are the input-output relation of the IIR synthesis filter (which relates b_k , e_{WM} and s_{WM}) replied for all the $n = 0, \dots, N-1$, and the orthogonality relations of the marked excitation/residual e_{WM} with respect to the prediction $\hat{s}_{WM}(n)$.

$$s_{WM}(n) - \sum_{k=1}^p b_k s_{WM}(n-k) = e_{WM}(n), \quad 0 \leq n \leq (N-1) \quad (27)$$

are the first N constraint functions. Moreover, it is still valid the relation of the mean power

of the marked residual

$$\begin{aligned}
 E_{WM}(bk, s_{WM}(n)) &= \sum_{n=0}^{N-1} e_{WM}^2(n) = \sum_{n=0}^{N-1} (s_{WM}(n) - \hat{s}_{WM}(n))^2 = \\
 &= \sum_{n=0}^{N-1} (s_{WM}(n) - \sum_{k=1}^p b_k s_{WM}(n-k))^2 = \\
 &= E_m = \sum_{n=0}^{N-1} e^2(n)
 \end{aligned} \tag{28}$$

which, as already said, is a known value, function of the $p + N$ unknowns.

In order to obtain the orthogonality relations, E_{WM} is minimized with respect to the LP parameters b_k and to the samples $s_{WM}(n)$.

$$\begin{aligned}
 \frac{\partial E_{WM}}{\partial b_k} &= \sum_{n=0}^{N-1} 2(s_{WM}(n) - \sum_{k=1}^p b_k s_{WM}(n-k)) \cdot (-s_{WM}(n-k)) = \\
 &= \sum_{n=0}^{N-1} -2e_{WM}(n)s_{WM}(n-k) = 0, \quad 1 \leq k \leq p,
 \end{aligned} \tag{29}$$

from which the p constraint functions:

$$\sum_{n=0}^{N-1} e_{WM}(n)s_{WM}(n-k) = 0, \quad 1 \leq k \leq p \tag{30}$$

This form is more useful for the LPC synthesis than that used for the analysis (subsection 3.2.), where autocorrelations were featured, since it contains linear combinations and introduces the known terms $e_{WM}(n)$.

The Lagrange function therefore is:

$$\begin{aligned}
 \Lambda(y) &= \sum_{k=1}^p (b_k - a_k)^2 + \\
 &+ \sum_{n=0}^{N-1} \alpha_n \left(s_{WM}(n) - \sum_{k=1}^p b_k s_{WM}(n-k) - e_{WM}(n) \right) + \\
 &+ \sum_{k=1}^p \beta_k \sum_{n=0}^{N-1} e_{WM}(n)s_{WM}(n-k) + \\
 &+ \sum_{n=0}^{N-1} \gamma_n \left(e_{WM}(n) - \sum_{k=1}^p b_k e_{WM}(n+k) \right)
 \end{aligned} \tag{31}$$

where $y = (b_1, \dots, b_p, s_{WM}(0), \dots, s_{WM}(N-1), \alpha_0, \dots, \alpha_{N-1}, \beta_1, \dots, \beta_p, \gamma_0, \dots, \gamma_{N-1})$ is the augmented unknowns vector and $\alpha_i, \beta_i, \gamma_i$ are the *Lagrange multipliers*. Computing the $p + N + N + p + N$ partial derivatives and posing them equal to 0, we

obtain the following non linear equations to be jointly solved:

$$\frac{\partial \Lambda}{\partial b_k} = 2(b_k - a_k) - \sum_{n=0}^{N-1} \alpha_n s_{WM}(n-k) + \sum_{n=0}^{N-1} \gamma_n e_{WM}(n+k) = 0, \quad 1 \leq k \leq p$$

$$\frac{\partial \Lambda}{\partial s_{WM}(n)} = \left(\alpha_n - \sum_{k=1}^p \alpha_{n+k} b_k \right) + \sum_{k=1}^p \beta_k e_{WM}(n-k) = 0, \quad 0 \leq n \leq (N-1)$$

Derivatives with respect to multipliers α_i , β_i and γ_i yield the imposed constraints:

$$\frac{\partial \Lambda}{\partial \alpha_n} = s_{WM}(n) - \sum_{k=1}^p b_k s_{WM}(n-k) - e_{WM}(n) = 0, \quad 0 \leq n \leq (N-1) \quad (32)$$

$$\frac{\partial \Lambda}{\partial \beta_k} = \sum_{n=0}^{N-1} e_{WM}(n) s_{WM}(n-k) = 0, \quad 1 \leq k \leq p \quad (33)$$

$$\frac{\partial \Lambda}{\partial \gamma_n} = e_{WM}(n) - \sum_{k=1}^p b_k e_{WM}(n+k) = 0, \quad 0 \leq n \leq (N-1) \quad (34)$$

Note that $s_{WM}(n)$ will be known from the previous frame (null or not) for $n < 0$; similarly $e_{WM}(n)$ is to be considered null for $n > N-1$ (future frame not available). Solving this non linear equation system with a suitable solver, it is possible to find the solution to the problem. The chosen solver has a high impact on method complexity, especially considering that most of the involved matrices are sparse. This method is also interesting because it is a new general analysis framework which may be customized in the Lagrange function keeping firm the constraints, to test different perceptual impact functions. Results in terms of recovered data quality are given in Section 9.4..

8. A Case Study: Air Traffic Control

As example simulation scenario, we chose the case of Air Traffic Control, where an aircraft equipped with a speech watermarking encoder transmits in air towards a ground station GS (*reverse link*, A/G) which is remotely connected by a terrestrial network to the airport control tower or ACC (Area Control Center). The focus on this link is arbitrary, the scenario could be equally studied for the inverse communication link (*forward link*, G/A). Currently, ATC relies mainly on radio communications between pilots and controllers; these are analogue transmissions realized with DSB-AM modulation (Double Side Band - Amplitude Modulation) in the 118 – 137 MHz (VHF) frequency band with frequency division multiple access (FDMA). The channel spacing is 25 kHz, but to cope with growing air traffic it has been introduced a new spacing of 8.33 kHz. Air space is divided in sectors, and each of them is covered by one or more radio stations, depending on the orographic features of territory. Since the nature of this link is analogue, speech watermarking techniques are a useful tool for short-medium term augmentation of the capabilities of ATC networks, especially in terms of authentication and new data services for the increase of safety and security of

air traffic control (see [11] and [2] for a deeper explanation of the issue). These techniques must be adapted to the low transmission quality given by noisy channels and narrow signal bandwidth, which constrain the rate of additional hidden data. High rate SpW has features that renders it well suited for this task, both for the actual 25 kHz RF channels and for the new 8.33 kHz RF channels.

The main performance requirements of a speech watermarking system for ATC are the following:

- *Audibility*: the data stream should not disturb the voice communication and still be reliably transmitted. This is important also for *compatibility with not watermarking-ready legacy RF facilities*. MOS should be ≥ 3.6 .
- *Data Rate*: the minimal required data rate is 40 bits/s⁷. Higher rates are useful to deploy new ATC data services.
- *Reliability*: watermarking message bits should be detected with a *Packet Error Rate* (PER) $\leq 10^{-2}$.
- *Delay*: Transmission/decoding delay should be kept < 1 s.
- *RF Channels*: the system must be adaptable to operate in both European RF channel bandwidths, 25 and 8.33 kHz.

9. Simulation Results

9.1. Work Hypotheses and Simplifying Assumptions

For the development and simulation of this system we used MATLAB/Simulink R2009a environment [31]. In deriving our simulation results we made the following assumptions:

- *Audio sources*: a group of proof signals, in .wav format, obtained recording ATC communications directly in the aircraft cockpit in order to not collect channel noise.
- *RF Channel*: direct link or VHF ATC channel [32]; thanks to the envelope-detector and low-pass filter in the RF section of the receiver, Doppler shift has no effect, and the fading effect is reduced, giving results which we found similar to a simple AWGN channel.
- *SNR*: evaluated at the decoder input, it has been varied from 20 to 50 dB, although low values in the range 20 – 30 dB are seldom encountered in the usual ATC operativity.
- 10000 s of speech + data transmitted per simulation run.

⁷A few tens of bits for each message are required in order to have a sufficient number of signatures for the European air traffic: 40 bits is set as the minimum target because assuming, as an example, 16 bits for ancillary fields (timestamp, CRC, and others), the other 24 bits can carry $2^{24} \approx 16$ million signatures, sufficient to satisfy current and future needs.

System performance have been evaluated in terms of BER and PER for the quality of received watermark data and subjective MOS measurements (involving a sufficiently high number of listeners) for watermarked speech quality estimation. Evaluation of voice quality is not easy, since extensive and expensive subjective listenings are needed to make it feasible. Hence, we started our analysis by performing a limited listening campaign, associated with automatic measurement of MOS through ITU's PESQ-LQ (*Perceptual Evaluation of Speech Quality - Listening Quality*) non-intrusive objective method [33][34]. PESQ-LQ is a score homogeneous with subjective MOS scale, and is a very useful method for speech quality assessment. However, *we found PESQ-LQ useless for this type of watermarking*: since the comparison between original and watermarked signals is made in a perceptive domain, and this type of watermark is embedded in the vocal tract excitation, the method fails in providing a realistic measure of the difference of speech quality. So the research of more suited methods of objective speech quality estimation will also be part of future developments. Therefore, the classical MOS measurements for watermarked speech quality are used.

9.2. Channel Coding Results

Figures 14 and 15 show the BER and PER results for uncoded and channel coded watermark. The gain is significant and may be quantified in 3 dB at $\text{BER}=10^{-4}$ and 4 dB at $\text{PER}=10^{-3}$. Suboptimal decoding is only 1 dB worst than optimal decoding, with considerable complexity reduction (about 50%).

Figures 16 and 17 show what are the results when LP parameters are estimated at the receiver and no equalization technique is used: BER and PER are highly degraded, with a floor for high SNRs due to channel filtering function $H(z)$. We can see that the gain derived from the application of coding is even higher due to the strong performance floor reduction.

9.3. Synchronization Results

Figures 18 and 19 show optimal performance for framing synchronization after an extensive campaign of parameters optimization (FS=60, SFR=-3 dB, FF=-24 dBW). SER (Synchronization Error Rate), that is the fraction of missed initial synchronizations on the total of simulated transmissions, is 0 above $\text{SNR}=25$ dB, and MOS is only slightly lower than original voice; note that the degradation effect is negligible, mainly for low SNRs, where it is masked by channel noise.

Figure 20 shows an example of detection probabilities and consequent BER and PER values for packet tagging, comparing them to similar results obtained for previously proposed decoder segmentation [2]. For $\text{SNR}=40$ dB, $\text{WTR}=1$ dB and $S=5$ are the optimal values, for which we have $P_{rs} = 0.9991$, $P_{md} = 8.6 \cdot 10^{-4}$ and $P_{fa} = 7.5 \cdot 10^{-5}$, better values than that for decoder segmentation. The enhancement in terms of BER and PER is clear, deducible also from Figure 22, where it can be seen that the performance are acceptable from 40 dB and only max 3 dB distant from ideal case.

Figures 21 and 22 show the performance comparison in terms of BER and PER in the case of ideal LP parameters among cases of ideal segmentation, V/U segmentation operated at the receiver and tagging technique application. We can see that packet tagging is very

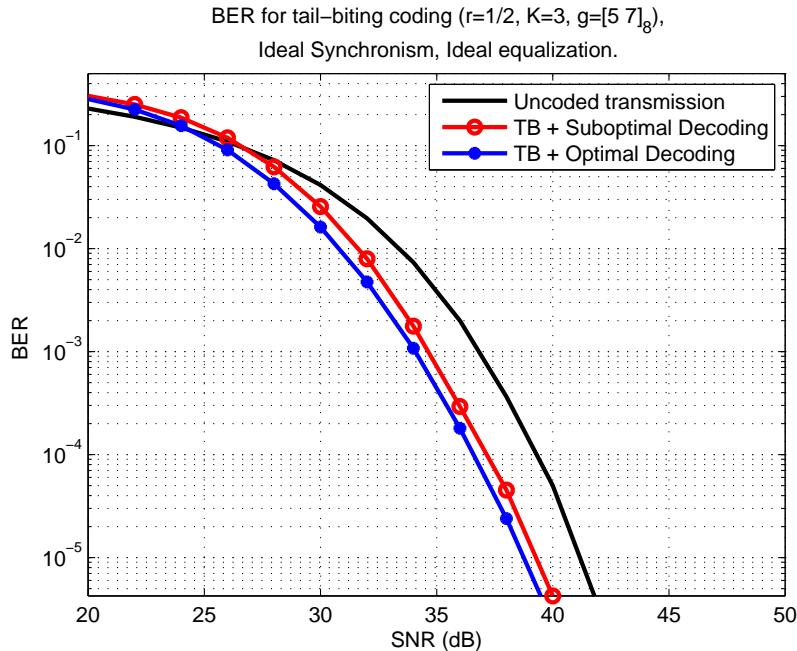


Figure 14. BER comparison for Tail-biting coding vs. uncoded transmission, ideal synchronization and equalization.

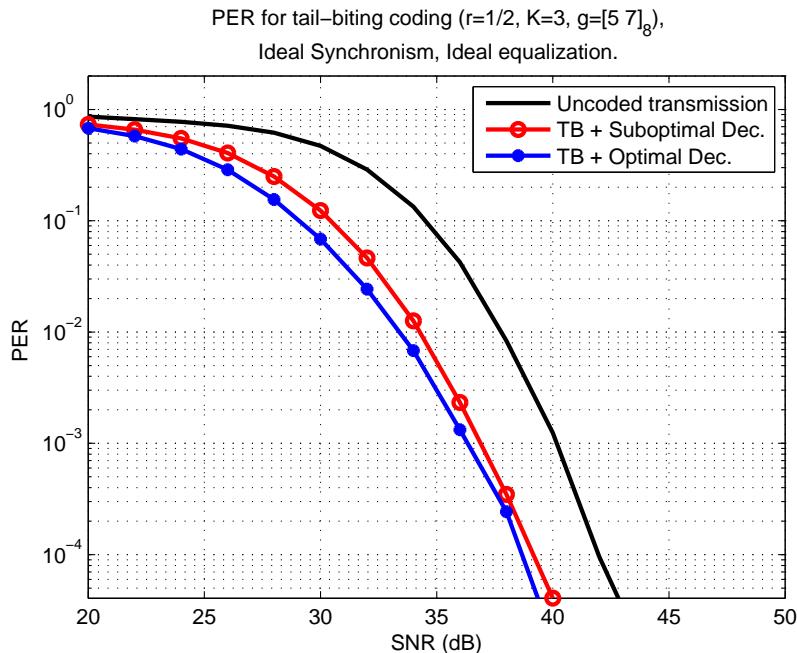


Figure 15. PER comparison for Tail-biting coding vs. uncoded transmission, ideal synchronization and equalization.

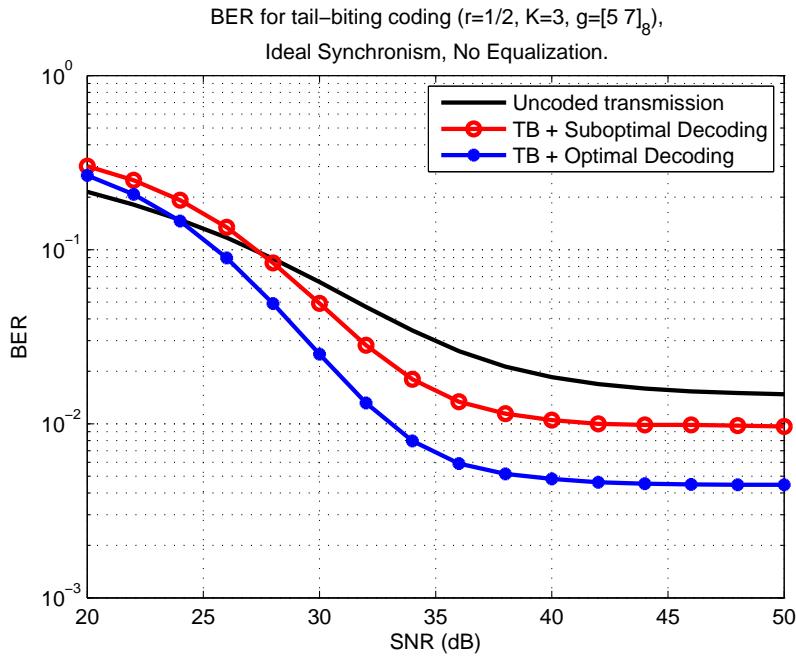


Figure 16. BER comparison for Tail-biting coding vs. uncoded transmission, ideal synchronization, no equalization.

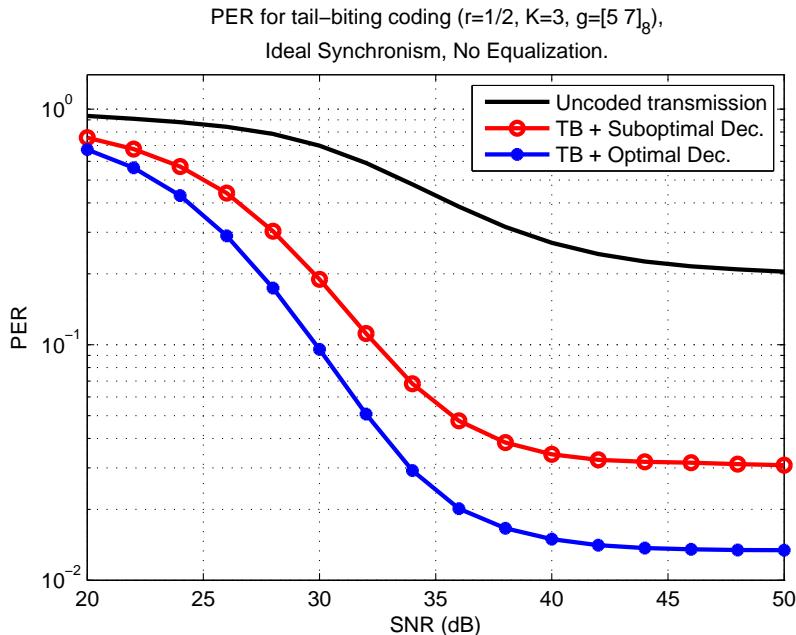


Figure 17. PER comparison for Tail-biting coding vs. uncoded transmission, ideal synchronization, no equalization.

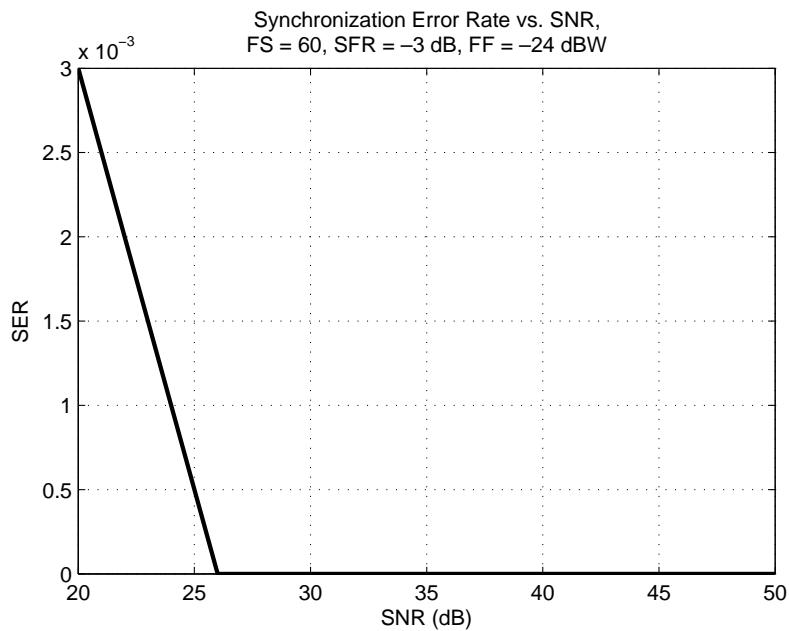


Figure 18. SER for Framing Synchronization technique vs. SNR, FS=60, SFR=-3 dB, FF=-24 dBW.

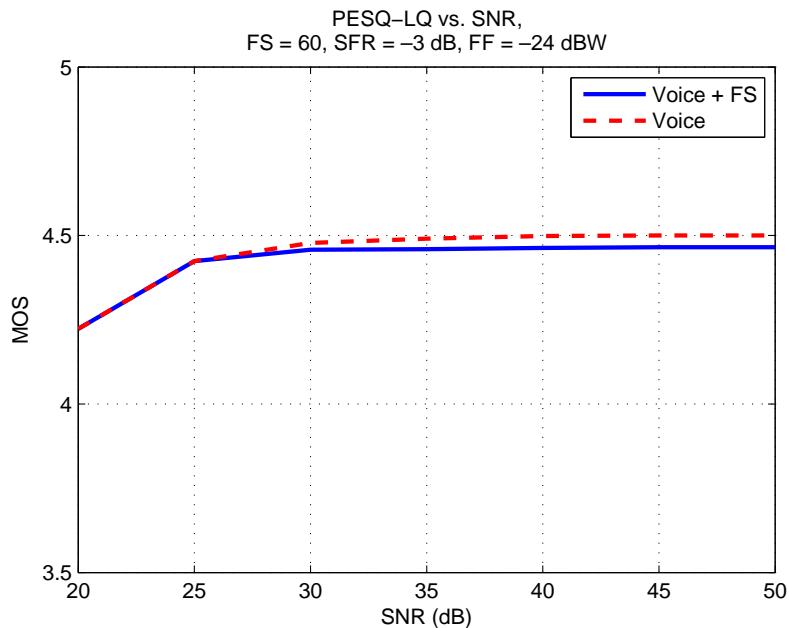


Figure 19. MOS for Framing Synchronization technique vs. SNR, FS=60, SFR=-3 dB, FF=-24 dBW.

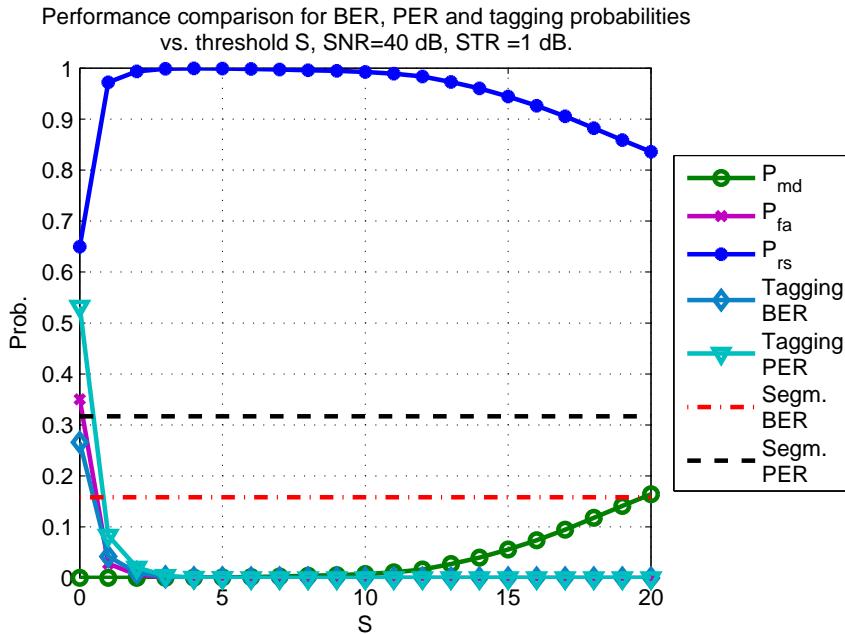


Figure 20. Performance comparison for BER, PER and tagging probabilities vs. threshold S, SNR=40 dB, WTR=1 dB.

effective in recovering performance, with a loss of only 2 dB with respect to ideal case and an acceptable floor located at high SNR values due to the interference between watermark data and tagging sequence. On the contrary, performance of receiver segmentation is unacceptable.

9.4. Equalization Results

Figure 23 shows the strong performance recovery when preemphasis technique is used, both in terms of absolute PER and of floor mitigation. However, a distance of 7 dB from ideal PER value at $\text{PER}=10^{-2}$ indicates that proposed algorithm is not completely effective: future developments will have to focus on equalization enhancements.

9.5. Perceptual Quality of Received Speech

Generally speaking, the estimated quality of received voice is very good ($\text{MOS} \approx 4.3 - 4.4$) in all estimated cases; the proposed modifications do not degrade significantly marked voice quality. Only when SNR is very low or when using preemphasis technique, careful subjective listening over quality headphones showed that a slight difference between the original and the processed speech sound is audible to the expert listener, but, especially for noisy speech, this difference is not disturbing and does not degrade the perceived speech quality. The requirement of $\text{MOS} \geq 3.6$ is always respected, and for most of time $\text{MOS} \geq 4.1$ at all SNR values, an excellent level for ATC communications.

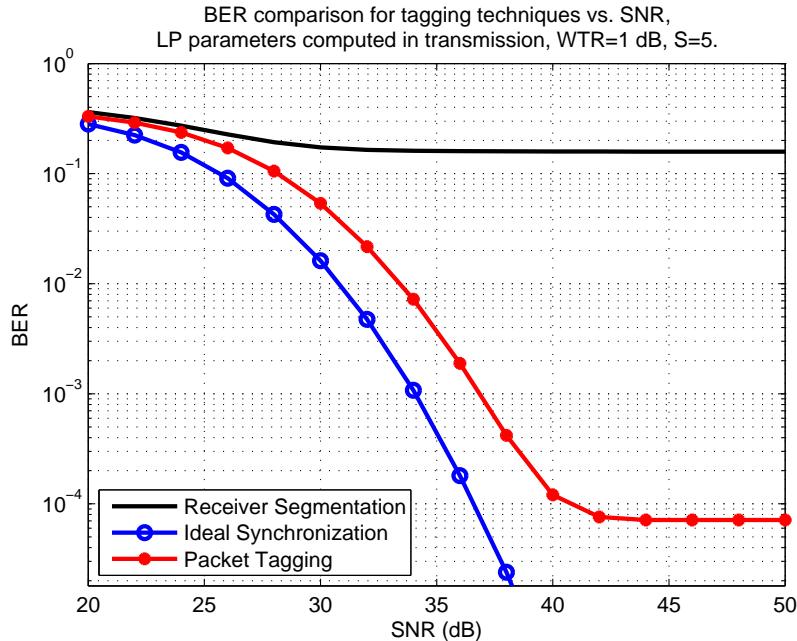


Figure 21. BER comparison for tagging techniques vs. SNR, LP parameters computed in transmission, WTR=1 dB, S=5.

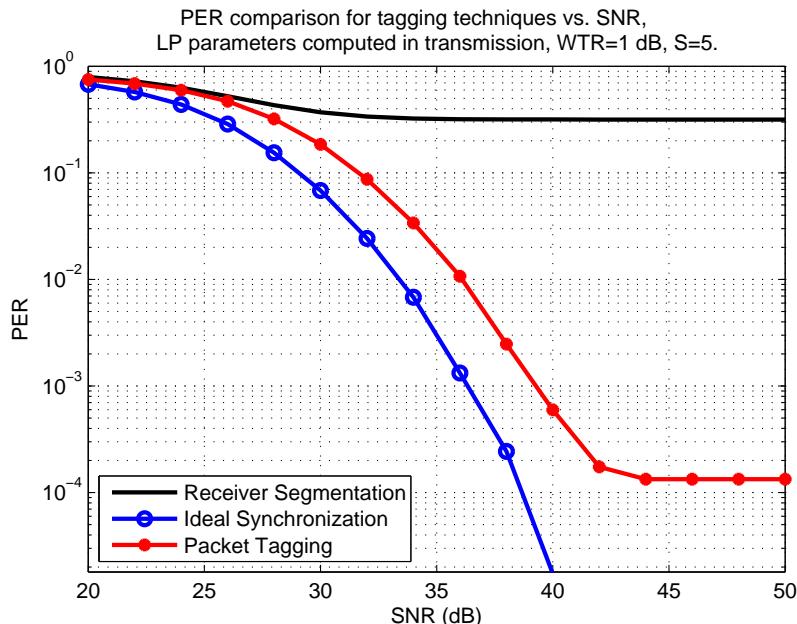


Figure 22. PER comparison for tagging techniques vs. SNR, LP parameters computed in transmission, WTR=1 dB, S=5.

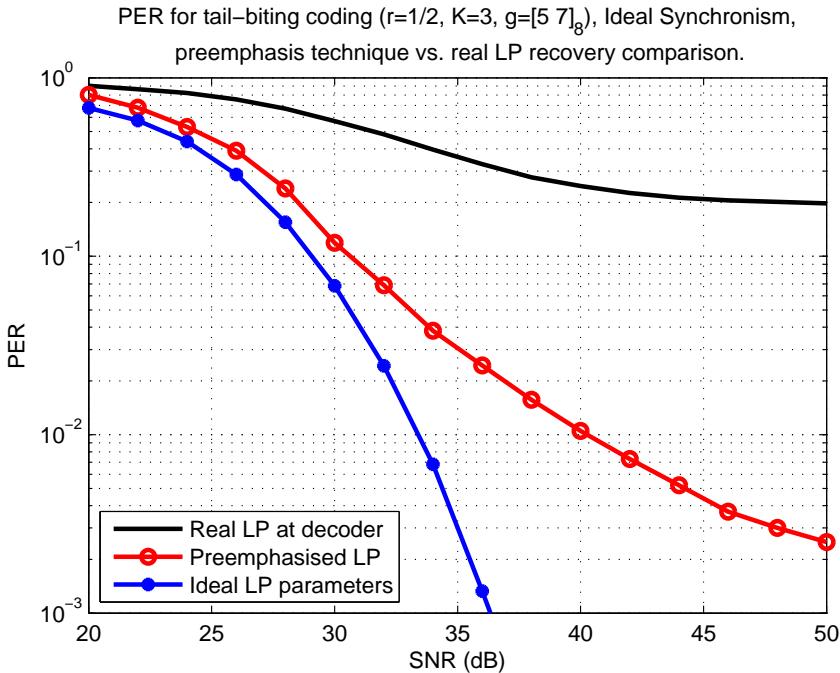


Figure 23. PER comparison for preemphasis technique vs. LP parameters recovery as a function of SNR, Ideal Synchronism.

9.6. Results Related to Choice of Prediction Order p

A performance study has been also performed as prediction order p varies, in order to justify the choice of its value. Simulations have been performed in the case of channel coding with optimal decoding, both with ideal and real LPC parameters, with values of p equal to 7, 8, 10 and 12. In the following we show the results related to the simulations performed for three SNR values and LPC coefficients computed at receiver (no preemphasis technique) in the cases of real and ideal synchronization.

Figure 24 shows the energy of the prediction error for linear prediction performed at the decoder: the trend is growing with p instead of decreasing as usual. This is due to the fact that the parameters estimation at the receiver is not perfect for the reasons explained in Section 7. and the errors affect mostly the coefficients with the highest order, so more coefficients mean a greater mean prediction error. This effect reflects on the system performance in terms of BER and PER, as it can be seen in Figures 25 and 26 respectively for the two extreme SNR values of 20 dB and 50 dB. Aside from the obvious differences in terms of absolute BER and PER values, both indicators are slightly increasing as p increase, showing that a high value of the order of prediction, while producing a lower prediction error at transmitter side, does not yield better overall system performance. Remembering that the operation related to linear prediction dominant in terms of complexity is the analysis, which has an order of $O(p^2)$ thanks to Levinson-Durbin algorithm, a good trade-off between performance and complexity suggests the choice of a low value of prediction order. Al-

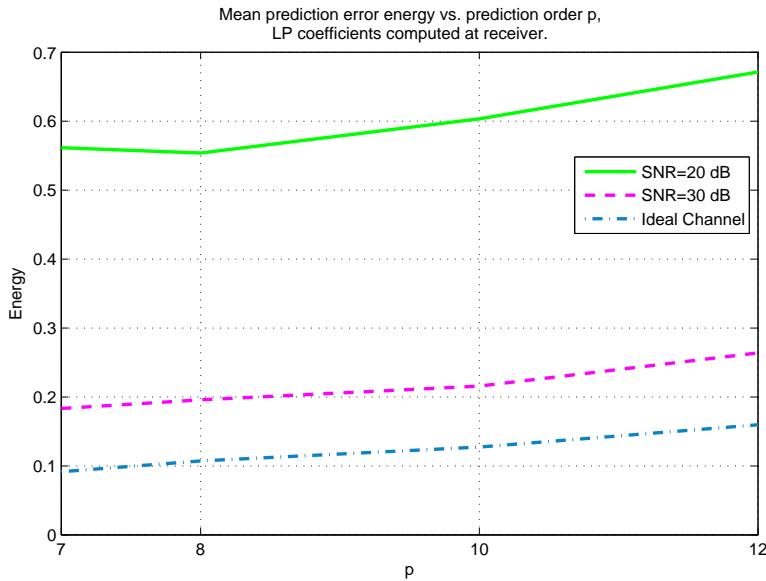


Figure 24. Mean prediction error energy vs. order of prediction p , LP parameters computed at receiver.

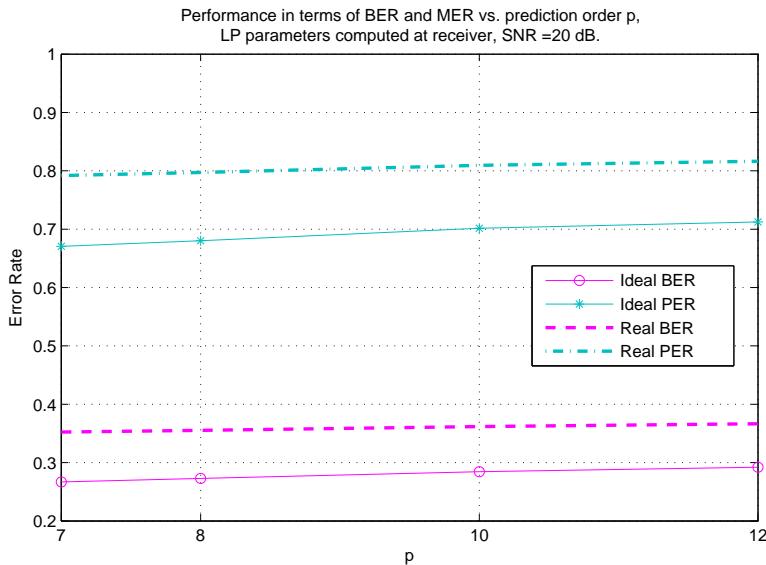


Figure 25. Performance in terms of BER and MER vs. prediction order p , LP parameters computed at receiver, SNR=20 dB.

though not shown in the figures, values under 7 are characterized by decreasing trends, so $p = 7$ is the optimal choice.

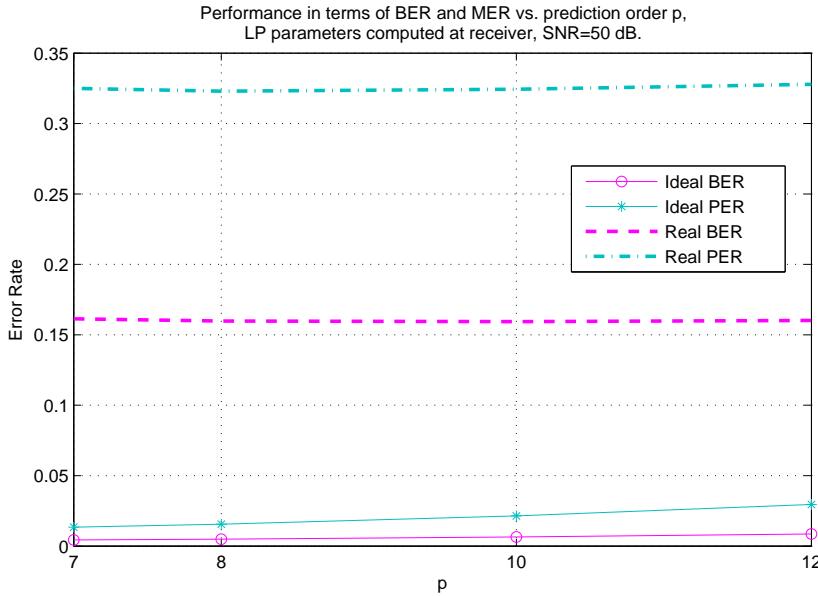


Figure 26. Performance in terms of BER and MER vs. prediction order p , LP parameters computed at receiver, SNR=50 dB.

10. High Rate Speech Watermarking for IP Networks

In this section we present a specific application for the augmented high rate watermarking system, partially related and integrated to the ATC scenario presented in the previous sections. This example is aimed to demonstrate the system's versatility even in a context normally considered far from usual application of watermarking techniques for analogue channels, the transmission of voice on a digital IP network. The reader is advised to first read the previous sections in order to better understand issues specific to this part.

Voice over Internet Protocol (VoIP) is the technology aimed to realize convergence between the old circuit-switched PSTN (Public Switched Telephone Network) telephony service and the packet-switched IP data networks [35]. Real-time VoIP is a challenge due to the packet-switching and connection-less nature of the IP networks and the strict QoS (Quality of Service) requirements of the conversational class of service: packet switching optimizes the use of resources but the handling of fragmentation and multiple routes is not always compatible with required quality of services. However, the integration is nowadays completely realized and it is able to guarantee a good QoS in the context of Local Area Networks (LAN). In Metropolitan Area Networks (MAN) or Wide Area Networks (WAN) it is more difficult to adhere to specific QoS requirements in terms of scalability (dependent from specific needs), reliability and interoperability (protocols, equipments, etc.). We can enumerate many advantages for VoIP over traditional telephony, as lower calling costs for long distances, lower infrastructure costs, implementation capability for new advanced functionalities (video calls, email answering, instant messaging, etc.), easy equipment expandability.

Though major advancements in quality and reliability have been made in the recent years, there are other important issues to consider, as security and bandwidth consumption. The first is usually addressed with the use of cryptography techniques [36] and digital signatures contained in the packet header of IP or higher protocol layers; however, computational costs may be too high and there is also the possibility of third parties malicious attacks such as identity spoofing. Moreover, the constantly growing use of these new technologies both by public and private users has raised considerably the centrality of IT security-related problems, becoming more and more important for privacy, frauds prevention and, speaking generally, criminal acts related to not secure and authenticated use of Internet telephony.

In this section we propose, as an interesting application example, the use of digital speech watermarking systems for the augmentation of VoIP networks as a mean to obtain a secure hidden data channel for an alternative faster, lightweight and more robust authentication architecture, for conversation-related data services or secure VoIP protocol signaling, without additional cost for bandwidth. If combined with existing security mechanisms, it can greatly improve VoIP system's security. Since this chapter is focused on high rate watermarking, we will present here only results related to this system, but obviously more traditional watermarking approaches as that presented in Section 3.1. could be combined with VoIP service. We evaluated High Rate SpW in conjunction with three popular VoIP codecs; their coexistence is evaluated in terms of specific QoS requirements, such as received voice quality and watermark extraction latency and reliability, showing the feasibility of authentication service in some of the evaluated conditions and the possibility of additional data services in most usage scenarios.

10.1. VoIP Physical and Security Protocols

10.1.1. Voice over IP Issues

There are many factors which impact on voice quality in IP networks, some related to their nature, architecture, topology and protocols, such as latency, packet loss, jitter; other factors are related to voice processing operations such as filtering and compression and, in particular, compression algorithms, transducers quality, echo-cancellation algorithms [35]. In this chapter we will focus on the second factors, though keeping in mind the importance of the first ones and considering them in the example scenario used for performance evaluation.

Packet-switching allows to use the same channel to transport at the same time heterogeneous data. If, in traditional telephony, time and distance are the main parameters with an impact on costs, in IP transmissions we get used bandwidth. So the costs reduce using less bandwidth. It is usually difficult to predict the QoS that we may expect from a VoIP session but, usually, this is directly proportional to the bandwidth available for the conversation. Practically, since IP protocol does not guarantee a reliable service, we can achieve this raising the available bandwidth, trying to reduce at minimum the causes of latency and packet loss, that are certainly more frequent in a congested network. Protocols like IntServ/RSVP and DiffServ [35] are also useful in ensuring the required QoS for VoIP service. It is clear that practical design considerations suggest a trade-off choice on bandwidth between voice quality and implementation costs.

Table 1. Characteristics of some ITU-T standardized speech codecs.

Codec	Bit rate (kbps)	Complexity (compared to G.726)	Algorithmic delay (ms)	Quality (MOS)
G.711	64	very low	0.125	4
G.723	5.3	8	37.500	3.9
G.723.1	6.3	8	37.500	3.6
G.726	32	1	0.125	3.9
G.728	16	15	0.625	3.6
G.729	8	10	15.000	3.9
G.729A	8	6	15.000	3.7

10.1.2. Speech Codecs

Audio *codecs* are the algorithms used for compressing audio captured from an external source and for decompressing and reproducing it when needed (compression and decompression operations) [35]. These tools are used in VoIP to reduce the bandwidth consumption. Here we focus on *speech codecs*: they exploit the features of human voice to achieve high compression factors.

In order to allow the performing of effective vocal communications, compression techniques must be framed in known standards. The institute that has done this in the voice field is mainly ITU-T [37], with the standards showed in Table 1. They have important differences basing on parameters like bit rate, computational complexity, algorithmic delay and voice quality. If we assume, as the same starting point for all codecs, a PCM voice signal sampled at 8 kHz with 16 bits quantization, for an original bit rate of 128 kbit/s, the codec output rate is an index of its compression ratio. Generally speaking, the higher the codec's compression, the lower the voice quality will be. Complexity and latency are also important parameters, the first particularly in energy-constrained applications scenarios, the second always when considering total system latency. So, the codec choice for a given communications system is always driven by a trade-off between these parameters. We do not consider here other free codecs like Speex [38] or iLBC [39] or proprietary codecs like that of GSM system [40] or proprietary programs like Skype [41], but the study could be extended to them in a straightforward way.

We recall here the main features of the three codecs evaluated in this paper, that is G.711, G.726, G.729 algorithms.

G.711 It is a very basic, widely used, codec; it accepts as input a PCM speech signal coded with 14 bits (μ -law, used in US and Japan) or 13 bits (A-law, used in Europe) bits. It outputs exponentially coded 8-bit samples, for a standard bit rate of 64 kbit/s.

G.726 It is an evolution of G.711 with the major addition of Adaptive Differential Pulse Code Modulation (ADPCM), which adaptively quantizes samples of the voice residual [35] based on an input 8-bit G.711-encoded speech signal. The output may be encoded with 2, 3, 4 or 5 bits yielding bit rates respectively of 16, 24, 32 or 40 kbit/s.

G.729 It is an advanced codec using Conjugate Structure - Algebraic Code-Excited Linear Prediction (CS-ACELP) encoding to compress blocks of 80 16 bit-samples (10 ms) in packets of 10 byte (8 kbit/s) for main payload or 2 bytes for CNG (Comfort Noise Generator). G.729 specifies also the algorithms for VAD (Voice Activity Detector) and CNG. G.729A is a G.729 variant which requires less computational resources with slightly lower sound quality.

This choice of codecs is driven not only by their relative popularity, but also from the need to test a high rate, a medium rate and a low rate codec in order to understand how this choice impacts the joint VoIP and watermark systems performance. For even a greater detail level, G.726 has been tested in all its output configurations.

10.1.3. Security and Authentication

Security for IP telephony is related, differently from traditional one, to many risk factors which may compromise the call quality or success [36][42]. Among the many types of risks for the security of modern data networks we may cite some notable ones:

- DOS (*Denial Of Service*), where a malicious party attacks an IT system providing a service (e.g. a web site) with a massive amount of service requests in order to saturate it, make it unstable and cause it not being able to provide service anymore. Any system connected to Internet and providing TCP (Transport Control Protocol)-based network services is subject to the risk of DOS attacks.
- *Man-in-the-middle*, attacks which allow third parties to monitor, record, block or even modify a data transmission. Usually packet sniffing or capturing accompanies this type of attacks. Its countermeasures are represented by cryptography and authentication.
- *Trojan horses and malware*: they are software processes designed for propagation through Internet and IP-based networks that infect hosts in order to replicate themselves. Trojans do not self-diffuse like viruses or worms, so they require a direct aggressor intervention to send the malicious executable to the victim.

For a VoIP network maintainer, security is divided in three categories:

- Access Control
- Software Maintenance
- Intrusion Prevention

In this work the most important aspect is the access control through user or software process authentication, that is the possibility to authorize the use of particular resources through credentials represented by secret data as a password, an encrypted string or some other secret key [42].

For a secure VoIP communication, it is usually adopted some asymmetric key encryption technique [36] coupled with a digital signature containing the information relative to the identification of the user and inserted within the IP (or higher protocol layer) packet.

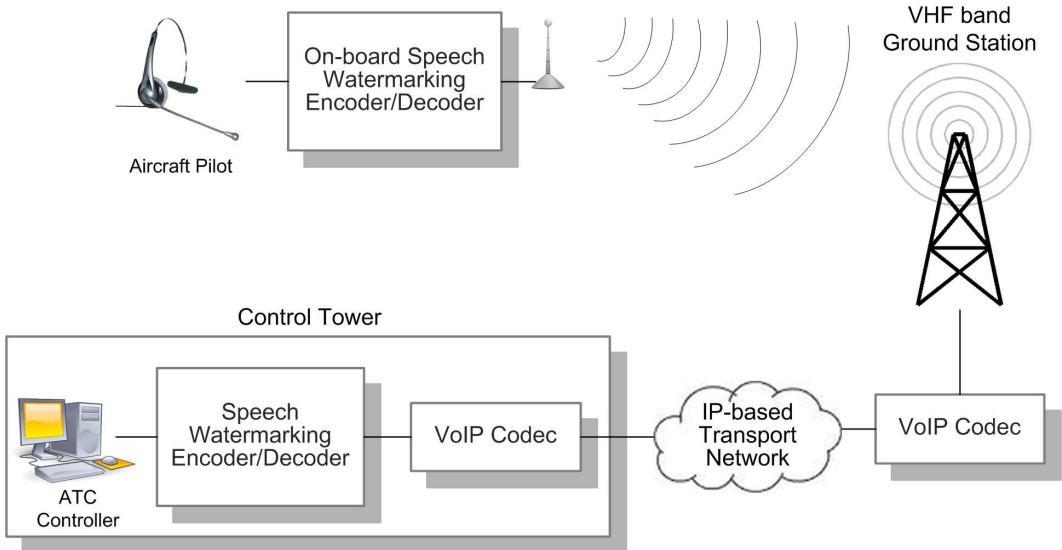


Figure 27. Scheme of evaluated Air/Ground ATC scenario.

These techniques involve computational costs which cause also additional transmission delays. Moreover, as above mentioned, there is the possibility that a third party substitutes itself to one of the authorized parties involved in the conversation (*identity spoofing*) thanks to a successful man-in-the-middle attack. Clearly this is a risk even more dangerous in communications among multiple parties.

We propose speech watermarking as an idea for an automatic authentication system for voice working to the lowest conceivable level, that is with digital signature/auxiliary data embedded as integral part of the samples of the voice signal itself. In this way data are invisible to packet sniffing and other attacks related to cryptography, IP layer or higher layers. Nonetheless, embedded data may be encrypted to further raise security level in case it is needed.

10.2. A Case Study: IP-Based Aeronautical Network

For better ease of reading, we report here from Section 8. the description of the main facts about the current civil air traffic communications. As example simulation scenario, we chose the case of an Aeronautical Terrestrial Network (ATN), as depicted in Figure 27, where an aircraft equipped with a SpW encoder transmits in air towards a ground station GS (*reverse link*, A/G), remotely connected by a terrestrial network to the airport control tower or ACC (Area Control Center). The focus on this link is arbitrary, the scenario could be equally studied for the inverse communication link (*forward link*, G/A). Currently, ATC relies mainly on radio communications between pilots and controllers; these are analogue transmissions realized with DSB-AM modulation (Double Side Band - Amplitude Modulation) in the 118 – 137 MHz (VHF) frequency band with frequency division multiple access (FDMA). The channel spacing is 25 kHz, but to cope with growing air traffic, it has been introduced a new spacing of 8.33 kHz. Air space is divided in sectors, and each of them

is covered by one or more radio stations, depending on the orographic features of territory. Since the nature of this link is analogue, speech watermarking techniques are a useful tool for short-medium term augmentation of the capabilities of ATC networks, especially in terms of authentication and new data services. These techniques must be adapted to the low transmission quality given by noisy channels and narrow signal bandwidth, which constrain the rate of additional hidden data.

Digitalization of the entire ATN is expected, at least in Europe, for the time frame 2015-2020, with SESAR (Single European Sky for ATM Research) project [43]. This is a long term for air interfaces, but the update for ground equipment and terrestrial network is more straightforward and expected in the next few years. While there are modernization problems for the aeronautical communication system (costs, security regulations, certifications), the shift to IP-based ground networks will yield lower maintenance costs of the airport network (both for equipment and personnel), bandwidth resources optimization and a greater flexibility. In the meantime, analogue-digital coexistence requires the designers to verify the *integrability* of the ATC speech watermarking models with VoIP technology, namely to verify that SpW does not degrade VoIP QoS with respect to minimal requirements and that VoIP codecs and networks impairments do not degrade too much SpW expected performance.

10.3. Performance Evaluation

10.3.1. Reference Scenario

In order to evaluate the performance of the proposed approach in terms of data and speech quality, extensive simulation experiments have been carried out by means of the MATLAB/Simulink [31] and OMNeT++ [44] development frameworks, with the addition of VoIPTool framework [45]. The experiments initially concerned three usage scenarios related to the case study illustrated in subsection 10.2.:

- Direct link between Ground Station and controller with speech watermarking but no codec: the performance limit is given only by the voice-watermark coexistence;
- Direct link between Ground Station and controller with speech watermarking and VoIP codec;
- Speech watermarking + VoIP transmission and AWGN channel for air interface.

In addition to this, a transport IP-based network (Figure 28) has been placed between the GS and controller site to simulate a private airport LAN. In this network, the “server” node uses static routing tables to route voice packets on UDP (User Datagram Protocol) datagrams towards “centralhost” node. The routers backbone handles packet transmission among network hosts; the links are realized with Ethernet technology [46]. This choice of a private local network permits us to freely variate network parameters starting with negligible delays, jitter and packet loss and then raising them to test overall performance and simulate changing network conditions.

Table 2 shows the main simulation parameters for the entire scenario. The channel code has been chosen after theoretic considerations and an extensive optimization campaign of

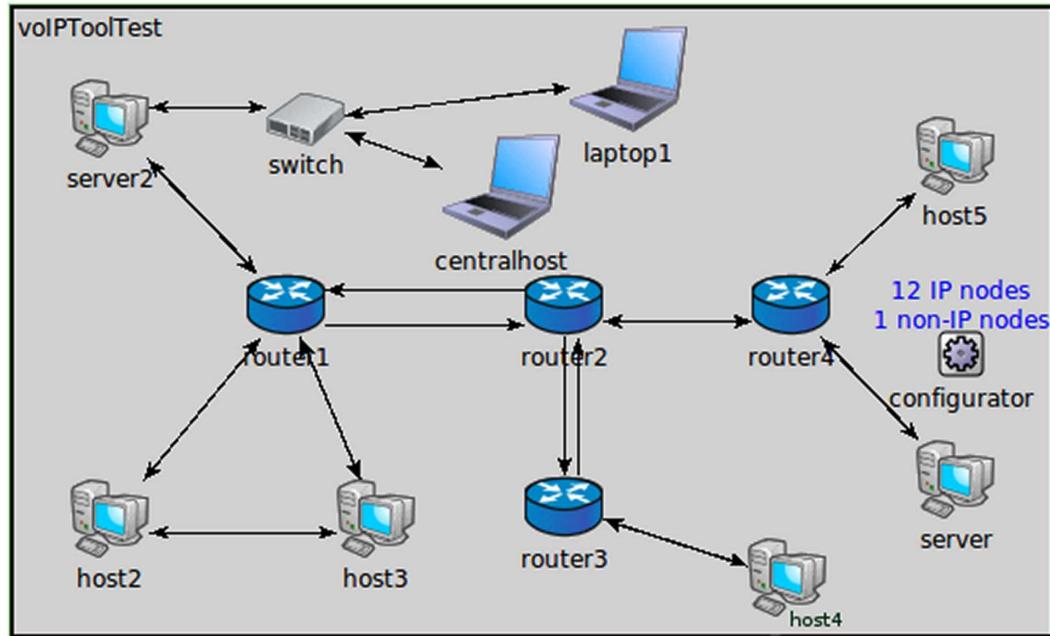


Figure 28. Test network topology as implemented in OMNeT++ framework.

Table 2. Parameters values adopted for system simulation.

Parameter	Value
Tail-biting codes	$r = 1/2, K = 3, g = [5 \ 7]_8$
Simulation duration	1000 s
A/G channel	AWGN (SNR=40 dB)

simulations (as reported in Section 5.). Two audio files (sampled at 8 kHz frequency and encoded with 16 bits) have been used: “takeoff” is the registration, in an aircraft cockpit, of the pilot ATC transmission, while the second (“phonecall”) is a registered common phone conversation, with lower noise level.

The evaluated quality parameters are BER and PER for watermark data and MOS and P.563⁸ [47] scores for voice quality. The last two were used instead of PESQ-LQ due to a lack of applicability for high rate model, as previously explained in subsection 9.1.. The acceptable work conditions are $\text{BER} \leq 10^{-3}$, $\text{PER} \leq 10^{-2}$ and $\text{MOS} \geq 3.6$, similar to requirements explained in Section 8..

⁸P.563 is a non-intrusive objective method for evaluation of speech quality, indicated when it is not possible to compare the degraded voice sample with the original one because this last one is not available. In this case it is used for its better score coherence than PESQ method.

Table 3. Data and voice performance for High Rate SpW model related to G.711 codec.

Output Parameter	BER	PER	MOS	P.563
Without codec	$8.0 \cdot 10^{-5}$	$1.1 \cdot 10^{-4}$	4.4	4.329
G.711	$8.0 \cdot 10^{-5}$	$1.1 \cdot 10^{-4}$	4.3	4.321
G.711 + AWGN	$1.2 \cdot 10^{-4}$	$9.0 \cdot 10^{-4}$	4.1	4.033

10.3.2. Numerical Results

Table 3 shows the watermark/voice performance for G.711 codec and “takeoff” file; the codec does not introduce data losses but only a negligible variation of voice quality. The AWGN channel yields a slight, although dominant, decrease of BER and PER and an acceptable decrease of voice quality indexes. They are predictable results, since G.711 is a low compression codec. P.563 scores are rather consistent with our subjective listening, showing it can be a good substitute for PESQ in this case. The usage of “phonecall” file has given almost identical results.

Table 4. Data and voice performance for High Rate SpW model related to G.726 codec (file “phonecall”, Silence Threshold=0).

File	BER	PER	MOS	P.563
<i>phonecall</i>	$1.3 \cdot 10^{-4}$	$1.5 \cdot 10^{-4}$	4.2	4.161
<i>phonecall_40</i>	$1.5 \cdot 10^{-4}$	$1.8 \cdot 10^{-4}$	4.2	4.120
<i>phonecall_32</i>	$6.9 \cdot 10^{-4}$	$7.3 \cdot 10^{-4}$	3.7	3.802
<i>phonecall_16</i>	$3.1 \cdot 10^{-3}$	$3 \cdot 10^{-3}$	3.5	3.697

For G.726 we show only the interesting worst-case results (codec + channel). Tables 4 and 5 show the results for HR+G.726 case for “phonecall” file and output bit rates 40 kbits/s, 32 kbits/s and 16 kbits/s, respectively for a Silence Threshold of 0 and 300, which determines the amplitude level under which voice is considered silence and is discarded in the compression phase. The case for 24 kbits/s is not showed due to results very similar to 32 kbits/s case. It can be seen that a growing compression factor yields a significant increase of BER and PER. If in the first case 40 kbits/s produces a negligible voice and data degradation, higher compression factors have a big impact on voice and data quality, until, at 16 kbits/s, data quality is a bit lower than intended targets. As expected, a silence threshold higher than 0 has severe detrimental effects on PER and BER, a phenomenon related to the elimination of silence data frames: although voice quality remains always acceptable, data quality is acceptable only for 40 kbits/s case. Therefore, high compression ratios with silence substitution are not compatible with high rate SpW. Similar, and slightly better, results are obtained with “takeoff” file.

Table 6 shows the model’s performance with G.729 codec with VAD deactivated (no silence substitutions). With the highest compression ratio of the considered codecs, HR model is far from the intended target performance. Their joint use is feasible only with a

Table 5. Data and voice performance for High Rate SpW model related to G.726 codec (file “phonecall”, Silence Threshold=300).

File	BER	PER	MOS	P.563
<i>phonecall</i>	$1.3 \cdot 10^{-4}$	$1.5 \cdot 10^{-4}$	4.2	4.161
<i>phonecall_40</i>	$7.4 \cdot 10^{-4}$	$4.6 \cdot 10^{-4}$	4.2	4.122
<i>phonecall_32</i>	$2.7 \cdot 10^{-3}$	$3.3 \cdot 10^{-3}$	3.7	3.653
<i>phonecall_16</i>	$8.1 \cdot 10^{-3}$	$8.9 \cdot 10^{-2}$	3.5	3.564

Table 6. Data and voice performance for High Rate SpW model related to G.729 codec.

File	BER	PER	MOS	P.563
<i>takeoff</i>	$6.8 \cdot 10^{-3}$	$7.9 \cdot 10^{-3}$	4.0	3.976
<i>phonecall</i>	$7.1 \cdot 10^{-3}$	$8 \cdot 10^{-3}$	3.9	3.921

strong enhancement of watermark protection (e.g. channel coding) in spite of net data rate. Results are very similar for the two audio files, showing a low influence on performance for the intrinsic noise registered in the aircraft.

Table 7. Data and voice performance for High Rate SpW model + codecs under Packet Loss conditions (file “takeoff”, Silence Threshold=0).

Packet Loss	Codec	BER	PER	MOS
1%	G.711	$8.8 \cdot 10^{-5}$	$2.2 \cdot 10^{-4}$	4.1
	G.726	$1.1 \cdot 10^{-3}$	$1.5 \cdot 10^{-3}$	3.5
	G.729	$1.8 \cdot 10^{-2}$	$2.0 \cdot 10^{-2}$	3.7
2%	G.711	$3.7 \cdot 10^{-4}$	$4.9 \cdot 10^{-4}$	3.8
	G.726	$4.7 \cdot 10^{-3}$	$6.3 \cdot 10^{-3}$	3.7
	G.729	$7.2 \cdot 10^{-2}$	$8.0 \cdot 10^{-2}$	3.7
10%	G.711	$7.3 \cdot 10^{-3}$	$9.5 \cdot 10^{-3}$	2.7
	G.726	$1.2 \cdot 10^{-2}$	$3.6 \cdot 10^{-2}$	2.4
	G.729	$2.6 \cdot 10^{-1}$	$6.5 \cdot 10^{-1}$	2.3

Network *packet loss* impact on performance is reported in Table 7 for values of 1%, 2% and 10%, assuming that Packet Loss Concealment (PLC) techniques are used for each codec to maintain an acceptable voice quality. It may be seen that unfortunately HR model is very sensitive to packet loss, making it unsuitable for networks with no guaranteed QoS. However, for low packet losses, typical of a reliable LAN network, it may give the desired performance.

11. Conclusion

This chapter is focused on the proposal of novel techniques to overcome flaws inherent in the original algorithm for high rate speech watermarking proposed in [1] by K. Hofbauer and G. Kubin, such as the lack of a suitable channel coding technique, an incomplete synchronization solution and the lack of hidden channel equalization. The tail biting coding gives at least 4 dB of PER gain and more when combined with non ideal segmentation conditions. Frame synchronism and packet tagging are the solution proposed to synchronize receiver with performance superior to the simple segmentation operated at the receiver. Preemphasis technique lowers the performance floor of about two orders of magnitude compared to a not equalized case. These techniques greatly simplify the decoder structure and complexity at the cost of a moderate increase of encoder complexity; the achieved performance gains make feasible the implementation of the system as a solution for storage or transmission of digital data with a higher rate than previous systems, integrating them in a speech signal without affecting significantly its quality. We also showed the feasibility of integration with a VoIP codec on an IP-based aeronautical terrestrial network for services of authentication or IP signaling transport, with reliable good data and voice quality in most every considered scenario, with exclusion of the highest compression ratios and packet loss ratios. Other possible envisioned applications are talker authentication, instant messaging, secure transmission of network signaling or, in the farther future, real-time text transcription of the current conversation.

Acknowledgments

This work has been partially supported by MIUR-FIRB *Integrated Systems for Emergency (“InSyEme”)* project under the grant RBIP063BPH and by the Italian National Project *Wireless multiplatfOrm mimo active access netwOrks for QoS-demanding muLti-media Delivery (“WORLD”)* under grant number 2007R989S.

References

- [1] K. Hofbauer, H. Hering. High-Rate Data Embedding in Unvoiced Speech. In *Proc. INTERSPEECH 2006 - Ninth International Conference on Spoken Language Processing (ICSLP)*, Pittsburgh, USA, Sep. 2006.
- [2] R. Fantacci, S. Menci, L. Micciullo, L. Pierucci. A secure radio communication system based on an efficient speech watermarking approach. *Wiley Security and Communication Networks Journal*, 2(4):305–314, 2008.
- [3] I. J. Cox, M. L. Miller, J. A. Bloom. *Digital Watermarking*. Morgan Kaufmann Publishers, 2001.
- [4] M. Hagmüller, H. Hering, A. Kröpfl, G. Kubin. Speech watermarking for air traffic control. In *Proc. European Signal Processing Conference (EUSIPCO) 2004*, volume 1, pages 1653–1656, Vienna, Austria, Sep. 2004.

- [5] R. Fantacci, S. Menci, L. Pierucci, C. Borgioli, P. Fantappiè, P. Maltese. Efficient Speech Watermarking for Air Traffic Control Narrow-Band Communications. In *Proc. IEEE WRECOM 2007*, Oct. 2007. Rome.
- [6] R. Fantacci, G. Maneschi, S. Menci, L. Pierucci. An Advanced Speech Watermarking System for Air Traffic Control Communications Based on MPEG Psychoacoustics Models. In *Proc. IEEE WRECOM 2007*, Oct. 2007. Rome, Italy.
- [7] M. Barni, F. Bartolini. *Watermarking Systems Engineering - Enabling Digital Assets Security and Other Applications*. Signal Processing and Communications Series. Marcel Dekker, 2004.
- [8] Q. Cheng, J. Sorensen. Spread Spectrum Signaling for Speech Watermarking. In *Proc. IEEE ICASSP 2001*, volume 3, pages 1337–1340, Salt Lake City, UT, USA, May 2001.
- [9] H. S. Malvar, D. A. F. Florencio. Improved spread spectrum: a new modulation technique for robust watermarking. *IEEE Transactions on Signal Processing*, 51(4):898–905, Apr. 2003.
- [10] B. Chen, G. W. Wornell. Quantization index modulation: a class of provably good methods for digital watermarking and information embedding. *IEEE Trans. on Information Theory*, 47(4):1423–1443, May 2001.
- [11] M. Hagemüller, G. Kubin, (TU Graz). Speech Watermarking for Air Traffic Control. Technical report, EUROCONTROL - European Organization for the Safety of Air Navigation, February 2005.
- [12] M. Hatada, T. Sakai, N. Komatsu, Y. Yamazaki. Digital watermarking based on process of speech production. In *Proc. of SPIE - Multimedia Syst. and Appl.* V, 2002.
- [13] M. Celik, G. Sharma, A. M. Tekalp. Pitch and duration modification for speech watermarking. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2005.
- [14] B. Geiser, P. Jax, P. Vary. Artificial bandwidth extension of speech supported by watermark-transmitted side information. In *Proc. of the 9th European Conf. on Speech Communication and Technology EUROSPEECH*, 2005.
- [15] S. Sakaguchi, T. Arai, Y. Murahara. The effect of polarity inversion of speech on human perception and data hiding as an application. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2002.
- [16] L. Girin, S. Marchand. Watermarking of speech signals using the sinusoidal model and frequency modulation of the partials. In *Proc. of the IEEE Int. Conf. on Acoustics, Speech, and Signal Processing (ICASSP)*, 2004.
- [17] T. Parsons. *Voice and Speech Processing*. Electrical and Computer Engineering. McGraw-Hill.

- [18] J. Makhoul. Linear prediction: A tutorial review. *Proceedings of the IEEE*, 63(4):561–580, Apr. 1975.
- [19] G. H. Golub, C. F. Van Loan. *Matrix Computations*. The John Hopkins University Press, 3rd edition, 1996.
- [20] P. Vary, R. Martin. *Digital Speech Transmission*. John Wiley and Sons Ltd., 2006.
- [21] B. S. Atal, M. R. Schroeder. Predictive coding of speech signals. In *Proc. the International Conference on Speech Communication and Processing*, pages 360–361, 1967.
- [22] A. S. Spanias. Speech coding: a tutorial review. In *Proceedings of the IEEE*, volume 82, no. 10, pages 1541–1582, 1994.
- [23] G. Kubit, B. S. Atal, W. B. Kleijn. Performance of noise excitation for unvoiced speech. In *Proc. the IEEE Workshop on Speech Coding for Telecommunications*, 1993.
- [24] International Telecommunication Union. *ITU-T Recommendation P.800: Methods for Subjective Determination of Transmission Quality*. ITU-T, Aug. 1996.
- [25] International Telecommunication Union. *ITU-T Recommendation P.800.1: Mean Opinion Score (MOS) terminology*. ITU-T, July 2006.
- [26] J. G. Proakis. *Digital Communications*. Mc-Graw-Hill, 4th edition, 2000.
- [27] P. Boersma, D. Weenink. Praat: doing Phonetics by Computer, Web Site. <http://www.fon.hum.uva.nl/praat/>.
- [28] W. E. Ryan, S. Lin. *Channel Codes: Classical and Modern*. Cambridge University Press, 1st edition, October 2009.
- [29] S. Lin, D. J. Costello Jr. *Error Control Coding: Fundamentals and Applications*. Prentice-Hall Inc., 2nd edition, 1983.
- [30] Q. Wang, V. K. Bhargava. An efficient maximum likelihood decoding algorithm for generalizedtail biting convolutional codes including quasicyclic codes. *IEEE Transactions on Communications*, 37(8):875–879, 1989.
- [31] The MathWorks Web Site. <http://www.mathworks.com>.
- [32] P. Hoher, E. Haas. Aeronautical Channel Modeling at VHF-Band. In *Proc. IEEE 50th Vehicular Technology Conference (VTC1999-Fall)*, pages 1961–1966, Amsterdam, Netherlands, Sep. 1999.
- [33] International Telecommunication Union. *ITU-T Recommendation P.862: Perceptual Evaluation of Speech Quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. ITU-T, Nov. 2005.

- [34] L. Sun. *Speech Quality Prediction for Voice over Internet Protocol Networks*. PhD thesis, University of Plymouth, Jan. 2004.
- [35] Nortel Networks. *VoIP Technologies: A Comprehensive Guide to Voice over Internet Protocol (VoIP)*. Nortel Press, 1st edition, March 2008.
- [36] W. Stallings. *Cryptography and Network Security*. Prentice Hall, 4th edition, 2006.
- [37] International Telecommunications Union, ITU-T Web Site. <http://www.itu.int/ITU-T>.
- [38] Speex free codec Web Site. <http://www.speex.org>.
- [39] iLBC codec Web Site. <http://ilbcfreeware.org>.
- [40] GSM page on ETSI Web Site. <http://www.etsi.org/WebSite/Technologies/gsm.aspx>.
- [41] Skype Web Site. <http://www.skype.com>.
- [42] D. Endler, M. Collier. *Hacking Exposed VoIP: Voice Over IP Security Secrets & Solutions*. McGraw-Hill, 1st edition, November 2006.
- [43] SESAR Joint Undertaking Web Site. <http://www.sesarju.eu>.
- [44] OMNeT++ Web Site. <http://www.omnetpp.org>.
- [45] M. Bohge, M. Renwanz. A realistic VoIP traffic generation and evaluation tool for OMNeT++. In *Proc. of 1st International Workshop on OMNeT++*, Marseille, France, March 2008.
- [46] A. S. Tanenbaum. *Computer Networks*. Prentice Hall, 4th edition, 2002.
- [47] International Telecommunication Union. *ITU-T Recommendation P.563: Single-ended method for objective speech quality assessment in narrow-band telephony applications*. ITU-T, May 2004.

Chapter 17

BROADBAND INTERNET ACCESS AND THE DIGITAL DIVIDE: FEDERAL ASSISTANCE PROGRAMS

Lennard G. Kruger^{1,a} and Angele A. Gilroy^{2,b}

¹Science and Technology Policy

²Telecommunications Policy

Summary

The “digital divide” is a term that has been used to characterize a gap between “information haves and have-nots,” or in other words, between those Americans who use or have access to telecommunications technologies (e.g., telephones, computers, the Internet) and those who do not. One important subset of the digital divide debate concerns high-speed Internet access and advanced telecommunications services, also known as *broadband*. Broadband is provided by a series of technologies (e.g., cable, telephone wire, fiber, satellite, wireless) that give users the ability to send and receive data at volumes and speeds far greater than traditional “dial-up” Internet access over telephone lines.

Broadband technologies are currently being deployed primarily by the private sector throughout the United States. While the numbers of new broadband subscribers continue to grow, studies and data suggest that the rate of broadband deployment in urban and high income areas are outpacing deployment in rural and low-income areas. Some policymakers, believing that disparities in broadband access across American society could have adverse economic and social consequences on those left behind, assert that the federal government should play a more active role to avoid a “digital divide” in broadband access. One approach is for the federal government to provide financial assistance to support broadband deployment in unserved and underserved areas.

Economic stimulus legislation enacted by the 111th Congress includes provisions that provides federal financial assistance for broadband deployment. On February 17, 2009, President Obama signed P.L. 111-5, the American Recovery and Reinvestment Act (ARRA). The ARRA provides a total of **\$7.2 billion** for broadband, consisting of \$4.7 billion to NTIA/DOC for a newly established Broadband Technology Opportunities Program and \$2.5 billion to existing RUS/USDA broadband programs.

Meanwhile, it is expected that the Obama Administration will ultimately develop a national broadband policy or strategy that will seek to reduce or eliminate the “digital divide”

^a E-mail address: lkruger@crs.loc.gov, 7-7070.

^b E-mail address: agilroy@crs.loc.gov, 7-7778.

with respect to broadband. It is likely that elements of a national broadband policy, in tandem with broadband investment measures in the American Recovery and Reinvestment Act, will significantly shape and expand federal policies and programs to promote broadband deployment and adoption. A key issue is how to strike a balance between providing federal assistance for unserved and underserved areas where the private sector may not be providing acceptable levels of broadband service, while at the same time minimizing any deleterious effects that government intervention in the marketplace may have on competition and private sector investment.

This report will be updated as events warrant.

Introduction

The “digital divide” is a term used to describe a perceived gap between perceived “information haves and have-nots,” or in other words, between those Americans who use or have access to telecommunications technologies (e.g., telephones, computers, the Internet) and those who do not.¹ Whether or not individuals or communities fall into the “information haves” category depends on a number of factors, ranging from the presence of computers in the home, to training and education, to the availability of affordable Internet access.

Broadband technologies are currently being deployed primarily by the private sector throughout the United States. While the numbers of new broadband subscribers continue to grow, studies and data suggest that the rate of broadband deployment in urban and high income areas are outpacing deployment in rural and low-income areas. Some policymakers, believing that disparities in broadband access across American society could have adverse economic and social consequences on those left behind, assert that the federal government should play a more active role to avoid a “digital divide” in broadband access. One approach—adopted in the American Recovery and Reinvestment Act of 2009 (P.L. 111-5)—is for the federal government to provide financial assistance, primarily grants, to support broadband deployment in unserved and underserved areas.

Status of Broadband Deployment in the United States

Prior to the late 1990s, American homes accessed the Internet at maximum speeds of 56 kilobits per second by dialing up an Internet Service Provider (such as AOL) over the same copper telephone line used for traditional voice service. A relatively small number of businesses and institutions used broadband or high speed connections through the installation of special “dedicated lines” typically provided by their local telephone company. Starting in the late 1990s, cable television companies began offering cable modem broadband service to homes and businesses. This was accompanied by telephone companies beginning to offer DSL service (broadband over existing copper telephone wireline). Growth has been steep, rising from 2.8 million high speed lines reported as of December 1999, to 121.2 million lines as of December 31, 2007. Of the 121.2 million high speed lines reported by the FCC, 74.0 million serve residential users.² Since the deployment of residential broadband in the United

¹ The term “digital divide” can also refer to international disparities in access to information technology. This report focuses on domestic issues only.

² FCC, *High-Speed Services for Internet Access: Status as of December 31, 2007*, January 2009. Available at http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-287962A1.pdf.

States, the primary residential broadband technologies deployed continue to be cable modem and DSL. A distinction is often made between “current generation” and “next generation” broadband (commonly referred to as next generation networks or NGN). “Current generation” typically refers to currently deployed cable, DSL, and many wireless systems, while “next generation” refers to dramatically faster download and upload speeds offered by fiber technologies and also potentially by future generations of cable, DSL, and wireless technologies.³ In general, the greater the download and upload speeds offered by a broadband connection, the more sophisticated (and potentially valuable) the application that is enabled.

December 2008 survey data from the Pew Internet and American Life Project found that 57% of Americans have broadband at home.⁴ It is estimated that less than 10% of U.S. households have no access to any broadband provider whatsoever (not including satellite).⁵ While the broadband *adoption* or *penetration* rate stands at close to 60% of U.S. households, broadband *availability* is much higher, at more than 90% of households. Thus, approximately 30% of households have access to some type of terrestrial (non-satellite) broadband service, but do not choose to subscribe. According to the FCC, possible reasons for the gap between broadband availability and subscribership include the lack of computers in some homes, price of broadband service, lack of content, and the availability of broadband at work.⁶ According to Pew, non-broadband users tend to be older, have lower incomes, have trouble using technology, and may not see the relevance of using the Internet to their lives. Between 2007 and 2008, low income Americans (under \$20,000 annual income) and African Americans showed no significant growth in home broadband adoption after strong growth in previous years.⁷ Pew also found that about one-third of adults without broadband cite price and availability as the reasons why they don’t have broadband in their homes, while two-thirds cite reasons such as usability and relevance.⁸

³ Initially, and for many years following, the FCC defined broadband (or more specifically “high-speed lines”) as over 200 kilobits per second (kbps) in at least one direction, which was roughly four times the speed of conventional dialup Internet access. In recent years, the 200 kbps threshold was considered too low, and on March 19, 2008, the FCC adopted a report and order (FCC 08-89) establishing new categories of broadband speed tiers for data collection purposes. Specifically, 200 kbps to 768 kbps will be considered “first generation,” 768 kbps to 1.5 Mbps as “basic broadband tier 1,” and increasingly higher speed tiers as broadband tiers 2 through 7 (tier seven is greater than or equal to 100 Mbps in any one direction). Tiers can change as technology advances.

⁴ Horrigan, John, Pew Internet & American Life Project, “Barriers to Broadband Adoption—The User Perspective,” December 19, 2008, available at http://otrans.3cdn.net/fe2b6b302960dbe0d7_bqm6ib242.pdf.

⁵ S. Derek Turner, Free Press, *Down Payment on Our Digital Future*, December 2008, p. 8.

⁶ Federal Communications Commission, *Fourth Report to Congress*, “Availability of Advanced Telecommunications Capability in the United States,” GN Docket No. 04-54, FCC 04-208, September 9, 2004, p. 38. Available at http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-04-208A1.pdf.

⁷ “Barriers to Broadband Adoption—The User Perspective,” p. 1.

⁸ Horrigan, John, Pew Internet & American Life Project, “Obama’s Online Opportunities II: If You Build It Will They Log On?” January 21, 2009, available at http://www.pewinternet.org/pdfs/PIP_Broadband%20Barriers.pdf.

Broadband in Rural and Underserved Areas⁹

While the number of new broadband subscribers continues to grow, the rate of broadband deployment in urban and high income areas appears to be outpacing deployment in rural and low-income areas. While there are many examples of rural communities with state of the art telecommunications facilities,¹⁰ recent surveys and studies have indicated that, in general, rural areas tend to lag behind urban and suburban areas in broadband deployment. Data (2008) from the Pew Internet & American Life Project indicate that while broadband adoption is growing in urban, suburban, and rural areas, broadband users make up larger percentages of urban and suburban users than rural users. Pew found that the percentage of all U.S. adults with broadband at home is 60% for suburban areas, 57% for urban areas, and 38% for rural areas.¹¹

Similarly, according to the latest FCC data on the deployment of high-speed Internet connections (released January 2009), high-speed subscribers were reported in 99% of the most densely populated zip codes, as opposed to 90% of zip codes with the lowest population densities. For zip codes ranked by median family income, high-speed subscribers were reported present in 99% of the top one-tenth of zip codes, as compared to 92% of the bottom one-tenth of zip codes.¹²

The comparatively lower population density of rural areas is likely the major reason why broadband is less deployed than in more highly populated suburban and urban areas. Particularly for wireline broadband technologies—such as cable modem and DSL—the greater the geographical distances among customers, the larger the cost to serve those customers. Thus, there is often less incentive for companies to invest in broadband in rural areas than, for example, in an urban area where there is more demand (more customers with perhaps higher incomes) and less cost to wire the market area.¹³

Some policymakers believe that disparities in broadband access across American society could have adverse consequences on those left behind, and that advanced telecommunications applications critical for businesses and consumers to engage in e-commerce are increasingly dependent on high speed broadband connections to the Internet. Thus, some say, communities and individuals without access to broadband could be at risk to the extent that e-commerce becomes a critical factor in determining future economic development and prosperity. A February 2006 study done by the Massachusetts Institute of Technology for the Economic Development Administration of the Department of Commerce marked the first attempt to quantitatively measure the impact of broadband on economic growth. The study found that “between 1998 and 2002, communities in which mass-market broadband was available by

⁹ For more information on rural broadband and broadband programs at the Rural Utilities Service, see CRS Report RL33816, *Broadband Loan and Grant Programs in the USDA's Rural Utilities Service*, by Lennard G. Kruger.

¹⁰ See for example: National Exchange Carrier Association (NECA), Trends 2006: Making Progress With Broadband, 2006, 26 p. Available at http://www.neca.org/media/trends_brochure_website.pdf.

¹¹ Horrigan, John B., Pew Internet & American Life Project, *Home Broadband Adoption 2008*, July 2008, p. 3. Available at http://www.pewinternet.org/pdfs/PIP_Broadband_2008.pdf.

¹² FCC, *High-Speed Services for Internet Access: Status as of December 31, 2007*, p. 4.

¹³ The terrain of rural areas can also be a hindrance to broadband deployment because it is more expensive to deploy broadband technologies in a mountainous or heavily forested area. An additional added cost factor for remote areas can be the expense of “backhaul” (e.g., the “middle mile”) which refers to the installation of a dedicated line which transmits a signal to and from an Internet backbone which is typically located in or near an urban area.

December 1999 experienced more rapid growth in employment, the number of businesses overall, and businesses in IT-intensive sectors, relative to comparable communities without broadband at that time.”¹⁴

Subsequently, a June 2007 report from the Brookings Institution found that for every one percentage point increase in broadband penetration in a state, employment is projected to increase by 0.2 to 0.3% per year. For the entire U.S. private non-farm economy, the study projected an increase of about 300,000 jobs.¹⁵

Some also argue that broadband is an important contributor to U.S. future economic strength with respect to the rest of the world. According to the International Telecommunications Union, the U.S. ranks 24th worldwide in broadband penetration (subscriptions per 100 inhabitants in 2007).¹⁶ Data from the Organization for Economic Cooperation and Development (OECD) found the U.S. ranking 15th among OECD nations in broadband access per 100 inhabitants as of June 2008.¹⁷ By contrast, in 2001 an OECD study found the U.S. ranking 4th in broadband subscribership per 100 inhabitants (after Korea, Sweden, and Canada).¹⁸ While many argue that the U.S. declining performance in international broadband rankings is a cause for concern,¹⁹ others maintain that the OECD and ITU data undercount U.S. broadband deployment,²⁰ and that cross-country broadband deployment comparisons are not necessarily meaningful and inherently problematic.²¹ Finally, an issue related to international broadband rankings is the extent to which broadband speeds and prices differ between the U.S. and the rest of the world.²²

¹⁴ Gillett, Sharon E., Massachusetts Institute of Technology, *Measuring Broadband's Economic Impact*, report prepared for the Economic Development Administration, U.S. Department of Commerce, February 28, 2006 p. 4.

¹⁵ Crandall, Robert, William Lehr, and Robert Litan, *The Effects of Broadband Deployment on Output and Employment: A Cross-sectional Analysis of U.S. Data*, June 2007, 20 pp. Available at <http://www3.brookings.edu/views/papers/crandall/200706litan.pdf>.

¹⁶ International Telecommunications Union, *Economies by broadband penetration*, 2007. Available at http://www.itu.int/ITU-D/ict/statistics/at_glance/top20_broad_2007.html.

¹⁷ OECD, *OECD Broadband Statistics*, June 2008. Available at <http://www.oecd.org/sti/ict/broadband>.

¹⁸ OECD, Directorate for Science, Technology and Industry, *The Development of Broadband Access in OECD Countries*, October 29, 2001, 63 pp. For a comparison of government broadband policies, also see OECD, Directorate for Science, Technology and Industry, *Broadband Infrastructure Deployment: The Role of Government Assistance*, May 22, 2002, 42 pp.

¹⁹ See Turner, Derek S., Free Press, *Broadband Reality Check II: The Truth Behind America's Digital Divide*, August 2006, pp 8-11. Available at <http://www.freepress.net/files/bbrc2-final.pdf>; and Turner, Derek S., Free Press, ‘Shooting the Messenger’ Myth vs. Reality: U.S. Broadband Policy and International Broadband Rankings, July 2007, 25 pp., available at http://www.freepress.net/files/shooting_the_messenger.pdf.

²⁰ National Telecommunications and Information Administration, *Fact Sheet: United States Maintains Information and Communication Technology (ICT) Leadership and Economic Strength*, at http://www.ntia.doc.gov/ntia/home/press/2007/ICTleader_042407.html.

²¹ See Wallsten, Scott, Progress and Freedom Foundation, *Towards Effective U.S. Broadband Policies*, May 2007, 19 pp. Available at <http://www.pff.org/issues-pubs/pops/pop14.7usbroadbandpolicy.pdf>. Also see Ford, George, Phoenix Center, *The Broadband Performance Index: What Really Drives Broadband Adoption Across the OECD?*, Phoenix Center Policy Paper Number 33, May 2008, 27 pp; available at <http://www.phoenix-center.org/pcpp/PCPP33Final.pdf>.

²² See price and services and speed data on OECD Broadband Portal, available at <http://www.oecd.org/sti/ict/broadband>; Turner, Derek S., Free Press, *Broadband Reality Check II: The Truth Behind America's Digital Divide*, August 2006, pp 5-9; Kende, Michael, Analysis Consulting Limited, *Survey of International Broadband Offerings*, October 4, 2006, 12 p, available at <http://www.analysys.com/pdfs/BroadbandPerformanceSurvey.pdf>; and Atkinson, Robert D., The International Technology and Innovation Foundation, *Explaining International Broadband Leadership*, May 2008, 108 p, available at <http://www.itif.org/files/ExplainingBBLedership.pdf>.

Is Broadband Deployment Data Adequate?

Obtaining an accurate snapshot of the status of broadband deployment is problematic. Anecdotes abound of rural and low-income areas which do not have adequate Internet access, as well as those which are receiving access to high-speed, state-of-the-art connections. Rapidly evolving technologies, the constant flux of the telecommunications industry, the uncertainty of consumer wants and needs, and the sheer diversity and size of the nation's economy and geography make the status of broadband deployment very difficult to characterize. The FCC periodically collects broadband deployment data from the private sector via "FCC Form 477"—a standardized information gathering survey. Statistics derived from the Form 477 survey are published every six months. Additionally, data from Form 477 are used as the basis of the FCC's (to date) five broadband deployment reports.

The FCC is working to refine the data used in future Reports in order to provide an increasingly accurate portrayal. In its March 17, 2004 Notice of Inquiry for the *Fourth Report*, the FCC sought comments on specific proposals to improve the FCC Form 477 data gathering program.²³ On November 9, 2004, the FCC voted to expand its data collection program by requiring reports from all facilities based carriers regardless of size in order to better track rural and underserved markets, by requiring broadband providers to provide more information on the speed and nature of their service, and by establishing broadband-over-power line as a separate category in order to track its development and deployment. The FCC Form 477 data gathering program was extended for five years beyond its March 2005 expiration date.²⁴

The Government Accountability Office (GAO) has cited concerns about the FCC's zip-code level data. Of particular concern is that the FCC will report broadband service in a zip code even if a company reports service to only one subscriber, which in turn can lead to some observers overstating broadband deployment. According to GAO, "the data may not provide a highly accurate depiction of local deployment of broadband infrastructures for residential service, especially in rural areas." The FCC has acknowledged the limitations in its zip code level data.²⁵

On April 16, 2007, the FCC announced a Notice of Proposed Rulemaking which sought comment on a number of broadband data collection issues, including how to develop a more accurate picture of broadband deployment; gathering information on price, other factors determining consumer uptake of broadband, and international comparisons; how to improve data on wireless broadband; how to collect information on subscribership to voice over Internet Protocol service (VoIP); and whether to modify collection of speed tier information.²⁶

²³ Federal Communications Commission, *Notice of Inquiry*, "Concerning the Deployment of Advanced Telecommunications Capability to All Americans in a Reasonable and Timely Fashion, and possible Steps to Accelerate Such Deployment Pursuant to Section 706 of the Telecommunications Act of 1996," FCC 04-55, March 17, 2004, p. 6.

²⁴ FCC News Release, *FCC Improves Data Collection to Monitor Nationwide Broadband Rollout*, November 9, 2004. Available at http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-254115A1.pdf.

²⁵ U.S. Government Accountability Office, *Broadband Deployment is Extensive throughout the United States, but It Is Difficult to Assess the Extent of Deployment Gaps in Rural Areas*, GAO-06-426, May 2006, p. 3.

²⁶ Federal Communications Commission, *Notice Proposed Rulemaking*, "Development of Nationwide Broadband Data to Evaluate Reasonable and Timely Deployment of Advanced Services to All Americans, Improvement of Wireless Broadband Subscribership Data, and Development of Data on Interconnected Voice Over Internet Protocol (VoIP) Subscribership," WC Docket No. 07-38, FCC 07-17, released April 16, 2007, 56 pp.

On March 19, 2008, the FCC adopted an Order that substantially expands its broadband data collection capability. Specifically, the Order expands the number of broadband reporting speed tiers to capture more information about upload and download speeds offered in the marketplace, requires broadband providers to report numbers of broadband subscribers by census tract, and improves the accuracy of information collected on mobile wireless broadband deployment. Additionally, in a Further Notice of Proposed Rulemaking, the FCC sought comment on broadband service pricing and availability.²⁷

During the 110th Congress, state initiatives to collect broadband deployment data in order to promote broadband in underserved areas were viewed as a possible model for governmental efforts to encourage broadband. In particular, an initiative in the Commonwealth of Kentucky—called ConnectKentucky—has developed detailed broadband inventory mapping which identifies local communities that lack adequate broadband service. Kentucky is using this data to promote public-private partnerships in order to reach a goal of universal broadband coverage in the state.²⁸ Other states are pursuing or considering similar approaches.

The 110th Congress explored ways to support or implement the types of broadband mapping and data collection efforts demonstrated by ConnectKentucky. The Broadband Data Improvement Act was enacted by the 110th Congress and became P.L. 110-385 on October 10, 2008. The law requires the FCC to collect demographic information on unserved areas, data comparing broadband service with 75 communities in at least 25 nations abroad, and data on consumer use of broadband. The act also directs the Census Bureau to collect broadband data, the Government Accountability Office to study broadband data metrics and standards, and the Department of Commerce to provide grants supporting state broadband initiatives.

P.L. 111-5, the American Recovery and Reinvestment Act, provides NTIA with an appropriation of \$350 million to implement the Broadband Data Improvement Act and to develop and maintain a national broadband inventory map, which shall be made accessible to the public no later than two years after enactment.

Broadband and the Federal Role

The Telecommunications Act of 1996 (P.L. 104-104) addressed the issue of whether the federal government should intervene to prevent a “digital divide” in broadband access. Section 706 requires the FCC to determine whether “advanced telecommunications capability [i.e., broadband or high-speed access] is being deployed to all Americans in a reasonable and timely fashion.” If this is not the case, the act directs the FCC to “take immediate action to accelerate deployment of such capability by removing barriers to infrastructure investment and by promoting competition in the telecommunications market.”

Since 1999, the FCC has issued and adopted five reports pursuant to Section 706. All five reports formally concluded that the deployment of advanced telecommunications capability to

²⁷ FCC, News Release, “FCC Expands, Improves Broadband Data Collection,” March 19, 2008. Available at http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-280909A1.pdf.

²⁸ Testimony of Brian Mefford, President and CEO, Connected Nation, Inc., before the Senate Committee on Commerce, Science and Transportation, April 24, 2007. Available at http://commerce.senate.gov/public/_files/DC_Committeetestimony_04_23_07.pdf.

all Americans is reasonable and timely. The fifth and most recent 706 report was adopted on March 19, 2008, and released on June 12, 2008.²⁹ Two FCC Commissioners (Michael Copps and Jonathan Adelstein) have repeatedly dissented from the reports' conclusions that broadband deployment is reasonable and timely, arguing that the relatively poor world ranking of United States broadband penetration indicates that deployment is insufficient, that the FCC's definition of broadband was outdated and not comparable to the much higher speeds available to consumers in other countries, that the use of zip code data (measuring the presence of at least one broadband subscriber within a zip code area) did not sufficiently characterize the availability of broadband across geographic areas, and that broadband deployment is impeded by the lack of a comprehensive national broadband policy.³⁰

Bush Administration

The Bush Administration pursued a broadband policy that emphasized deregulation, non-intervention by government in the marketplace, and general tax policies intended to foster overall economic growth. On March 26, 2004, President Bush endorsed a goal of "universal broadband access by 2007," and on April 26, 2004, announced a broadband initiative which included promoting legislation which would permanently prohibit all broadband taxes, making spectrum available for wireless broadband and creating technical standards for broadband over power lines, and simplifying rights-of-way processes on federal lands for broadband providers.³¹ Subsequently, on January 31, 2008, NTIA released a report, entitled, *Networked Nation: Broadband in America, 2007* which characterized the Bush Administration's broadband initiative as follows:

From its first days, the Administration has implemented a comprehensive and integrated package of technology, regulatory, and fiscal policies designed to lower barriers and create an environment in which broadband innovation and competition can flourish.³²

The Bush Administration broadband policy embraced the view that a minimum of government intervention would create an economic climate favorable to private sector investment in the broadband market. According to NTIA, the report showed "that the Administration's technology, regulatory, and fiscal policies have stimulated innovation and competition, and encouraged investment in the U.S. broadband market contributing to significantly increased accessibility of broadband services."³³

²⁹ Federal Communications Commission, *Fifth Report*, "In the Matter of Inquiry Concerning the Deployment of Advanced Telecommunications Capability to All Americans in a Reasonable and Timely Fashion, and Possible Steps to Accelerate Such Deployment Pursuant to Section 706 of the Telecommunications Act of 1996," GN Docket No. 07- 45, FCC 08-88, Adopted March 19, 2008, Released June 12, 2008. 76 pp. Available at http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-08-88A1.pdf.

³⁰ *Ibid.*, pp. 5, 7.

³¹ See White House, A New Generation of American Innovation, April 2004. Available at http://www.whitehouse.gov/infocus/technology/economic_policy200404/innovation.pdf.

³² U.S. Department of Commerce, National Telecommunications and Information Administration, Networked Nation: Broadband in America 2007, January 2008, p. I. Available at <http://www.ntia.doc.gov/reports/2008/NetworkedNationBroadbandinAmerica2007.pdf>.

³³ NTIA, *Press Release*, "Gutierrez Hails Dramatic U.S. Broadband Growth," January 31, 2008. Available at http://www.ntia.doc.gov/ntiahome/press/2008/NetworkedNation_013108.html.

During the 110th Congress, some policymakers disagreed with the Bush Administration's assessment and asserted that the federal government should play a more active role to avoid a "digital divide" in broadband access. Bills were introduced seeking to provide federal financial assistance for broadband deployment in the form of grants, loans, subsidies, and/or tax credits.

Obama Administration

It is expected that the Obama Administration will ultimately develop a national broadband policy or strategy that will seek to reduce or eliminate the "digital divide" with respect to broadband. One of the key elements of the Obama transition's technology agenda is to "deploy next-generation broadband," and specifically:

Work towards true broadband in every community in America through a combination of reform of the Universal Service Fund, better use of the nation's wireless spectrum, promotion of next-generation facilities, technologies and applications, and new tax and loan incentives. America should lead the world in broadband penetration and Internet access.³⁴

The Obama campaign released a policy blueprint for technology and innovation that includes policy proposals intended to result in full broadband penetration and deployment of next-generation broadband. Specifically, policy proposals include:

- redefining broadband at speeds "demanded by 21st century business and communications;"
- reforming universal service to support affordable broadband specifically focusing on unserved communities;
- creating incentives for more efficient use of government spectrum and new standards for commercial spectrum to bring affordable broadband to rural communities;
- ensuring that schools, libraries and hospitals have access to next-generation networks and that adequate training and resources are available to enable these institutions to take full advantage of broadband connectivity; and
- encouraging public/private partnerships at the local level that result in broadband to unserved communities.³⁵

It is likely that these and other potential elements of a national broadband policy, in tandem with broadband investment measures and development of a national broadband strategy by the FCC as directed by the American Recovery and Reinvestment Act of 2009, will significantly shape and expand federal policies and programs intended to promote broadband deployment and adoption.

Current Federal Broadband Programs

Aside from the broadband programs newly established by the American Recovery and Reinvestment Act of 2009 (P.L. 111-5),³⁶ the Rural Broadband Access Loan and Loan

³⁴ Office of the President-Elect, *Technology Agenda*, available at http://change.gov/agenda/technology_agenda.

³⁵ Barack Obama, *Connecting and Empowering All Americans Through Technology and Innovation*, 2008, available at <http://obama.3cdn.net/780e0e91ccb6cdbc6e6udymvin7.pdf>.

Guarantee Program and the Community Connect Broadband Grants, both at the Rural Utilities Service of the U.S. Department of Agriculture, are currently the only federal programs *exclusively* dedicated to deploying broadband infrastructure. However, there exist other federal programs that provide financial assistance for various aspects of telecommunications development. The major vehicle for funding telecommunications development, particularly in rural and low-income areas, is the Universal Service Fund (USF). While the USF's High Cost Program does not *explicitly* fund broadband infrastructure, subsidies are used, in many cases, to upgrade existing telephone networks so that they are capable of delivering high-speed services. Additionally, subsidies provided by USF's Schools and Libraries Program and Rural Health Care Program are used for a variety of telecommunications services, including broadband access.

Table 1 (at the end of this report) shows selected federal domestic assistance programs throughout the federal government that can be associated with telecommunications development. Many (if not most) of these programs can be related, if not necessarily to the deployment of broadband technologies in particular, then to telecommunications and the "digital divide" issue generally.

Table 2 (also at the end of this report) presents selected federal programs that have provided financial assistance for broadband. These programs are broken down into three categories: first, programs that fund access to telecommunications services in unserved or underserved areas; second, general economic development programs that have funded broadband-related projects; and third, applications-specific programs which will typically fund some aspect of broadband access as a means towards supporting a particular application, such as distance learning or telemedicine.

Rural Utilities Service Programs

The Rural Electrification Administration (REA), subsequently renamed the Rural Utilities Service (RUS), was established by the Roosevelt Administration in 1935. Initially, it was established to provide credit assistance for the development of rural electric systems. In 1949, the mission of REA was expanded to include rural telephone providers. Congress further amended the Rural Electrification Act in 1971 to establish within REA a Rural Telephone Account and the Rural Telephone Bank (RTB). Rural Telephone Loans and Loan Guarantees provide long-term direct and guaranteed loans for telephone lines, facilities, or systems to furnish and improve telecommunications service in rural areas. The RTB—liquidated in FY2006—was a public-private partnership intended to provide additional sources of capital that would supplement loans made directly by RUS. Another program, the Distance Learning and Telemedicine Program, specifically addresses health care and education needs of rural America.

RUS implements two programs specifically targeted at providing assistance for broadband deployment in rural areas: the Rural Broadband Access Loan and Loan Guarantee Program and Community Connect Broadband Grants. The 110th Congress reauthorized and reformed the Rural Broadband Access Loan and Loan Guarantee program as part of the 2008 farm bill (P.L. 110- 234). For further information on rural broadband and the RUS broadband

³⁶ See CRS Report R40436, *Broadband Infrastructure Programs in the American Recovery and Reinvestment Act*, by Lennard G. Kruger.

programs, see CRS Report RL33816, *Broadband Loan and Grant Programs in the USDA's Rural Utilities Service*, by Lennard G. Kruger.

The Universal Service Concept and the FCC³⁷

Since its creation in 1934 the Federal Communications Commission (FCC) has been tasked with "... mak[ing] available, so far as possible, to all the people of the United States, ... a rapid, efficient, Nation-wide, and world-wide wire and radio communications service with adequate facilities at reasonable charges.... "³⁸ This mandate led to the development of what has come to be known as the universal service concept.

The universal service concept, as originally designed, called for the establishment of policies to ensure that telecommunications services are available to all Americans, including those in rural, insular and high cost areas, by ensuring that rates remain affordable. Over the years this concept fostered the development of various FCC policies and programs to meet this goal. The FCC offers universal service support through a number of direct mechanisms that target both providers of and subscribers to telecommunications services.³⁹

The development of the federal universal service high cost fund is an example of provider-targeted support. Under the high cost fund, eligible telecommunications carriers, usually those serving rural, insular and high cost areas, are able to obtain funds to help offset the higher than average costs of providing telephone service.⁴⁰ This mechanism has been particularly important to rural America where the lack of subscriber density leads to significant costs. FCC universal service policies have also been expanded to target individual users. Such federal programs include two income-based programs, Link Up and Lifeline, established in the mid-1980s to assist economically needy individuals. The Link Up program assists low-income subscribers pay the costs associated with the initiation of telephone service and the Lifeline program assists low-income subscribers pay the recurring monthly service charges. Funding to assist carriers providing service to individuals with speech and/or hearing disabilities is also provided through the Telecommunications Relay Service Fund. Effective January 1, 1998, schools, libraries, and rural health care providers also qualified for universal service support.

Universal Service and the Telecommunications Act of 1996

Passage of the Telecommunications Act of 1996 (P.L. 104-104) codified the long-standing commitment by U.S. policymakers to ensure universal service in the provision of telecommunications services.

³⁷ The section on universal service was prepared by Angele Gilroy, Specialist in Telecommunications, Resources, Science and Industry Division. For more information on universal service, see CRS Report RL33979, *Universal Service Fund: Background and Options for Reform*, by Angele A. Gilroy.

³⁸ Communications Act of 1934, As Amended, Title I sec.1 [47 U.S.C. 151].

³⁹ Many states participate in or have programs that mirror FCC universal service mechanisms to help promote universal service goals within their states.

⁴⁰ Additional FCC policies such as rate averaging and pooling have also been implemented to assist high cost carriers.

The Schools and Libraries, and Rural Health Care Programs

Congress, through the 1996 Act, not only codified, but also expanded the concept of universal service to include, among other principles, that elementary and secondary schools and classrooms, libraries, and rural health care providers have access to telecommunications services for specific purposes at discounted rates. (See Sections 254(b)(6) and 254(h) of the 1996 Telecommunications Act, 47 U.S.C. 254.)

1. The Schools and Libraries Program. Under universal service provisions contained in the 1996 Act, elementary and secondary schools and classrooms and libraries are designated as beneficiaries of universal service discounts. Universal service principles detailed in Section 254(b)(6) state that “Elementary and secondary schools and classrooms ... and libraries should have access to advanced telecommunications services....” The act further requires in Section 254(h)(1)(B) that services within the definition of universal service be provided to elementary and secondary schools and libraries for education purposes at discounts, that is at “rates less than the amounts charged for similar services to other parties.”

The FCC established the Schools and Libraries Division within the Universal Service Administrative Company (USAC) to administer the schools and libraries or “E (education)-rate” program to comply with these provisions. Under this program, eligible schools and libraries receive discounts ranging from 20 to 90 percent for telecommunications services depending on the poverty level of the school’s (or school district’s) population and its location in a high cost telecommunications area. Three categories of services are eligible for discounts: internal connections (e.g., wiring, routers and servers); Internet access; and telecommunications and dedicated services, with the third category receiving funding priority. According to data released by program administrators, \$21.3 billion in funding has been committed over the first ten years of the program with funding released to all states, the District of Columbia and all territories. Funding commitments for funding Year 2008 (July 1, 2008 - June 30, 2009), the eleventh and current year of the program, totaled \$2.2 billion as of March 10, 2009.⁴¹

2. The Rural Health Care Program. Section 254(h) of the 1996 Act requires that public and nonprofit rural health care providers have access to telecommunications services necessary for the provision of health care services at rates comparable to those paid for similar services in urban areas. Subsection 254(h)(1) further specifies that “to the extent technically feasible and economically reasonable” health care providers should have access to advanced telecommunications and information services. The FCC established the Rural Health Care Division (RHCD) within the USAC to administer the universal support program to comply with these provisions. Under FCC established rules only public or non-profit health care providers are eligible to receive funding. Eligible health care providers, with the exception of those requesting only access to the Internet, must also be located in a rural area. The funding ceiling, or cap, for this support was established at \$400 million annually. The

⁴¹ For additional information on this program, including funding commitments, see the E-rate website: <http://www.universalservice.org/sl/>.

funding level for Year One of the program (January 1998 - June 30, 1999) was set at \$100 million. Due to less than anticipated demand, the FCC established a \$12 million funding level for the second year (July 1, 1999 to June 30, 2000) of the program but has since returned to a \$400 million yearly cap. As of March 17, 2009, covering the first 11 years of the program, a total of \$284.7 million has been committed to 4,167 rural health care providers. The primary use of the funding is to provide reduced rates for telecommunications and information services necessary for the provision of health care.⁴²

The Telecommunications Development Fund

Section 714 of the 1996 Act created the Telecommunications Development Fund (TDF). The TDF is a private, non-governmental, venture capital corporation currently overseen by a five-member board of directors and fund management. The TDF focuses on seed, early stage, and select later stage investments in communications and has \$90 million under management in two funds. Fund I is no longer making new investments. Fund II remains active and currently has 13 companies in its investment portfolio. Funding is largely derived from the interest earned from the upfront payments bidders submit to participate in FCC auctions. The TDF also provides entrepreneur education, training, management and technical assistance in underserved rural and urban communities through the TDF Foundation.⁴³

Universal Service and Broadband

One of the policy debates surrounding universal service is whether access to advanced telecommunications services (i.e. broadband) should be incorporated into universal service objectives. The term universal service, when applied to telecommunications, refers to the ability to make available a basket of telecommunications services to the public, across the nation, at a reasonable price. As directed in the 1996 Telecommunications Act [Section 254(c)] a federal-state Joint Board was tasked with defining the services which should be included in the basket of services to be eligible for federal universal service support; in effect using and defining the term “universal service” for the first time. The Joint Board’s recommendation, which was subsequently adopted by the FCC in May 1997, included the following in its universal service package: voice grade access to and some usage of the public switched network; single line service; dual tone signaling; access to directory assistance; emergency service such as 911; operator services; access and interexchange (long distance) service.

Some policy makers expressed concern that the FCC-adopted definition is too limited and does not take into consideration the importance and growing acceptance of advanced services such as broadband and Internet access. They point to a number of provisions contained in the Universal Service section of the 1996 Act to support their claim. Universal service principles contained in Section 254(b)(2) state that “Access to advanced telecommunications services should be provided to all regions of the Nation.” The subsequent principle (b)(3) calls for consumers in all regions of the nation including “low-income” and those in “rural, insular,

⁴² For additional information on this program, including funding commitments, see the RHCD website: <http://www.universalservice.org/rhc/>.

⁴³ For additional information on the TDF fund and TDF Foundation see the TDF website at <http://www.tdfund.com>.

and high cost areas” to have access to telecommunications and information services including “advanced services” at a comparable level and a comparable rate charged for similar services in urban areas. Such provisions, they state, dictate that the FCC expand its universal service definition.

Others caution that a more modest approach is appropriate given the “universal mandate” associated with this definition and the uncertainty and costs associated with mandating nationwide deployment of such advanced services as a universal service policy goal. Furthermore they state the 1996 Act does take into consideration the changing nature of the telecommunications sector and allows for the universal service definition to be modified if future conditions warrant. Section 254(c) of the act states that “universal service is an evolving level of telecommunications services” and the FCC is tasked with “periodically” reevaluating this definition “taking into account advances in telecommunications and information technologies and services.” Furthermore, the Joint Board is given specific authority to recommend “from time to time” to the FCC modification in the definition of the services to be included for federal universal service support. The Joint Board, on November 19, 2007, concluded such an inquiry and recommended that the FCC change the mix of services eligible for universal service support. The Joint Board recommended, among other things, that “the universal availability of broadband Internet services” be included in the nation’s communications goals and hence be supported by federal universal service funds.⁴⁴ In response to the Joint Board recommendation, the FCC, on January 29, 2008, released three notices of proposed rulemaking dealing with specific aspects of universal service, including an examination of the scope of the definition. The FCC is still examining proposals for universal service reform, including expanding the program to include broadband, but has not taken action.

Legislation in the 110th Congress

In the 110th Congress, legislation was introduced that would provide financial assistance for broadband deployment. Of particular note is the reauthorization of the Rural Utilities Service (RUS) broadband loan program, which was enacted as part of the 2008 farm bill (P.L. 110-234). In addition to reauthorizing and reforming the RUS broadband loan program, P.L. 110-234 contains provisions establishing a National Center for Rural Telecommunications Assessment and requiring the FCC and RUS to formulate a comprehensive rural broadband strategy.

The Broadband Data Improvement Act (P.L. 110-385) was enacted by the 110th Congress and requires the FCC to collect demographic information on unserved areas, data comparing broadband service with 75 communities in at least 25 nations abroad, and data on consumer use of broadband. The act also directs the Census Bureau to collect broadband data, the Government Accountability Office to study broadband data metrics and standards, and the Department of Commerce to provide grants supporting state broadband initiatives.

⁴⁴ The Joint Board recommended that the definition of those services that qualify for universal service support be expanded and that the nation’s communications goals include the universal availability of: mobility services (i.e., wireless voice); broadband Internet services; and voice services at affordable and comparable rates for all rural and non-rural areas. For a copy of this recommendation see http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-07J4A1.pdf.

Meanwhile, the America COMPETES Act (H.R. 2272) was enacted (P.L. 110-69) and contains a provision authorizing the National Science Foundation (NSF) to provide grants for basic research in advanced information and communications technologies. Areas of research include affordable broadband access, including wireless technologies. P.L. 110-69 also directs NSF to develop a plan that describes the current status of broadband access for scientific research purposes.

The following is a listing of broadband related bills enacted in the 110th Congress.

P.L. 110-69 (H.R. 2272)

America COMPETES Act. Authorizes the National Science Foundation (NSF) to provide grants for basic research in advanced information and communications technologies. Areas of research include affordable broadband access, including wireless technologies. Also directs NSF to develop a plan that describes the current status of broadband access for scientific research purposes. Introduced May 10, 2007; referred to House Committee on Science and Technology. Passed House May 21, 2007. Passed Senate July 19, 2007. Signed into law, August 9, 2007.

P.L. 110-161 (H.R. 2764)

Consolidated Appropriations Act, 2008. For Rural Utilities Service, U.S. Department of Agriculture, provides \$6.45 million to support a loan level of \$300 million for the broadband loan program, and \$13.5 million for broadband community connect grants. Signed by President, December 26, 2007.

P.L. 110-234 (H.R. 2419)

Food, Conservation, and Energy Act of 2008. Reauthorizes broadband program at the Rural Utilities Service through FY2012. Establishes a National Center for Rural Telecommunications Assessment. Directs USDA and the FCC to submit to Congress a comprehensive rural broadband strategy. Introduced May 22, 2007; referred to Committee on Agriculture, and in addition to Committee on Foreign Affairs. Subcommittee on Specialty Crops, Rural Development, and Foreign Agriculture held markup of Title VII (Rural Development) on June 6, 2007. Reported by House Committee on Agriculture (H.Rept. 110-256) on July 23, 2007. Passed House July 27, 2007. Passed Senate with an amendment, December 14, 2007. Conference report (H.Rept. 110- 627) approved by the House May 14, 2008, and by the Senate May 15, 2008. Vetoed by the President, May 21, 2008. House and Senate overrode veto on May 21 and May 22, 2008. Became P.L. 110-234, May 22, 2007.

P.L. 110-329 (H.R. 2638)

Consolidated Security, Disaster Assistance, and Continuing Appropriations Act, 2009. Continuing resolution funds RUS broadband loan and grant program at FY2008 levels through March 6, 2009. Signed by President September 30, 2008.

P.L. 110-385 (S. 1492)

Broadband Data Improvement Act. Seeks to improve the quality of federal broadband data collection and encourage state initiatives that promote broadband deployment. Requires the FCC to collect demographic information on unserved areas, data comparing broadband service with 75 communities in at least 25 nations abroad, and data on consumer use of broadband. Directs the Census Bureau to collect broadband data, the Government Accountability Office to study broadband data metrics and standards, and the Department of Commerce to provide grants supporting state broadband initiatives. Introduced May 24, 2007; referred to Committee on Commerce, Science, and Transportation. Ordered to be reported July 19, 2007; reported by Committee (S.Rept. 110-204) and placed on Senate Legislative Calendar, October 24, 2007. Passed by Senate with an amendment September 26, 2008. Passed by House September 29, 2008. Became P.L. 110-385, October 10, 2008.

Broadband Stimulus Legislation in the 111th

In December 2008, leadership in the House and Senate, as well as the Obama transition team, announced their intention to include a broadband component in the infrastructure portion of the economic stimulus package. At the same time, numerous interested parties, including: broadband equipment manufacturers; large, mid-sized, and small wireline and wireless service providers; satellite operators; telecommunications unions; consumer groups; education groups; public safety organizations; think tanks; and others unveiled a multitude of specific proposals for government support of broadband infrastructure.⁴⁵

House

On January 21, 2009, the House Appropriations Committee approved legislative language for the spending portion of the economic stimulus package (H.R. 1, American Recovery and Reinvestment Act of 2009). The legislation would provide \$6 billion to support deployment of broadband and wireless services in rural, unserved, and underserved areas of the nation. Of the total, \$2.825 billion would be provided to the Rural Utilities Service (RUS) of the Department of Agriculture, and \$3. 175 billion to the National Telecommunications and Information Administration of the Department of Commerce. The House broadband stimulus provisions are included within Title II (under Rural Utilities Service), Title III (under

⁴⁵ See CRS Report R40149, *Infrastructure Programs: What's Different About Broadband?*, by Charles B. Goldfarb and Lennard G. Kruger, p. 21.

National Telecommunications and Information Administration), and Title VI (Broadband Communications) of H.R. 1. Specifically, the legislation breaks down as follows:

- **\$2.825 billion** to the Rural Utilities Service for additional loans, loan guarantees, and grants to finance “open access” broadband infrastructure. Specifies that at least 75% of the area to be served by a project receiving funds shall be in a rural area without sufficient access to high speed broadband service to facilitate economic development, as determined by the Secretary of Agriculture. Priority is given to projects that provide service to the most rural residents that do not have access to broadband services. Priority is also given to borrowers and former borrowers of rural telephone loans.
- **\$350 million** to the National Telecommunications and Information Administration to establish the State Broadband Data and Development Grant Program, as authorized by the recently enacted Broadband Data Improvement Act (P.L. 110-385). Grants would be used to develop and implement statewide initiatives to identify and track the availability and adoption of broadband within each state. Would also be used to develop and maintain a nationwide broadband inventory map.
- **\$1 billion** to NTIA for Wireless Deployment Grants for wireless voice service and advanced wireless broadband service (at least 3 Mbps downstream, 1 Mbps upstream). To the extent possible, 25% of the grants are to be awarded for providing wireless voice service in unserved areas, and 75% for advanced wireless broadband service in underserved areas. Grant recipients are required to operate on an “wireless open access” basis.
- **\$1.825 billion** to NTIA for Broadband Deployment Grants for basic broadband service (at least 5 Mbps downstream, 1 Mbps upstream) or advanced broadband service (at least 45 Mbps downstream, 15 Mbps upstream). To the extent possible, 25% of the grants are to be awarded for providing basic broadband in unserved areas, and 75% for advanced broadband service in underserved areas. Grant recipients are required to operate on an “open access” basis.

For the Wireless and Broadband Deployment Grants, the terms “unserved,” “underserved,” “open access,” and “wireless open access” shall be defined by the FCC not later than 45 days after enactment of the legislation. Also for these grants, each state planning to participate is required to submit to NTIA a report detailing which geographic areas within that state are most in need of wireless voice, advanced wireless broadband, basic broadband, and advanced broadband services in both unserved and underserved areas. Unserved and underserved areas identified by a state shall not constitute more than 20% of the population or geographic area of that state.

While the RUS broadband programs and the State Broadband Data and Development Grant Program were previously authorized (the RUS programs have operated for seven years, while the state grants is newly established by P.L. 110-385, the Broadband Data Improvement Act, and not yet funded), the Broadband Deployment Grants and the Wireless Deployment Grants would be newly authorized.

On January 22, 2009, the House Committee on Energy and Commerce marked up and approved sections 3101 (nationwide broadband inventory map to be developed by NTIA) and 3102 (authorizing wireless and broadband deployment grants at NTIA). An amendment

in the nature of a substitute, offered by the Chairman, additionally requires NTIA to issue an annual report assessing the impact and effectiveness of the grants, and expands the list of eligible entities to include satellite companies. Other amendments agreed to by the Committee would:

- include the improvement of interoperable broadband communications systems used for public safety and emergency response among factors to be considered in award decisions;
- direct the FCC to review and revise its definitions of unserved and underserved areas after completion of NTIA's nationwide broadband inventory map;
- direct the FCC to submit to Congress a national broadband plan; and
- direct NTIA to consider whether an eligible entity is a socially and economically disadvantaged small business.

On January 28, 2009, the House passed H.R. 1. An amendment agreed to by the House would make projects funded by the newly established NTIA broadband and wireless grant programs eligible for worker training grant money (under Title IX, Subtitle A of H.R. 1).

Senate

On February 7, 2009, a substitute amendment to H.R. 1 (S.Amdt. 570, the "Collins-Nelson amendment") was submitted in the Senate. S.Amdt. 570 would provide \$7 billion to NTIA for establishment of a national broadband service development and expansion program called the Broadband Technology Opportunities Program. This is \$2 billion less than what was provided in the Senate Appropriations Committee mark (S. 336, S.Rept. 111-3). The program, as provided in S.Amdt. 570, consists of:

- **\$6.650 billion** for competitive broadband grants, of which not less than \$200 million shall be available for competitive grants for expanding public computer center capacity (including at community colleges and public libraries); not less than \$250 million to encourage sustainable adoption of broadband service; and \$10 million transferred to the Department of Commerce Office of Inspector General for audits and oversight.
- **\$350 million** for funding the Broadband Data Improvement Act (P.L. 110-385) and for the purpose of developing and maintaining a broadband inventory map, which shall be made accessible to the public no later than two years after enactment. Funds deemed necessary and appropriate by the Secretary of Commerce may be transferred to the FCC for the purposes of developing a national broadband plan, which shall be completed one year after enactment.

Significant language related to Broadband Technology Opportunities Program grants includes the following:

- 50% of the total grant funding shall be used to support projects in rural communities, and funds may be transferred for this purpose to USDA's Rural Utilities Service if

deemed appropriate by the Secretary of Commerce and in consultation with the Secretary of Agriculture. In cases where this funding is made available to the RUS broadband loan, loan guarantee, and grant programs, at least 75% of the area to be served by a project receiving funds shall be in a rural area without sufficient access to high speed broadband service to facilitate economic development, as determined by the Secretary of Agriculture. Priority is given to projects that provide service to the highest proportion of rural residents that do not have sufficient access to broadband services.

- Among the purposes of the grant program is to provide broadband to citizens residing in unserved and underserved areas. NTIA may consult with the chief executive officer of any state with respect to identifying unserved and underserved areas within that state, and with respect to the allocation of grant funds within that state.
- NTIA shall, in coordination with the FCC, develop nondiscrimination and network interconnection obligations that shall be contractual conditions of grants awarded.
- The federal share of any project may not exceed 80% unless NTIA determines financial hardship.
- Grant eligibility includes: a state or political subdivision, a nonprofit foundation, corporation, institution or association, Indian tribe, Native Hawaiian organization or other nongovernmental entity in partnership with a state or political subdivision, Indian tribe, or native Hawaiian organization.
- Grants may be used to acquire equipment and technology necessary for broadband infrastructure, to construct and deploy broadband service related infrastructure, to ensure access to broadband by community anchor institutions, to facilitate broadband service by vulnerable populations, to construct and deploy broadband facilities to improve public safety communications.

S.Amdt. 570 also includes an investment tax credit for qualified broadband expenditures. The provision would establish a 10% tax credit for investment in current generation broadband in rural and underserved areas, a 20% tax credit for investment in current generation broadband in unserved areas, and a 20% tax credit for investment in next generation broadband in rural, underserved, and unserved areas. Current generation is defined as at least 5 Mbps download speed and 1 Mbps upload, or for wireless broadband, 3Mbps download and 768 kbps upload. Next generation is defined as at least 100 Mbps download and 20 Mbps upload.

On February 10, 2009, the Senate passed H.R. 1 as amended by S.Amdt. 570.

Comparison of House and Senate Bills

The following are some key similarities and differences between the House-passed and Senate- passed broadband provisions of H.R. 1:

- both the House and Senate bills would rely primarily on grants as a strategy to stimulate broadband deployment – House total funding is \$6 billion, Senate total is \$7 billion;

- House would provide \$3.175 billion to NTIA and \$2.825 billion to RUS broadband programs; Senate provides all funding to NTIA, but directs that 50% should finance projects in rural areas, DOC can transfer this funding in part to the RUS broadband loan and grant programs if deemed appropriate;
- both the House and Senate bills would provide \$350 million to NTIA for funding the Broadband Data Improvement Act and to develop a national broadband inventory map;
- Senate specifically sets aside not less than \$200 million for competitive grants for expanding public computer center capacity (including at community colleges and public libraries) and not less than \$250 million to encourage sustainable adoption of broadband service; funding is not specifically set aside for these purposes in the House bill;
- Senate has a 20% matching requirement for grant recipients (which can be waived in case of financial hardship); House doesn't have a matching requirement but cites a 20% match as a positive consideration when assessing grant applications;
- House sets funding allocation percentages for Broadband Technology Opportunity grants based on minimum broadband speed requirements (download and upload) and whether area is unserved or underserved, directs FCC to define "unserved" and "underserved" within 45 days; Senate doesn't prescribe allocations based on minimum download/upload speeds and whether an area is unserved or underserved, instead directs NTIA to consult with each state to identify unserved and underserved areas as well as the appropriate allocation of grant funds within that state;
- House mandates "open access" requirement for grant projects and requires that projects adhere to FCC's net neutrality principles, directs FCC to define "open access" and "wireless open access" within 45 days; Senate directs that NTIA shall, in coordination with the FCC, develop nondiscrimination and network interconnection obligations that shall be contractual conditions of grants awarded;
- House defines entities eligible for grants as essentially any provider of wireless or broadband service, including states or local governments; Senate defines an eligible applicant as a state or political subdivision thereof, a nonprofit foundation, corporation, institution or association, Indian tribe, Native Hawaiian organization or other nongovernmental entity in partnership with a state or political subdivision, Indian tribe, or native Hawaiian organization;
- Senate includes broadband investment tax credits; House does not include broadband tax incentives;
- House directs that 50% of grant funds are to be awarded no later than September 30, 2009; Senate directs all funds to be awarded by the end of FY20 10; and
- both the House and Senate bills direct FCC to develop a national broadband plan in one year.

P.L. 111-5: The American Recovery and Reinvestment Act of 2009

On February 17, 2009, President Obama signed P.L. 111-5, the American Recovery and Reinvestment Act (ARRA). Broadband provisions of the ARRA provide a total of **\$7.2 billion**, primarily for broadband grants. The total consists of \$4.7 billion to NTIA/DOC for a newly established Broadband Technology Opportunities Program and \$2.5 billion to existing RUS/USDA broadband programs.⁴⁶

Regarding the \$2.5 billion to RUS/USDA broadband programs, the ARRA specifies that at least 75% of the area to be served by a project receiving funds shall be in a rural area without sufficient access to high speed broadband service to facilitate economic development, as determined by the Secretary of Agriculture. Priority is given to projects that provide service to the most rural residents that do not have access to broadband services. Priority is also given to borrowers and former borrowers of rural telephone loans.

Of the \$4.7 billion appropriated to NTIA:

- \$4.35 billion is directed to a competitive broadband grant program, of which not less than \$200 million shall be available for competitive grants for expanding public computer center capacity (including at community colleges and public libraries); not less than \$250 million to encourage sustainable adoption of broadband service; and \$10 million transferred to the Department of Commerce Office of Inspector General for audits and oversight; and
- \$350 million is directed for funding the Broadband Data Improvement Act (P.L. 110-385) and for the purpose of developing and maintaining a broadband inventory map, which shall be made accessible to the public no later than two years after enactment. Funds deemed necessary and appropriate by the Secretary of Commerce may be transferred to the FCC for the purposes of developing a national broadband plan, which shall be completed one year after enactment.

The Broadband Technology Opportunities Program within NTIA is authorized by Division B, Title VI of the ARRA. Specific implementation requirements and guidelines for the new NTIA broadband grants are as follows:

- Directs NTIA to consult with each state to identify unserved and underserved areas (with respect to access to broadband service) as well as the appropriate allocation of grant funds within that state. The Conferees (H.Rept. 111-16) “intend that the NTIA has discretion in selecting the grant recipients that will best achieve the broad objectives of the program.”
- The substitute does not define “unserved area,” “underserved area,” and “broadband.” The Conferees instructed NTIA to coordinate its understanding of these terms with the FCC, and in defining “broadband service” to take into consideration technical differences between wireless and wireline networks and to

⁴⁶ For information on existing broadband programs at RUS, see: CRS Report RL33816, *Broadband Loan and Grant Programs in the USDA's Rural Utilities Service*, by Lennard G. Kruger.

consider the actual speeds these networks are able to deliver to consumers under a variety of circumstances.

- Directs NTIA, in coordination with the FCC, to publish “non-discrimination and network interconnection obligations” that shall be contractual conditions of awarded grants, and specifies that these obligations should adhere, at a minimum, to the FCC’s broadband principles to promote the openness and interconnected nature of the Internet (FCC 05-151, adopted August 5, 2005).
- Directs NTIA, when considering applications for grants, to consider whether the project will provide the greatest broadband speed possible to the greatest population of users in the area. There are no specific speed thresholds that applicants must meet to be eligible for a grant. The Conferees acknowledged that while speed thresholds could have the unintended effect of thwarting broadband deployment in some areas, deploying next-generation speeds would likely result in greater job creation and job preservation. NTIA is instructed to “seek to fund, to the extent practicable, projects that provide the highest possible, next- generation broadband speeds to consumers.”
- Defines entities eligible for grants as: a state or political division thereof; the District of Columbia; a territory or possession of the United States; an Indian tribe or native Hawaiian organization; a nonprofit foundation, corporation, institution or association; or any other entity, including a broadband service or infrastructure provider, that NTIA finds by rule to be in the public interest. It was the intent of the Conferees that as many entities as possible be eligible to apply for a grant, including wireless carriers, wireline carriers, backhaul providers, satellite carriers, public-private partnerships, and tower companies.
- Requires NTIA to consider whether a grant applicant is a socially and economically disadvantaged small business as defined under the Small Business Act.
- Directs NTIA to ensure that all awards are made before the end of FY2010. Grantees will be required to substantially complete projects within two years after the grant is awarded.
- Directs that the federal share of any project cannot exceed 80% unless the applicant petitions NTIA and demonstrates financial need.

The Conference Agreement and public law bill did not include the broadband investment tax credit provisions that were contained in the Senate bill. For more information on implementation of the broadband provisions of the ARRA, see CRS Report R40436, *Broadband Infrastructure Programs in the American Recovery and Reinvestment Act*, by Lennard G. Kruger.

Other Legislation in the 111 Congress

P.L. 111-8 (H.R. 1105). Omnibus Appropriations Act, 2009. Appropriates to RUS/USDA \$15.619 million to support a loan level of \$400.487 million for the Rural Broadband Access Loan and Loan Guarantee Program, and \$13.406 million for the Community Connect Grant Program. To the FCC, designates not less than \$3 million to establish and administer a State Broadband Data and Development matching grants program

for State-level broadband demand aggregation activities and creation of geographic inventory maps of broadband service to identify gaps in service and provide a baseline assessment of statewide broadband deployment. Passed House February 25, 2009. Passed Senate March 10, 2009. Signed by President, March 12, 2009.

H.R. 691 (Meeks). Broadband Access Equality Act of 2009. Amends the Internal Revenue Code of 1986 to provide credit against income tax for businesses furnishing broadband services to underserved and rural areas. Introduced January 26, 2009; referred to Committee on Ways and Means.

H.R. 760 (Eshoo). Advanced Broadband Infrastructure Bond Initiative of 2009. Amends the Internal Revenue Code of 1986 to provide an income tax credit to holders of bonds financing new advanced broadband infrastructure. Introduced January 28, 2009; referred to Committee on Ways and Means and in addition to Committee on Energy and Commerce.

Concluding Observations

As Congress considers various options for encouraging broadband deployment, a key issue is how to strike a balance between providing federal assistance for unserved and underserved areas where the private sector may not be providing acceptable levels of broadband service, while at the same time minimizing any deleterious effects that government intervention in the marketplace may have on competition and private sector investment. In addition to loans, loan guarantees, and grants for broadband infrastructure deployment, a wide array of policy instruments are available to policymakers including universal service reform, tax incentives to encourage private sector deployment, broadband bonds, demand-side incentives (such as assistance to low income families for purchasing computers), regulatory and deregulatory measures, and spectrum policy to spur roll-out of wireless broadband services. In assessing federal incentives for broadband deployment, Congress will likely consider the appropriate mix of broadband deployment incentives to create jobs in the short and long term, the extent to which incentives should target next-generation broadband technologies, the extent to which “underserved” areas with existing broadband providers should receive federal assistance, and how broadband stimulus measures of the ARRA might fit into the context of overall goals for a national broadband policy.

Table 1. Selected Federal Domestic Assistance Programs Related to Telecommunications Development

Program	Agency	Description	FY2008 (obligations)	Web Links for More Information http://12.46.245.173/cfda/cfda.html : Go to “All Programs Listed Numerically” and search by program
Public Telecommunications Facilities—Planning and Construction	National Telecommunications and Information Administration, Dept. of Commerce	Assists in planning, acquisition, installation and modernization of public telecommunications facilities	\$19.5 million	http://www.ntia.doc.gov/otiahom/e/ptfp/index.html
Investments for Public Works and Economic Development Facilities	Economic Development Administration, Dept. of Commerce	Provides grants to economically distressed areas for construction of public facilities and infrastructure, including broadband deployment and other types of telecommunications enabling projects	\$249 million	http://www.eda.gov/
Rural Telephone Loans and Loan Guarantees	Rural Utilities Service, U.S. Dept. of Agriculture	Provides long-term direct and guaranteed loans to qualified organizations for the purpose of financing the improvement, expansion, construction, acquisition, and operation of telephone lines, facilities, or systems to furnish and improve telecommunications service in rural areas	\$145 million (hardship loans); \$250 million (cost of money loans); \$295 million (FFB Treasury loans)	http://www.usda.gov/rus/telecom/index.htm
Distance Learning and Telemedicine Loans and Grants	Rural Utilities Service, U.S. Dept. of Agriculture	Provides seed money for loans and grants to rural community facilities (e.g., schools, libraries, hospitals) for advanced telecommunications systems that can provide health care and educational benefits to rural areas	\$24.7 million (grants) \$28 million (loans and loan-grant combinations)	http://www.usda.gov/rus/telecom/dlt/dlt.htm
Rural Broadband Access Loan and Loan Guarantee Program	Rural Utilities Service, U.S. Dept. of Agriculture	Provides loan and loan guarantees for facilities and equipment providing broadband service in rural communities	\$300 million (cost of money loans)	http://www.usda.gov/rus/telecom/broadband.htm

Table 1. Continued

Program	Agency	Description	FY2008 (obligations)	Web Links for More Information http://12.46.245.173/cfda/cfda.html : Go to "All Programs Listed Numerically" and search by program
Community Connect Broadband Grants	Rural Utilities Service, U.S. Dept. of Agriculture	Provides grants to applicants proposing to provide broadband service on a "community-orientated connectivity" basis to rural communities of under 20,000 inhabitants.	\$13.4 million	http://www.usda.gov/rus/telecom/index.htm
Education Technology State Grants	Office of Elementary and Secondary Education, Dept. of Education	Grants to State Education Agencies for development of information technology to improve teaching and learning in schools	\$267 million	http://www.ed.gov/Technology/TLCF/index.html
Ready to Teach	Office of Assistant Secretary for Educational Research and Improvement, Dept. of Education	Grants to carry out a national telecommunication-based program to improve the teaching in core curriculum areas.	\$10.7 million	http://www.ed.gov/programs/readytoe/index.html
Special Education — Technology and Media Services for Individuals with Disabilities	Office of Special Education and Rehabilitative Services, Dept. of Education	Supports development and application of technology and education media activities for disabled children and adults	\$39.3 million	http://www.ed.gov/about/offices/list/osers/index.html?src=mr/
Telehealth Network Grants	Health Resources and Services Administration, Department of Health and Human Services	Grants to develop sustainable telehealth programs and networks in rural and frontier areas, and in medically unserved areas and populations.	\$3.9 million	http://www.hrsa.gov/telehealth/
Medical Library Assistance	National Library of Medicine, National Institutes of Health, Department of Health and Human Services	Provides funds to train professional personnel; strengthen library and information services; facilitate access to and delivery of health science information; plan and develop advanced information	\$67.5 million	http://www.nlm.nih.gov/ep/extramural.html

Table 1. Continued

Program	Agency	Description	FY2008 (obligations)	Web Links for More Information http://12.46.245.173/cfda/cfda.html: Go to “All Programs Listed Numerically” and search by program
		networks; support certain kinds of biomedical publications; and conduct research in medical informatics and related sciences		
State Library Program	Office of Library Services, Institute of Museum and Library Services, National Foundation on the Arts and the Humanities	Grants to state library administrative agencies for promotion of library services that provide all users access to information through State, regional, and international electronic networks	\$171.5 million	http://www.imls.gov/grants/library/lib_gsla.asp#po
Native American and Native Hawaiian Library Services	Office of Library Services, Institute of Museum and Library Services, National Foundation on the Arts and the Humanities	Supports library services including electronically linking libraries to networks	\$3.7 million	http://www.imls.gov/grants/library/lib_nat.asp
Appalachian Area Development	Appalachian Regional Commission	Provides project grants for Appalachian communities to support the physical infrastructure necessary for economic development and improved quality of life.	\$73 million	http://www.arc.gov/index.do?no deId = 21
Delta Area Economic Development	Delta Regional Authority	Grants to support self-sustaining economic development of eight states in Mississippi Delta region.	\$7.8 million	http://www.dra.gov/programs/informationtechnology
Denali Commission Program	Denali Commission	Provides grants through a federal and state partnership designed to provide critical infrastructure and utilities throughout Alaska, particularly in distressed communities	\$106 million	http://www.denali.gov/

Source: Prepared by CRS based on information from the Catalog of Federal Domestic Assistance.

Table 2. Selected Federal Programs Funding Broadband Access

Program	Comments
Programs Funding Access to Telecommunications in Underserved Areas	
Rural Broadband Access Loan and Loan Guarantee Program (Rural Utilities Service, U.S. Department of Agriculture)	Provides loan and loan guarantees for facilities and equipment providing broadband service in rural communities.
Community Connect Broadband Grants (Rural Utilities Service, U.S. Department of Agriculture)	Provides grants to applicants proposing to provide broadband service on a “community-oriented connectivity” basis to rural communities of under 20,000 inhabitants.
Rural Telephone Loans and Loan Guarantees (Rural Utilities Service, U.S. Department of Agriculture)	Since 1995, the RUS Rural Telephone Loan and Loan Guarantee program—which has traditionally financed telephone voice service in rural areas under 5,000 inhabitants—has required that all telephone facilities receiving financing must be capable of providing DSL broadband service at a rate of at least 1 megabyte per second.
Universal Service Fund: High Cost Program (Federal Communications Commission)	While the USF’s High Cost Program does not explicitly fund broadband infrastructure, subsidies are used, in many cases, to upgrade existing telephone networks.
Federal Economic Development Programs Funding Broadband Access	
Community Development Block Grants (Department of Housing and Urban Development)	In Michigan, a Digital Divide Investment Program (DDIP) combined Michigan Broadband Development Authority loans (initially \$12 million) and CDBG grant funding (\$4 million) to deploy a hybrid fixed wireless and fiber network in two rural counties which would make broadband affordable for low to moderate income residents.
Indian Community Development Block Grants (Department of Housing and Urban Development)	In 2005, HUD awarded the Coquille Indian Tribe a \$421,354 grant used to fund the Coquille Broadband Technology Infrastructure Project. The project will allow for improved connectivity for reservation residents, improvements in rural community access, and potentially increased wireless Internet access for the Tribal and surrounding communities.
Grants for Public Works and Economic Development Facilities (Economic Development Administration, Department of Commerce)	Supports the proliferation of broadband networks as a key priority for regional economic growth. Examples: \$6 million grant to a company in Virginia for investment in 300 miles of fiber optic cable in nine counties and three cities; \$2 million grant to companies in Vermont to help build a 424 mile fiber optic broadband network in rural northern Vermont; and \$270 thousand to support a Rhode Island Wireless Innovation Networks project. EDA encourages communities eligible for RUS programs to access that first before applying for EDA investment dollars.
Appalachian Regional Commission	The Appalachian Regional Development Act Amendments of 2002 reauthorized ARC for five years and created specific authority for a Regionwide initiative to bridge the telecommunications and technology gap between the Appalachian Region and the rest of the United States.
	Supported a telecommunications initiative (\$33 million over five year period) which includes projects such as: a regional fiber network across northeast Mississippi; wireless demonstrations in rural New York, Ohio, Pennsylvania, Virginia, West Virginia, and Georgia; and a regionwide effort in Kentucky to compile an inventory of broadband

Table 2. Continued

Program	Comments
	access across the 51 Appalachian counties and work with the private sector to substantially increase broadband coverage. In Maryland, a county-wide high-speed wireless network, funded by ARC over several years, now serves over 4,500 customers.
Delta Regional Authority	During a strategic planning retreat in February 2005, the DRA board determined that one of the authority's three top policy priorities would be information technology. To support its policy position, the authority devoted \$150,000 to create an information technology plan for the region.
Denali Commission	Funded Telecommunications Survey in 2000 which was used to determine the state of broadband deployment in Alaska and used as basis for applying for RUS broadband assistance.
Applications-Based Federal Programs Related to Broadband	
Universal Service Fund: Schools and Libraries or "E-Rate" Program (Federal Communications Commission)	Used to fund broadband access for schools and libraries.
Universal Service Fund: Rural Health Care Program (Federal Communications Commission)	Used to fund broadband access for rural health care centers.
Distance Learning and Telemedicine Program (Rural Utilities Service, U.S. Department of Agriculture)	Provides seed money for loans and grants to rural community facilities (e.g., schools, libraries, hospitals) for advanced telecommunications systems that can provide health care and educational benefits to rural areas.
Public Safety Interoperable Communications Grant Program (National Telecommunications and Information Administration, Department of Commerce)	Provides funding to states and territories to enable and enhance public safety agencies' interoperable communications capabilities.
Telehealth Network Grants (Health Resources and Services Administration, Department of Health and Human Services)	Grants to develop sustainable telehealth programs and networks in rural and frontier areas, and in medically unserved areas and populations.
Public Telecommunications Facilities Program (National Telecommunications and Information Administration, Department of Commerce)	Grants for public television, public radio, and nonbroadcast distance learning projects.
Education technology programs (Department of Education)	Examples include Education Technology State Grants, Ready to Teach.
State Library Grants (Office of Library Services, Institute of Museum and Library Services, National Foundation on the Arts and the Humanities)	Grants to state library administrative agencies for promotion of library services that provide all users access to information through State, regional, and international electronic networks.
Medical Library Assistance (National Library of Medicine, National Institutes of Health, Department of Health and Human Services)	Provides funds to train professional personnel; strengthen library and information services; facilitate access to and delivery of health science information; plan and develop advanced information networks; support certain kinds of biomedical publications; and conduct research in medical informatics and related sciences.

Chapter 18

NET NEUTRALITY: BACKGROUND AND ISSUES

*Angele A. Gilroy**
Telecommunications Policy

Summary

As congressional policymakers continue to debate telecommunications reform, a major point of contention is the question of whether action is needed to ensure unfettered access to the Internet. The move to place restrictions on the owners of the networks that compose and provide access to the Internet, to ensure equal access and non-discriminatory treatment, is referred to as “net neutrality.” There is no single accepted definition of “net neutrality.” However, most agree that any such definition should include the general principles that owners of the networks that compose and provide access to the Internet should not control how consumers lawfully use that network; and should not be able to discriminate against content provider access to that network. Concern over whether it is necessary to take steps to ensure access to the Internet for content, services, and applications providers, as well as consumers, and if so, what these should be, is a major focus in the debate over telecommunications reform. Some policymakers contend that more specific regulatory guidelines may be necessary to protect the marketplace from potential abuses which could threaten the net neutrality concept. Others contend that existing laws and Federal Communications Commission (FCC) policies are sufficient to deal with potential anti-competitive behavior and that such regulations would have negative effects on the expansion and future development of the Internet.

A consensus on this issue has not yet formed, and the 111th Congress, to date, has not introduced stand-alone legislation to address this issue. However, the net neutrality issue has been narrowly addressed within the context of the economic stimulus package (P.L. 111-5). Provisions in that law require the National Telecommunications and Information Administration (NTIA), in consultation with the FCC, to establish “... nondiscrimination and network interconnection obligations” as a requirement for grant participants in the Broadband Technology Opportunities Program (BTOP). This report will be updated as events warrant.

* E-mail address: agilroy@crs.loc.gov, 7-7778

Network Neutrality

As congressional policymakers continue to debate telecommunications reform, a major point of contention is the question of whether action is needed to ensure unfettered access to the Internet. The move to place restrictions on the owners of the networks that compose and provide access to the Internet, to ensure equal access and non-discriminatory treatment, is referred to as “net neutrality.” There is no single accepted definition of “net neutrality.” However, most agree that any such definition should include the general principles that owners of the networks that compose and provide access to the Internet should not control how consumers lawfully use that network; and should not be able to discriminate against content provider access to that network.

What, if any, action should be taken to ensure “net neutrality” has become a major focal point in the debate over broadband, or high-speed Internet access, regulation. As the marketplace for broadband continues to evolve, some contend that no new regulations are needed, and if enacted will slow deployment of and access to the Internet, as well as limit innovation. Others, however, contend that the consolidation and diversification of broadband providers into content providers has the potential to lead to discriminatory behaviors which conflict with net neutrality principles. The two potential behaviors most often cited are the network providers’ ability to control access to and the pricing of broadband facilities, and the incentive to favor network-owned content, thereby placing unaffiliated content providers at a competitive disadvantage.¹

In 2005 two major actions dramatically changed the regulatory landscape as it applied to broadband services, further fueling the net neutrality debate. In both cases these actions led to the classification of broadband Internet access services as Title I information services, thereby subjecting them to a less rigorous regulatory framework than those services classified as telecommunications services. In the first action, the U.S. Supreme Court, in a June 2005 decision (*National Cable & Telecommunications Association v. Brand X Internet Services*), upheld the Federal Communications Commission’s (FCC) 2002 ruling that the provision of cable modem service (i.e., cable television broadband Internet) is an interstate information service and is therefore subject to the less stringent regulatory regime under Title I of the Communications Act of 1934.² In a second action, the FCC in an August 5, 2005 decision, extended the same regulatory relief to telephone company Internet access services (i.e., wireline broadband Internet access, or DSL), thereby also defining such services as information services subject to Title I regulation.³ As a result neither telephone companies nor cable companies, when providing broadband services, are required to adhere to the more stringent regulatory regime for telecommunications services found under Title II (common

¹ The practice of consumer tiering, that is the charging of different rates to subscribers based on access speed, is not the concern. Access tiering, that is the charging of different fees, or the establishment of different terms and conditions to content, services, or applications providers for access to the broadband pipe is the focus of the net neutrality policy debate.

² 47U.S.C. 151 et seq. For a full discussion of the Brand X decision see CRS Report RL32985, *Defining Cable Broadband Internet Access Service: Background and Analysis of the Supreme Court's Brand X Decision*, by Angie A. Welborn and Charles B. Goldfarb.

³ See http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-260433A2.pdf for a copy of FCC Chairman Martin’s statement. For a summary of the final rule see Appropriate Framework for Broadband Access to the Internet Over Wireline Facilities. *Federal Register*, Vol. 70, No. 199, October, 17, 2005, p. 60222.

carrier) of the 1934 Act.⁴ However, classification as an information service does not free the service from regulation. The FCC continues to have regulatory authority over information services under its Title I, ancillary jurisdiction.⁵

Simultaneous to the issuing of its August 2005 information services classification order, the FCC also adopted a policy statement outlining the following four principles to “encourage broadband deployment and preserve and promote the open and interconnected nature of [the] public Internet:” (1) consumers are entitled to access the lawful Internet content of their choice; (2) consumers are entitled to run applications and services of their choice (subject to the needs of law enforcement); (3) consumers are entitled to connect their choice of legal devices that do not harm the network; and (4) consumers are entitled to competition among network providers, application and service providers, and content providers. Then FCC Chairman Martin did not call for their codification. However, he stated that they will be incorporated into the policymaking activities of the Commission.⁶ For example, one of the agreed upon conditions for the October 2005 approval of both the Verizon/MCI and the SBC/AT&T mergers was an agreement made by the involved parties to commit, for two years, “... to conduct business in a way that comports with the Commission’s (September 2005) Internet policy statement.... ”⁷ In a further action AT&T included in its concessions to gain FCC approval of its merger to BellSouth to adhering, for two years, to significant net neutrality requirements. Under terms of the merger agreement, which was approved on December 29, 2006, AT&T agreed to not only uphold, for 30 months, the FCC’s Internet policy statement principles, but also committed, for two years (expired December 2008), to stringent requirements to “... maintain a neutral network and neutral routing in its wireline broadband Internet access service.”⁸

In perhaps one of its most significant actions relating to its Internet policy statement to date, the FCC, on August 1, 2008, ruled that Comcast Corp., a provider of Internet access over cable lines, violated the FCC’s policy statement, when it selectively blocked peer-to-peer connections in an attempt to manage its traffic.⁹ This practice, the FCC concluded, “... unduly interfered with Internet users’ rights to access the lawful Internet content and to use the applications of their choice.” While no monetary penalties were imposed, Comcast is required to stop these practices by the end of 2008. Comcast stated that it will comply with the order, but it has filed an appeal in the U.S. DC Court of Appeals.¹⁰

Separately, in an April 2007 action, the FCC released a notice of inquiry (WC Docket No. 07-52), which is still pending, on broadband industry practices seeking comment on a wide range of issues including whether the August 2005 Internet policy statement should be

⁴ For example, Title II regulations impose rigorous anti-discrimination, interconnection and access requirements. For a further discussion of Title I versus Title II regulatory authority see CRS Report RL32985, cited above.

⁵ Title I of the 1934 Communications Act gives the FCC such authority if assertion of jurisdiction is “reasonably ancillary to the effective performance of [its] various responsibilities.” The FCC in its order, cites consumer protection, network reliability, or national security obligations as examples of cases where such authority would apply (see paragraph 36 of the final rule summarized in the *Federal Register* cite in footnote 3, above).

⁶ See <http://www.fcc.gov/headlines2005.html>. August 5, 2005. *FCC Adopts Policy Statement on Broadband Internet Access.*

⁷ See http://hraunfoss.FCC.gov/edocs_public/attachmatch/DOC-261936A1.pdf. It should be noted that applicants offered certain voluntary commitments, of which this was one.

⁸ See http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-269275A1.pdf.

⁹ See http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-08-183A1.pdf.

¹⁰ For a legal discussion of the FCC’s Comcast decision see CRS Report R40234, Net Neutrality: The Federal Communications Commission’s Authority to Enforce Its Network Management Principles , by Kathleen Ann Ruane.

amended to incorporate a new principle of nondiscrimination and if so, what form it should take. On January 14, 2008 the FCC issued three public notices seeking comment on issues related to network management (including the now-completed Comcast ruling) and held two (February 25 and April 17, 2008) public hearings specific to broadband network management practices.

Network Prioritization

As consumers expand their use of the Internet and new multimedia and voice services become more commonplace, control over network quality also becomes an issue. In the past, Internet traffic has been delivered on a “best efforts” basis. The quality of service needed for the delivery of the most popular uses, such as email or surfing the Web, is not as dependent on guaranteed quality. However, as Internet use expands to include video, online gaming, and voice service, the need for uninterrupted streams of data becomes important. As the demand for such services continues to expand, network broadband operators are moving to prioritize network traffic to ensure the quality of these services. Prioritization may benefit consumers by ensuring faster delivery and quality of service and may be necessary to ensure the proper functioning of expanded service options. However, the move on the part of network operators to establish prioritized networks, while embraced by some, has led to a number of policy concerns.

There is concern that the ability of network providers to prioritize traffic may give them too much power over the operation of and access to the Internet. If a multi-tiered Internet develops where content providers pay for different service levels, the potential to limit competition exists, if smaller, less financially secure content providers are unable to afford to pay for a higher level of access. Also, if network providers have control over who is given priority access, the ability to discriminate among who gets such access is also present. If such a scenario were to develop, the potential benefits to consumers of a prioritized network would be lessened by a decrease in consumer choice and/or increased costs, if the fees charged for premium access are passed on to the consumer. The potential for these abuses, however, is significantly decreased in a marketplace where multiple, competing broadband providers exist. If a network broadband provider blocks access to content or charges unreasonable fees, in a competitive market, content providers and consumers could obtain their access from other network providers. As consumers and content providers migrate to competitors, market share and profits of the offending network provider will decrease leading to corrective action or failure. However, this scenario assumes that every market will have a number of equally competitive broadband options from which to choose, and all competitors will have equal access to, if not identical, at least comparable content.

Despite the FCC’s ability to regulate broadband services under its Title I ancillary authority and the issuing of its broadband principles, some policymakers feel that more specific regulatory guidelines may be necessary to protect the marketplace from potential abuses; a consensus on what these should specifically entail, however, has yet to form. Others feel that existing laws and FCC policies regarding competitive behavior are sufficient to deal with potential anti-competitive behavior and that no action is needed and if enacted at this time, could result in harm.

The Congressional Debate

The issue of net neutrality, and whether legislation is needed to ensure access to broadband networks and services, has become a major focal point in the debate over telecommunications reform.¹¹ Those opposed to the enactment of legislation to impose specific Internet network access or “net neutrality” mandates claim that such action goes against the long standing policy to keep the Internet as free as possible from regulation. The imposition of such requirements, they state, is not only unnecessary, but would have negative consequences for the deployment and advancement of broadband facilities. For example, further expansion of networks by existing providers and the entrance of new network providers, would be discouraged, they claim, as investors would be less willing to finance networks that may be operating under mandatory build-out and/or access requirements. Application innovation could also be discouraged, they contend, if, for example, network providers are restricted in the way they manage their networks or are limited in their ability to offer new service packages or formats. Such legislation is not needed, they claim, as major Internet access providers have stated publicly that they are committed to upholding the FCC’s four policy principles.¹² Opponents also state that advocates of regulation cannot point to any widespread behavior that justifies the need to establish such regulations and note that competition between telephone and cable system providers, as well as the growing presence of new technologies (e.g., satellite, wireless, and power lines) will serve to counteract any potential anti-discriminatory behavior. Furthermore, opponents claim, even if such a violation should occur, the FCC already has the needed authority to pursue violators. They note that the FCC has not requested further authority¹³ and has successfully used its existing authority, in the August 1, 2008, Comcast decision (see above) as well as in a March 3, 2005, action against Madison River Communications. In the latter case, the FCC intervened and resolved, through a consent decree, an alleged case of port blocking by Madison River Communications, a local exchange (telephone) company.¹⁴ The full force of antitrust law is also available, they claim, in cases of discriminatory behavior.

¹¹ For a more lengthy discussion regarding proponents’ and opponents’ views see, for example, testimony from Senate Commerce Committee hearings on Net Neutrality, February 7, 2006; http://commerce.senate.gov/public/index.cfm?FuseAction=Hearings.Hearing&Hearing_ID=1708.

¹² See testimony of Kyle McSlarrow, President and CEO of the National Cable and Telecommunications Association and Walter McCormick, President and CEO of the United States Telecom Association, hearing on Net Neutrality before the Senate Commerce Committee, February 7, 2006, cited above.

¹³ Former FCC Chairman Martin indicated that the FCC has the necessary tools to uphold the FCC’s stated policy principles and did not request additional authority. Furthermore, former Chairman Martin stated that he was “... confident that the marketplace will continue to ensure that these principles are maintained” and is “... confident therefore, that regulation is not, nor will be, required.” See former *Chairman Kevin J. Martin Comments on Commission Policy Statement*, at http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-260435A2.pdf. However, FCC Commissioner Copps, in an April 3, 2006 speech, did express concerns over the concentration in broadband facilities providers and their “... ability, and possibly even the incentive, to act as Internet gatekeepers ...” and called for a “national policy” on “... issues regarding consumer rights, Internet openness, and broadband deployment.” See http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC-264765A1.pdf, for a copy of Commissioner Copps’ speech.

¹⁴ The FCC entered into a consent decree with Madison River Communications to settle charges that the company had deliberately blocked the ports on its network that were used by Vonage Corp. to provide voice over Internet protocol (VoIP) service. Under terms of the decree Madison River agreed to pay a \$15,000 fine and not block ports used for VoIP applications. See http://hraunfoss.fcc.gov/edocs_public/attachmatch/DA-05-543A2.pdf, for a copy of the consent decree.

Proponents of net neutrality legislation, however, feel that absent some regulation, Internet access providers will become gatekeepers and use their market power to the disadvantage of Internet users and competing content and application providers. They cite concerns that the Internet could develop into a two-tiered system favoring large, established businesses or those with ties to broadband network providers. While market forces should be a deterrent to such anti-competitive behavior, they point out that today's market for residential broadband delivery is largely dominated by only two providers, the telephone and cable television companies, and that, at a minimum, a strong third player is needed to ensure that the benefits of competition will prevail.¹⁵ The need to formulate a national policy to clarify expectations and ensure the "openness" of the Internet is important to protect the benefits and promote the further expansion of broadband, they claim. The adoption of a single, coherent, regulatory framework to prevent discrimination, supporters claim, would be a positive step for further development of the Internet, by providing the marketplace stability needed to encourage investment and foster the growth of new services and applications. Furthermore, relying on current laws and case-by-case anti-trust-like enforcement, they claim, is too cumbersome, slow, and expensive, particularly for small start-up enterprises.¹⁶

Congressional Activity

The 110th Congress addressed the debate over net neutrality largely within the broader issue of telecommunications reform. Then House Telecommunications and the Internet Subcommittee Chairman Markey, a strong advocate of net neutrality legislation, introduced legislation (H.R. 5353) to address this issue and held a May 6, 2008 hearing on the measure. House Judiciary Chairman Conyers introduced H.R. 5994, a bill which establishes an antitrust approach to address anticompetitive and discriminatory practices by broadband providers as a follow-up to a March 11, 2008 hearing on net neutrality held by the House Judiciary Antitrust Task Force. A standalone net neutrality measure (S. 215) was introduced and referred to the Senate Commerce, Science, and Transportation Committee where an April 22, 2008 hearing on the "Future of the Internet" was held. No further activity was undertaken in the 110th Congress.

A consensus on this issue has not yet formed, and no stand-alone measures addressing net neutrality have been introduced in the 111th Congress, to date. House Communications, Technology, and the Internet Subcommittee Chairman Boucher has stated that he continues to work with broadband providers and content providers to seek common ground on network management practices, and at this time, is pursuing this approach.

However, the net neutrality issue has been narrowly addressed within the context of the economic stimulus package. H.R. 1 (P.L. 111-5) contains provisions that require the National Telecommunications and Information Administration (NTIA), in consultation with the FCC, to establish "... nondiscrimination and network interconnection obligations" as a requirement

¹⁵ For FCC market share data for high-speed connections see *High-Speed Services for Internet Access: Status as of June 30, 2007*, Federal Communications Commission, Industry Analysis and Technology Division, Wireline Competition Bureau, released March 2008. View report at http://hraunfoss.fcc.gov/edocs_public/attachmatch/DOC280906A1.pdf.

¹⁶ For example, see testimony of Vint Cerf, VP Google, Earl Comstock, President and CEO of CompTel, and Jeffrey Citron, Chairman and CEO Vonage, hearing on Net Neutrality, before the Senate Commerce Committee, February 7, 2006, cited above.

for grant participants in the Broadband Technology Opportunities Program (BTOP). The law further directs that the FCC's four broadband policy principles, issued in August 2005, are the minimum obligations to be imposed.¹⁷ The NTIA has not, as of yet, issued these requirements.

¹⁷ For a further more detailed discussion of the broadband infrastructure programs contained in P.L. 111-5 see CRS Report R40436, *Broadband Infrastructure Programs in the American Recovery and Reinvestment Act*, by Lennard G. Kruger.

INDEX

A

abstraction, 8, 18, 115, 128, 130, 137
accessibility, 9, 110, 129
acute lymphoblastic leukemia, x, 126, 131, 220
adaptability, 95
adaptation, vii, 1, 7, 8, 13, 17, 26, 267
adolescence, 162
adolescents, xi, 85, 91, 138, 157, 158, 159, 160, 162, 163
ADP, 19, 25, 26
adverse event, 130
advertising, 9, 10
Africa, 106
African Americans, 311
age, xii, 82, 90, 91, 99, 117, 159, 160, 161, 191, 209, 210, 235, 242, 243, 246
aggregation, 9, 13, 247, 331
AIDS, 192, 211
Alaska, 334, 336
albumin, 127
algorithm, viii, xiii, 8, 33, 38, 40, 41, 42, 46, 51, 58, 71, 101, 103, 166, 176, 182, 194, 195, 198, 212, 218, 238, 239, 240, 241, 263, 264, 266, 272, 273, 274, 277, 292, 294, 305, 307
alienation, 160
allocated time, 83
alternatives, 25, 105, 106, 107, 242
ambiguity, x, 125, 128, 130, 131
America Online, 142
American Educational Research Association, 95
American Psychological Association, 123
American Recovery and Reinvestment Act, xiv, 125, 309, 310, 315, 317, 318, 329, 330, 343
amino acids, 181
amplitude, 38, 41, 42, 43, 44, 46, 280, 303
anaclitic depression, 116
anatomy, 247
anger, 116
annotation, 204, 219, 220, 221, 227
antibody, 226
antitrust, 341, 342
anxiety, 123

apoptosis, 186
ARC, 335, 336
architecture design, 231, 233
arginine, 203, 219
artificial intelligence, 203
Asia, 191, 219
assessment, 6, 27, 59, 123, 213, 227, 242, 288, 307, 308, 317, 331
assets, 23, 94
assignment, 182
assumptions, 287
asthma, 130
AT&T, 29, 339
atoms, 197
attacker, 151, 173
attacks, xi, 34, 151, 152, 165, 166, 171, 172, 173, 267, 297, 299, 300
attitudes, 153
attractiveness, 28
Australia, 82, 85, 93, 95, 108, 189
Austria, 305
authentication, xi, xiii, 8, 11, 13, 34, 77, 78, 150, 165, 168, 171, 173, 175, 176, 263, 264, 265, 266, 270, 275, 276, 286, 297, 299, 300, 301, 305
authenticity, 166
authority, 240, 241, 322, 335, 336, 339, 340, 341
authors, xiii, 82, 105, 117, 121, 154, 160, 204, 263, 264
automation, 18, 62, 75, 76
autonomy, 124
average costs, 319
averaging, 173, 247, 319
awareness, viii, 81, 95, 102, 106, 155, 231

B

bandwidth, xiii, 8, 11, 146, 263, 264, 266, 269, 270, 275, 276, 279, 280, 287, 297, 298, 301, 306
bandwidth resources, 301
banking, 176
barriers, 93, 129, 315, 316
basic research, 323
BEA, 22, 24

- Beck Depression Inventory, 118
 behavior, ix, xiv, 35, 92, 97, 99, 101, 104, 108, 337,
 340, 341, 342
 Beijing, 28
 BellSouth, 339
 belongingness, 121
 benchmarking, 16, 23, 27, 153
 benign, 151
 bias, 187
 binding, 199, 200, 216
 biochemistry, 180
 bioinformatics, xi, 179, 190, 226
 biological processes, 185
 biological systems, xii, 180, 189, 207
 biomarkers, 187
 biotechnology, 211
 blocks, 265, 273, 299, 340
 blog, 157, 236
 blogs, 158, 163, 236
 blood, 126, 187, 203
 blood plasma, 187
 blood pressure, 126
 Bluetooth, 66
 bonding, 113, 114, 116, 120, 121, 122
 bonds, 36, 112, 331
 boredom, 119, 159
 borrowers, 325, 329
 bounds, 206
 branching, 182, 199
 breast cancer, 201
 breathing, 130
 broadband, vii, xiv, 309, 310, 311, 312, 313, 314,
 315, 316, 317, 318, 321, 322, 323, 324, 325, 326,
 327, 328, 329, 330, 331, 332, 333, 335, 336, 338,
 339, 340, 341, 342, 343
 browser, 98, 186
 browsing, 98
 buffer, 152
 building blocks, 21
 bullying, 90, 91, 92, 93, 94, 95
 business management, 23
 business model, 28, 30, 101
 buttons, 51, 103
- C**
- cabbage, 224
 cable system, 341
 cable television, 310, 338, 342
 campaigns, 9, 11, 13, 273
 Canada, 107, 109, 110, 111, 117, 155, 159, 313
 cancer, 122, 130, 132, 180, 185, 187, 197, 214, 215,
 219, 223
 cancer cells, 185
 candidates, 202, 219
 carcinoma, 132
 caregivers, 99
 carrier, 225, 339
 case study, xiii, 5, 29, 31, 95, 155, 212, 251, 257,
 263, 264, 265, 275, 301
 categorization, 171, 173, 192, 233
 category d, 171, 310
 cDNA, 225
 cell, 23, 85, 133, 134, 144, 180, 182
 cell line, 180, 182
 cell phones, 144
 census, 315, 322, 324
 channels, vii, xiii, 1, 2, 7, 10, 17, 18, 19, 25, 26, 27,
 64, 66, 72, 73, 223, 263, 264, 266, 270, 287, 296,
 301
 chemical interaction, 201, 218
 chemotherapeutic agent, 131
 chemotherapy, 131
 children, 85, 89, 92, 94, 95, 130, 138, 333
 China, 28, 176
 chronic diseases, 99
 chunking, 198, 202, 204
 classes, 205, 256, 265, 273
 classification, vii, 1, 3, 4, 5, 6, 12, 16, 17, 21, 22, 23,
 26, 27, 129, 180, 195, 211, 221, 228, 338, 339
 classroom, 82, 84, 85, 86, 89, 91, 92, 94
 classrooms, 85, 87, 92, 93, 320
 cleavage, 205
 clients, 65, 66, 75, 76, 78, 143, 144, 146, 147, 148,
 150, 151, 152, 232
 clinical trials, 130
 clone, 225
 clustering, 105, 192, 193, 211, 227
 clusters, 22, 200, 220
 CMC, 115, 116, 117, 122, 141, 142, 154
 CNN, 236
 codes, 19, 129, 131, 251, 258, 277, 302, 307, 312
 coding, xiii, 34, 138, 174, 263, 264, 265, 268, 273,
 274, 275, 276, 277, 288, 289, 290, 294, 304, 305,
 307
 cognition, 85
 cognitive abilities, 102, 106
 coherence, 199, 302
 cohesion, 226
 cohort, 129
 collaboration, 29, 85, 134, 135, 152, 153, 155
 college students, xi, 142, 157, 158, 159, 161, 162,
 163
 colleges, 326, 328, 329
 collusion, 172
 combined effect, 25
 commerce, 30, 315, 341
 communication, ix, x, xiii, 18, 27, 35, 61, 62, 64, 65,
 66, 67, 68, 69, 71, 72, 73, 76, 82, 84, 85, 93, 111,
 112, 113, 114, 115, 117, 119, 120, 124, 141, 142,
 148, 152, 153, 154, 155, 157, 160, 161, 162, 163,
 166, 167, 236, 252, 264, 265, 286, 287, 299, 300,
 301, 305
 communication systems, 76
 communication technologies, 82, 93
 Communications Act, 319, 338, 339
 Communications Act of 1934, 319, 338

- communications channel, 64, 67, 70, 72, 73
community, viii, x, 12, 18, 81, 82, 83, 84, 91, 101, 106, 114, 123, 124, 141, 147, 149, 159, 205, 221, 229, 317, 323, 326, 327, 328, 329, 332, 333, 335, 336
compatibility, 2, 287
compensation, 34, 35
competition, xiv, 66, 310, 315, 316, 331, 339, 340, 341, 342
competitor, 20
competitors, 18, 27, 340
compiler, 252
complexity, vii, xiii, 1, 2, 3, 17, 18, 19, 129, 134, 137, 194, 250, 252, 253, 263, 264, 272, 276, 277, 281, 286, 288, 294, 298, 305
compliance, 136
components, xii, xiii, 2, 8, 11, 18, 63, 69, 72, 104, 107, 114, 115, 118, 120, 143, 166, 168, 197, 200, 203, 237, 238, 241, 249, 250, 251, 252, 253, 254, 255, 257, 262
composition, 180
compression, xiii, 8, 54, 59, 171, 173, 174, 175, 263, 264, 297, 298, 303, 305
computation, xi, 62, 78, 126, 179
computational grid, 262
computer simulations, 276, 277
computer systems, 64, 153
computing, viii, xii, 61, 64, 70, 72, 79, 176, 229, 231, 249, 251, 255, 262
concentration, 67, 341
conceptual model, 247
conceptualization, 116, 119
condor, 255
confidence, 196, 197, 209
confidentiality, 150, 166
configuration, xii, xiii, 8, 13, 69, 72, 78, 116, 229, 231, 249, 250, 252, 253, 257, 258, 259, 261
conflict, 160, 170, 338
Congress, xiv, xv, 109, 219, 309, 311, 315, 317, 318, 320, 322, 323, 326, 330, 331, 337, 342
conjugation, 244
connectivity, xiii, 119, 199, 264, 265, 317, 333, 335
conscious awareness, 120
consensus, xv, 337, 340, 342
consent, 341
consolidation, 338
Constitution, 43
construct validity, 118
construction, 212, 217, 218, 220, 233, 332
constructivism, 94
consumer choice, 340
consumers, xiv, 312, 316, 321, 330, 337, 338, 339, 340
consumption, 255, 297, 298
continuity, 266, 276
control, viii, xiv, 2, 8, 9, 13, 18, 47, 61, 64, 65, 66, 67, 72, 74, 75, 77, 78, 92, 116, 117, 121, 128, 130, 147, 148, 166, 171, 176, 180, 229, 257, 286, 287, 299, 300, 305, 337, 338, 340
convergence, 296
conversion, 34, 172
coping strategies, 119
copper, 310
corporations, viii, 61, 64
correlation, 25, 120, 173, 208, 242, 266, 272, 279
correlations, 119, 160, 226
cost saving, 125
costs, 2, 19, 62, 66, 67, 68, 70, 75, 126, 127, 153, 250, 275, 296, 297, 300, 301, 319, 322, 340
cough, 127
coughing, 130
counseling, 128
Court of Appeals, 339
covering, 23, 321
CPU, 151
credentials, 143, 299
credit, 318, 331
criminal acts, 297
cross-sectional study, 159
cryptography, xi, 77, 165, 166, 167, 176, 297, 299, 300
crystal structure, 205
cues, 160, 243
currency, 242, 243, 244
curriculum, 83, 88, 89, 333
customers, 3, 4, 5, 17, 18, 21, 25, 26, 153, 171, 202, 232, 312, 336
cyclophosphamide, 131
cytochrome, 201
cytokines, 192

D

- danger, 89
dangerous behaviour, 89
data analysis, xi, 5, 179, 209
data collection, 160, 311, 314, 315, 324
data gathering, 314
data mining, 200, 211, 226, 227
data processing, 180
data set, 118, 186, 189, 204, 207
data structure, 77, 78
data transfer, 64, 65, 67, 68, 72
database, xiii, 36, 37, 126, 128, 180, 182, 184, 186, 193, 199, 201, 203, 205, 207, 209, 215, 219, 221, 222, 223, 224, 225, 237, 238, 241, 247, 249, 251, 252, 256, 257, 258, 260, 262
dating, 162
decision makers, 154
decision making, x, 3, 27, 141, 142, 241, 242, 247
decisions, 98, 99, 154, 326
decoding, xiii, 53, 263, 264, 266, 267, 269, 270, 275, 276, 277, 287, 288, 294, 307
deconvolution, 182
deductive reasoning, 203
defense, xi, 165

E

- definition, xiv, 82, 93, 113, 204, 207, 212, 213, 266, 316, 320, 321, 322, 337, 338
degradation, 273, 279, 288, 303
delivery, vii, 1, 2, 7, 8, 9, 10, 13, 15, 17, 19, 25, 27, 30, 31, 68, 69, 70, 333, 336, 340, 342
Denmark, 93, 108
denoising, 172
density, 266, 319
Department of Agriculture, 318, 324, 335, 336
Department of Commerce, 312, 313, 315, 316, 322, 324, 326, 329, 335, 336
Department of Defense, 190
dependent variable, 27
depression, ix, 104, 111, 116, 118, 119, 120, 121, 122, 123, 124, 137, 139
deregulation, 316
derivatives, 285
designers, 232, 301
detection, viii, xi, xiii, 33, 34, 35, 39, 40, 41, 42, 43, 45, 46, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 58, 59, 74, 78, 130, 151, 165, 167, 168, 169, 170, 171, 263, 264, 267, 268, 273, 274, 276, 278, 279, 288
developing countries, 81
DFT, 175
diabetes, 99
differentiation, 10, 224
digital cameras, 236
digital communication, 90, 167
digital divide, vii, xiii, xiv, 309, 310, 315, 317, 318
dimensionality, 236
directors, 321
discipline, 167
disclosure, xi, 158, 161
discourse, 124
discrimination, 342
disease gene, 219, 224
disorder, 118, 222
dispersion, 267
distance learning, 318, 336
distortions, 174, 266, 276, 281
distribution, xii, 7, 9, 11, 17, 18, 25, 28, 35, 36, 47, 49, 50, 54, 102, 166, 182, 184, 197, 199, 229, 231, 232, 233, 257
diversification, 338
diversity, 219, 314
division, 12, 286, 300, 330
DNA, 186
doctors, ix, 97, 98
drawing, 23, 104, 105, 191
Drosophila, 224
drug discovery, 187
drugs, 126, 131, 180, 199
DSL, 310, 311, 312, 335, 338
duplication, 9
duration, 127, 267, 271, 302, 306
eavesdropping, 150
e-commerce, 98, 106, 176, 229, 312
economic development, 312, 318, 327, 334
economic growth, 312, 316, 335
economics, 17, 19, 112
education, 81, 82, 84, 91, 95, 176, 333, 336
educational experience, 153
elaboration, 281
election, 66, 67
e-mail, 84, 89, 90, 91, 94, 108, 114, 142, 153, 154, 157, 158, 296, 340
emergency response, 326
emission, 270
emotion, 119, 122
emotion regulation, 122
emotional disorder, 112
emotions, 87, 88
employees, 142, 152, 153
employment, 313
encapsulation, 73
encoding, xiii, 39, 47, 50, 51, 52, 53, 54, 55, 56, 57, 264, 265, 266, 267, 270, 273, 274, 275, 276, 277, 279, 299
encryption, 34, 150, 166, 167, 299
endothelial cells, 186
end-users, 149
energy, 173, 186, 266, 267, 270, 281, 283, 284, 294, 295
energy transfer, 186
England, 107
entropy, 101, 201, 202
environment, 2, 11, 12, 17, 19, 25, 53, 70, 74, 85, 91, 92, 94, 114, 116, 117, 127, 154, 231, 242, 247, 250, 251, 252, 255, 287, 316
error detection, 34
esophageal cancer, 224
EST, 225
ethnicity, 124, 160, 161
Europe, 123, 185, 298, 301
European Union, 185
evolution, 58, 210, 298
excitation, xiii, 263, 264, 270, 273, 282, 284, 288, 307
exclusion, 90, 158, 305
execution, 2, 11, 12, 72, 74, 77, 79, 249
expenditures, 327
expertise, 82, 84, 86, 94, 99, 128, 137, 208
exploitation, 75
expressed sequence tag, 205
external validity, 27
extraction, xi, xiii, 165, 167, 168, 171, 172, 189, 190, 191, 194, 195, 196, 198, 199, 200, 201, 202, 203, 204, 206, 207, 209, 210, 211, 212, 214, 215, 216, 217, 218, 219, 220, 226, 227, 228, 265, 276, 277, 297
extrapolation, 209

F

Facebook, 82, 117
 face-to-face interaction, 112, 113
 factor analysis, 123
 failure, 116, 119, 120, 121, 131, 340
 fairness, 278
 false negative, 206
 false positive, 132, 206
 family, viii, 19, 81, 82, 83, 85, 92, 112, 113, 115, 119, 122, 129, 130, 221, 312
 family income, 312
 family members, viii, 81, 82, 92, 119, 221
 farms, 239
 faster delivery, 340
 fear, 99, 116
 Federal Communications Commission, xiv, xv, 310, 311, 312, 314, 315, 316, 317, 319, 320, 321, 322, 323, 324, 325, 326, 327, 328, 329, 330, 337, 338, 339, 340, 341, 342, 343
 feedback, 11, 101, 102, 106
 feelings, 91, 113, 116, 161
 females, 103, 158, 159, 160, 161
 fever, 127
 filters, 5, 41, 51, 52, 53, 103
 finance, 176, 325, 328, 341
 financial capital, 112
 financial resources, 112
 financing, 331, 332, 335
 fine tuning, 185
 Finland, 142, 155
 firms, 4, 5, 22, 27, 142
 first generation, 2
 fish, 190, 203
 fish oil, 190, 203
 flexibility, 2, 3, 17, 18, 19, 26, 75, 127, 154, 250, 301
 flooding, 90, 151, 243
 fluorescence, 181
 focus groups, 5
 focusing, 5, 6, 23, 82, 84, 93, 191, 317
 Ford, 313
 framing, 278, 288
 France, 99, 308
 fraud, 75
 freedom, 83
 friendship, 90, 113, 159, 160, 162
 fulfillment, 65
 functional architecture, vii, 1, 4, 7, 12, 22
 funding, 318, 320, 321, 326, 327, 328, 329, 335, 336
 funds, 319, 321, 322, 324, 325, 326, 327, 328, 329, 333, 336
 fusion, 241

G

gambling, 98

gender, 159, 160, 162
 gene, 180, 190, 192, 193, 195, 196, 197, 198, 199, 200, 201, 204, 208, 211, 212, 213, 214, 215, 216, 217, 220, 221, 223, 224, 226, 227, 228
 gene expression, 211, 215, 217, 223, 226
 gene promoter, 224
 general practitioner, 126
 generation, 2, 10, 30, 107, 115, 121, 168, 180, 185, 191, 204, 207, 308, 311, 317, 327, 330, 331
 genes, xii, 182, 185, 189, 190, 193, 194, 195, 196, 198, 199, 205, 207, 209, 214, 215, 220, 221, 223, 224
 genetic marker, 224
 genome, 220, 225
 genomics, 189, 190, 205, 225
 genre, 175
 geography, 314
 Georgia, 335
 girls, 84, 91
 goals, 63, 81, 115, 117, 264, 319, 322, 331
 gold, 137, 204
 google, 109, 110, 247
 gossip, 159
 government, xiv, 256, 309, 310, 313, 315, 316, 317, 318, 324, 331
 government intervention, xiv, 310, 316, 331
 GPS, 19, 64, 65
 grants, 10, 19, 310, 315, 317, 322, 323, 324, 325, 326, 327, 328, 329, 330, 331, 332, 333, 334, 335, 336
 graph, 135, 182, 199, 228, 253
 grouping, 23, 131, 262, 277
 groups, xii, 74, 84, 86, 90, 91, 99, 106, 113, 114, 122, 134, 137, 146, 154, 172, 181, 189, 195, 324
 growth, 153, 172, 193, 311, 313, 342
 growth rate, 153
 guessing, 218
 guidelines, xiv, 27, 106, 220, 252, 329, 337, 340

H

harassment, 90, 92, 95
 harm, 339, 340
 Hawaii, 109
 health, vii, x, 82, 99, 101, 103, 106, 123, 125, 126, 127, 128, 129, 130, 133, 134, 137, 138, 318, 319, 320, 321, 332, 333, 336
 health care, x, 125, 318, 319, 320, 321, 332, 336
 health care system, x, 125
 health information, 99, 128, 129, 138
 health insurance, 129
 heart failure, 187
 heat, 132, 180
 height, 126
 heterogeneity, 250
 high school, 83, 94
 higher quality, 274
 HIV, 132

- homogeneity, 274
Honda, 186
Hong Kong, 220
hormone, 216, 217
hospitals, 317, 332, 336
host, 70, 86, 87, 151, 173, 240, 260, 261, 264, 265, 266, 267, 268, 276, 277
House, 4, 16, 17, 20, 21, 23, 25, 26, 27, 112, 119, 123, 177, 323, 324, 325, 326, 327, 328, 331, 342
households, 85, 311
hub, 240, 241
human genome, 224
human kinase, 223
human motivation, 122
human subjects, 129
Hunter, 210, 211, 215, 216, 227
hybrid, 192, 199, 201, 202, 214, 273, 335
hypermedia, 108
hypertext, 8, 247
hypothesis, 62, 191, 204, 207, 220
- identification, 3, 4, 16, 34, 76, 77, 78, 139, 180, 198, 200, 202, 212, 213, 214, 237, 279, 299
identity, xi, 83, 88, 89, 91, 150, 179, 297, 300
image, ix, xi, xiii, 47, 48, 49, 50, 51, 53, 54, 55, 56, 57, 97, 102, 105, 150, 165, 168, 169, 170, 171, 173, 174, 175, 176, 177, 249, 251, 258, 259, 260, 261, 265, 266, 268
images, ix, 17, 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 97, 98, 102, 105, 144, 167, 170, 172, 173, 175, 176, 177, 236, 256, 260, 264
immersion, 107
immunodeficiency, 132
immunoglobulin, 226
impairments, 266, 269, 301
implementation, x, xi, xiii, 23, 26, 51, 54, 125, 126, 128, 129, 142, 165, 167, 196, 207, 231, 232, 233, 260, 263, 264, 265, 275, 276, 296, 297, 305, 329, 330
impulsive, 276
in vitro, 181
in vivo, 181
incentives, 317, 331
inclusion, 106, 115, 116, 158
income, xiv, 309, 310, 311, 312, 319, 331, 335
income tax, 331
increased access, 316
independence, 275
independent variable, 27
index numbers, 38, 42
indexing, 193, 197, 211
India, 165, 176, 235
indication, 100, 101
indicators, 294
induction, 218
industry, 31, 149, 176, 314, 339
- infancy, 204
infection, 130, 137
inferences, 26
infinite, 198
information exchange, 114, 167
information retrieval, xii, 98, 108, 128, 137, 189, 190, 191, 204, 206, 207, 211, 212, 215, 226, 247
information technology, 31, 233, 310, 333, 336
infrastructure, 2, 4, 12, 17, 19, 64, 66, 67, 69, 71, 72, 73, 74, 76, 77, 78, 250, 296, 315, 318, 324, 325, 327, 330, 331, 332, 334, 335, 343
initiation, 159, 319
injuries, 203
innovation, 316, 317, 338, 341
insertion, 167, 168, 173, 175, 276, 277, 279, 280
insight, vii, 1, 3, 27
institutions, 310, 317, 327
insurance, xi, 165, 176
integration, viii, viii, 1, 2, 10, 12, 13, 18, 25, 26, 61, 62, 63, 64, 70, 73, 214, 225, 226, 265, 266, 275, 278, 296, 305
integrity, 150, 166, 171
intellectual property, 9, 101, 167, 264, 265
intellectual property rights, 167, 264
intelligence, 23, 231
interaction, viii, 10, 11, 61, 64, 75, 76, 87, 88, 98, 103, 110, 114, 119, 121, 123, 150, 196, 200, 202, 204, 205, 214, 215, 216, 219, 220, 221, 222, 223, 224, 227, 228, 229, 231
interactions, xii, 70, 83, 85, 86, 87, 88, 89, 90, 91, 93, 112, 113, 121, 122, 126, 182, 183, 189, 196, 199, 200, 201, 205, 208, 209, 210, 214, 215, 218, 221, 223, 227
interdependence, 124
interestingness, 197
interface, viii, ix, 33, 35, 40, 51, 52, 53, 57, 70, 71, 73, 97, 98, 102, 103, 104, 105, 106, 107, 158, 192, 220, 231, 237, 254, 255, 259, 301
interference, 268, 292
interferon, 192
internal validity, 27
internalizing, 23
internet, vii, viii, 34, 35, 81, 82, 90, 91, 123, 124, 155, 229
Internet Relay Chat, 142
interoperability, vii, xiii, 1, 2, 3, 17, 18, 19, 20, 25, 26, 74, 225, 249, 296
interpersonal communication, 115
interpersonal interactions, 112
interpersonal relations, 114, 115, 121, 161
interpersonal relationships, 114, 115, 121, 161
interval, 283
intervention, 92, 151, 175, 299
intimacy, x, xi, 115, 157, 158, 160, 161, 162, 163
intonation, 41, 44, 46
introversion, 124
inventors, 149
inversion, 170, 272, 306

investment, xiv, 66, 93, 310, 315, 316, 317, 321, 327, 328, 330, 331, 335, 342
 investors, 341
 ionization, 182
 ions, 181
 IP address, 143
 IP networks, vii, 8, 296, 297
 ISC, 59
 isolation, x, 125
 isotope, 182
 ISPs, 144
 Italy, 1, 28, 306

J

Japan, 4, 33, 43, 59, 95, 155, 229, 298
 Java, 199
 job creation, 330
 jobs, 313, 331
 jurisdiction, 339

K

knowledge acquisition, 220
 Korea, 176, 313

L

labeling, 186, 218, 228
 labor, 253
 lactation, 226
 Lagrange multipliers, 284, 285
 landscape, 88, 236, 338
 language, ix, xii, 37, 88, 97, 103, 107, 128, 147, 190, 196, 197, 198, 199, 200, 209, 210, 212, 213, 216, 218, 219, 227, 230, 231, 235, 237, 239, 243, 244, 247, 261, 324, 326
 language processing, xii, 103, 107, 128, 196, 210, 216, 218, 235, 243, 244
 latency, 265, 270, 275, 276, 277, 297, 298
 Latin America, 109
 law enforcement, 339
 laws, xiv, 337, 340, 342
 leadership, 324
 learners, 86, 153
 learning, viii, 81, 82, 83, 84, 85, 86, 92, 93, 94, 95, 98, 101, 105, 123, 126, 201, 202, 203, 212, 213, 227, 229, 250, 333
 learning environment, 82, 95
 learning outcomes, 82
 legal issues, 149
 legislation, xiv, xv, 309, 316, 322, 324, 325, 337, 341, 342
 leisure, 75
 lending, 58
 leukemia, 132

librarians, 153
 library services, 334, 336
 licenses, 260
 life cycle, 63
 life sciences, 219, 220
 limitation, xi, 27, 165, 173
 line, 5, 23, 44, 51, 103, 136, 243, 255, 266, 269, 310, 312, 314, 321
 linguistics, 191, 209
 links, 9, 82, 85, 86, 93, 101, 144, 190, 205, 220, 237, 239, 240, 301
 liquid chromatography, xi, 179, 186
 listening, 41, 44, 45, 57, 95, 152, 273, 288, 292, 303
 literacy, 87, 88, 89, 108
 loans, 317, 318, 325, 329, 331, 332, 335, 336
 local area networks, 142
 local community, 84
 local government, 328
 localization, 19, 64, 177, 205
 logging, 92, 145
 loneliness, 123, 159
 long distance, 64, 66, 67, 71, 74, 296, 321
 longitudinal study, 159, 163
 Louisiana, 108, 247

M

machine learning, 199, 202, 203, 212, 214, 219, 226
 magnesium, 190, 203, 210
 maintenance, 12, 65, 66, 72, 74, 75, 115, 162, 200, 254, 301
 males, 158, 160
 malware, 299
 management, vii, viii, 1, 2, 8, 9, 10, 11, 12, 13, 17, 18, 19, 23, 29, 31, 61, 63, 65, 92, 162, 255, 257, 321, 340, 342
 mandates, 129, 328, 341
 manipulation, 76, 166, 172, 175
 manufacturing, 26
 mapping, 23, 27, 180, 182, 184, 212, 213, 315
 market, 2, 4, 5, 7, 18, 23, 26, 27, 28, 30, 31, 77, 117, 153, 312, 315, 316, 340, 342
 market share, 117, 342
 market structure, 31
 marketing, 4, 9, 28
 marketplace, xiv, 310, 315, 316, 331, 337, 338, 340, 341, 342
 markets, 23, 314
 masking, 174, 270
 mass spectrometry, xi, 179, 181, 185, 186, 189
 matrix, 19, 20, 30, 228, 272
 measurement, 273, 288
 measures, xiv, 118, 126, 152, 159, 190, 191, 196, 197, 202, 206, 239, 310, 317, 331, 342
 media, vii, x, xi, 1, 2, 8, 17, 23, 26, 34, 62, 75, 83, 90, 94, 102, 104, 105, 107, 161, 162, 163, 165, 170, 171, 231, 236, 312, 333
 median, 312

- medical expertise, 134
 medication, 104, 130
 membership, 123
 membranes, 205, 225
 memory, 78, 102, 106, 250, 277
 men, 99
 mental model, 99
 mental state, 120
 mergers, 339
 messages, 10, 66, 68, 69, 70, 88, 92, 142, 143, 144, 145, 146, 148, 149, 150, 151, 152, 154, 157, 158, 159, 167, 236
 meta-analysis, 215
 metabolic pathways, 186, 215
 metabolites, 196
 methylation, 197, 215
 microarray technology, 208
 microscopy, 181, 185
 Microsoft, 6, 23, 149
 migration, 68, 69, 70, 258, 260
 mining, 128, 154, 194, 209, 210, 212, 213, 214, 215, 217, 219, 223, 227
 minorities, 159
 minority, 159
 mobile communication, viii, xi, 30, 64, 66, 68, 76, 165
 mobile device, 2, 65, 66, 71, 75, 102, 106, 107
 mobile phone, 85, 89
 mobility, 68, 322
 model, vii, 1, 3, 4, 5, 6, 7, 12, 13, 16, 21, 22, 23, 26, 27, 28, 29, 34, 47, 48, 49, 50, 57, 62, 63, 64, 66, 69, 75, 77, 78, 88, 101, 116, 143, 146, 160, 161, 163, 168, 185, 192, 198, 203, 217, 225, 226, 241, 242, 247, 252, 253, 262, 269, 270, 272, 273, 302, 303, 304, 306, 315
 modeling, viii, 27, 61, 153, 177, 253, 270
 models, 2, 8, 12, 62, 101, 108, 116, 176, 185, 195, 201, 202, 214, 215, 241, 242, 247, 273, 301
 modernization, 301, 332
 modules, 2, 4, 5, 7, 12, 16, 17, 23, 27, 199, 200, 250, 257, 258
 modulus, 268
 molecular biology, xii, 189, 211, 220
 molecules, 182, 183, 184, 185, 190
 money, 78, 93, 112, 326, 332, 336
 mood, 119
 morbidity, 137
 morphology, 195
 mortality, 137
 motion, xii, 229, 231, 233
 motivation, 117, 120, 158
 motives, vii, ix, 111, 115, 116, 117, 118, 119, 120, 121, 122, 123
 motor skills, 102, 106
 movement, 48
 mRNA, 183
 multidimensional, 12
 multimedia, xi, 18, 19, 28, 148, 165, 167, 175, 236, 237, 264, 265, 266, 340
 multimedia services, 18
 multiplication, 264, 266, 268, 274
 multivariate statistics, 124
 music, ix, 17, 28, 43, 45, 85, 86, 97, 98, 99, 105, 236
 mutation, 197, 216, 217
 myocardial infarction, 132
- N**
- narratives, 127, 131
 nasopharynx, 127
 nation, 256, 314, 317, 321, 322, 324
 National Institutes of Health, 138
 national policy, 341, 342
 national security, 339
 needy, 319
 negative relation, 19
 neglect, 54
 Netherlands, 93, 108, 159, 160, 307
 network elements, 17
 neural network, 201, 202, 203, 218
 neural networks, 201, 202, 203
 neurosurgery, 130
 New South Wales, 82
 newspapers, 89
 next generation, vii, 1, 2, 311
 nodes, xi, 179, 180, 183, 184
 noise, xiii, 41, 47, 167, 171, 173, 180, 263, 264, 265, 266, 267, 268, 273, 274, 275, 276, 278, 279, 280, 281, 287, 288, 302, 304, 307
 nonverbal cues, 162
 Norway, 126
 novelty, 136
 NTIA, xiv, xv, 309, 315, 316, 325, 326, 327, 328, 329, 330, 337, 342, 343
 nurses, 129, 130
 nursing, 126
- O**
- Obama Administration, xiv, 309, 317
 Obama, President, xiv, 309, 329
 objectives, 4, 190, 321, 329
 observations, xi, 165, 175, 208
 OECD, 313
 oil, 210
 omission, 117
 Omnibus Appropriations Act, 330
 one dimension, 116
 online information, 99
 openness, 330, 342
 operating system, xiii, 77, 142, 229, 249, 250, 251, 257, 258, 259
 operator, 241, 242, 321
 Operators, vii, 1, 2, 5, 10, 31
 optimal performance, 277, 288
 optimization, 101, 274, 277, 283, 288, 301

orchestration, 11, 13, 18
organelles, 205
orientation, 102, 205
orthogonality, 272, 280, 281, 284, 285
overlay, 28
overload, 106
oversight, 326, 329
ownership, xi, 165, 168, 171, 261, 265

P

Pacific, 190, 191, 210, 211, 212, 213, 215, 219, 227
parameters, xiii, 41, 44, 50, 51, 53, 54, 57, 58, 68,
76, 78, 126, 128, 180, 209, 249, 250, 253, 256,
259, 260, 264, 265, 269, 270, 271, 272, 274, 278,
279, 280, 281, 282, 283, 284, 285, 288, 293, 294,
295, 296, 297, 298, 301, 302
parents, viii, 81, 85, 87, 89, 90, 92, 93
partnership, 318, 327, 328, 334
password, 34, 90, 150, 299
pathology, 116, 132, 201
pathways, xi, 179, 180, 182, 183, 184, 186, 201, 205,
210, 225, 226
patient care, 137
PCA, 118, 119
PCM, 43, 273, 298
peers, 83, 95
penalties, 129, 339
peptides, 181, 182, 222
perceptions, 123
permit, 70, 73, 76, 78
personality, ix, 89, 91, 111, 116, 117, 121, 122
personality dimensions, ix, 111
phonemes, 197, 270
phosphorylation, 201, 219, 222
physical environment, 73, 103
physical health, 84, 111
piracy, 33
pitch, xiii, 41, 263, 264, 269, 273, 274
planning, 4, 64, 65, 66, 72, 74, 75, 149, 332
plants, 223
plasticity, 180
Poland, 149
polarity, 269, 306
police, 70
policy instruments, 331
policy makers, 321
polymorphism, 224
polymorphisms, 193
poor, vii, 1, 2, 26, 27, 121, 192, 208, 316
population, 124, 224, 312, 320, 325, 330
population density, 312
portfolio, vii, 1, 2, 11, 13, 17, 18, 23, 25, 26, 321
portfolio management, 18
ports, 93, 341
positive correlation, 25, 121, 160
positive interactions, 121
positive mood, 119
potassium, 222
poverty, 320
power, x, 65, 90, 92, 106, 125, 151, 183, 269, 270,
273, 274, 278, 280, 284, 316, 340, 341, 342
prediction, xiii, 180, 185, 218, 222, 263, 264, 270,
271, 272, 273, 280, 284, 294, 295, 296, 307
prednisone, 131
preference, 173
prevention, 171, 265, 297
prices, 105, 313
primary data, 6
primary school, vii, viii, 81, 82, 83, 84, 86, 89, 90,
91, 92, 93, 94, 95
privacy, x, 92, 125, 129, 142, 145, 167, 297
private sector, xiv, 309, 310, 316, 331, 336
probability, 34, 66, 134, 169, 180, 184, 196, 198
product design, 4
production, 4, 12, 62, 63, 64, 65, 67, 72, 74, 75, 255,
273, 282, 306
productivity, 120, 153
professions, 84
profits, 340
program, 58, 69, 71, 72, 74, 115, 151, 190, 210, 212,
230, 274, 314, 318, 319, 320, 321, 322, 323, 324,
326, 327, 329, 330, 332, 333, 334, 335
programming, 19, 101, 108, 209
programming languages, 19
proliferation, 335
promoter, 224
propagation, 299
proportionality, 182
proposition, vii, 1, 5, 28
prosperity, 312
prostate, 226
prostate cancer, 226
protein structure, 224
protein-protein interactions, 199, 201, 202, 203, 205,
210, 215, 216, 217, 218
proteins, xi, xii, 179, 180, 181, 182, 183, 184, 186,
189, 190, 194, 196, 199, 200, 201, 205, 207, 208,
213, 214, 215, 216, 221, 222, 223, 225
proteome, 179, 180, 181, 182
proteomics, vii, xi, 179, 180, 183, 185, 186, 187,
189, 205
protocol, 68, 69, 70, 146, 147, 148, 149, 150, 154,
166, 172, 173, 297, 299, 341
prototype, viii, 4, 33, 41, 46, 51, 53, 110
psychological well-being, 123
psychology, 112, 117
psychopathology, 123
public health, 126
public interest, 330
public radio, 66, 336
public safety, 324, 326, 327, 336
public service, 62
public television, 336
public-private partnerships, 315
pulse, 43, 126

Q

qualitative research, 6, 27
 quality assurance, 130
 quality improvement, x, 125
 quality of life, 334
 quality of service, 10, 29, 265, 296, 297, 301, 304, 340
 quantization, 171, 172, 176, 268, 298
 query, ix, 97, 98, 99, 101, 102, 103, 104, 107, 131, 136, 192, 193, 206, 209, 212, 217, 228, 236, 237, 238, 239, 241, 243, 246
 questioning, 238

R

race, 124, 160, 161
 radio, 66, 89, 236, 286, 300, 301, 305, 319
 range, vii, viii, 2, 4, 17, 19, 23, 25, 81, 82, 84, 85, 86, 142, 171, 196, 268, 269, 287, 339
 rating scale, 45
 ratings, 100
 REA, 318
 reaction mechanism, 221
 reading, 51, 99, 128, 134, 152, 158, 170, 196, 300
 reagents, 186
 real time, 61, 64, 66, 67, 70, 154
 reality, 95, 128
 reason, xii, 40, 55, 66, 68, 73, 77, 85, 126, 131, 235, 236, 281, 312
 reasoning, 203
 recall, 106, 128, 190, 191, 194, 195, 200, 201, 202, 203, 204, 206, 207, 208, 209, 298
 reception, 65
 receptors, 66, 216, 217
 reciprocal relationships, 120, 121
 reciprocity, 113, 162
 recognition, 30, 105, 190, 192, 193, 195, 204, 207, 211, 213, 214, 220, 228
 recovery, 278, 280, 292, 294
 redistribution, 47, 48, 50
 redundancy, 266
 regeneration, 83
 region, xi, 39, 83, 165, 334, 336
 regression, 159
 regulation, 120, 182, 185, 201, 338, 339, 341, 342
 regulations, xiv, 129, 136, 301, 337, 338, 339, 341
 regulators, 26, 221
 regulatory framework, 338, 342
 reinforcement, 112
 rejection, 116
 relationship, ix, 4, 11, 13, 17, 19, 20, 26, 27, 37, 47, 54, 111, 113, 114, 115, 158, 159, 161, 193, 196, 197, 201, 202, 204, 208, 226
 relationship maintenance, 161
 relationship management, 13
 relatives, 84, 153

relaxation, 115
 relevance, 17, 27, 62, 74, 100, 102, 190, 237, 238, 239, 241, 242, 311

reliability, 6, 27, 70, 107, 123, 147, 206, 265, 269, 296, 297, 339

relief, 338

rent, 40

replication, 151, 176

Requirements, 155, 169, 170

resistance, 131

resolution, 273, 324

resource allocation, 255

resources, xii, 8, 10, 11, 23, 64, 68, 75, 89, 93, 95, 113, 128, 150, 151, 191, 205, 222, 225, 226, 249, 250, 254, 255, 256, 296, 299, 317

respiratory, 130, 131

respiratory failure, 131

retention, 181

reticulum, 221

retinal disease, 220

retirement, 9, 13

returns, 241, 243, 260

reusability, 18

revenue, 12

rewards, 203

rheumatic fever, 131

rice, 222, 224, 225

risk, 5, 75, 112, 116, 120, 122, 130, 151, 299, 300, 312

risk factors, 299

risk management, 130

RNA, 214, 224

robustness, xi, 34, 39, 41, 44, 45, 54, 165, 169, 170, 171, 173, 174, 175, 266, 268, 274

romantic relationship, 160

routines, 209

routing, 301, 339

rubber, 226

rural areas, 312, 314, 318, 328, 331, 332, 335, 336

S

safety, viii, 81, 86, 89, 90, 91, 92, 93, 94, 126, 286

sales, 4, 34, 35, 36, 37

sampling, 43, 72

satellite, xiv, 17, 309, 311, 324, 326, 330, 341

satisfaction, 4, 106, 137

savings, 17, 137

scaling, 50, 175, 267, 278

school, viii, 81, 82, 83, 84, 85, 86, 88, 89, 91, 92, 93, 94, 95, 159, 160, 201, 202, 208, 243, 317, 319, 320, 332, 333, 336

school community, 95

scores, xi, 54, 106, 118, 119, 121, 179, 180, 183, 184, 239, 241, 302, 303

search, vii, ix, x, xi, xii, 29, 97, 98, 99, 100, 101, 102, 103, 104, 105, 106, 107, 108, 109, 125, 126, 128, 129, 130, 131, 132, 133, 134, 136, 137, 138,

- 139, 157, 158, 173, 179, 180, 185, 192, 193, 196, 197, 202, 208, 210, 215, 220, 235, 236, 237, 238, 239, 241, 242, 245, 246, 247, 278, 332, 333, 334
- search terms, x, 104, 131, 134, 157, 193
- searches, 100, 102, 103, 104, 105, 108, 130, 131, 132, 134, 182, 192, 239, 246
- searching, x, 98, 99, 101, 102, 103, 104, 105, 106, 107, 125, 126, 128, 130, 132, 158, 207, 237, 242
- second generation, 2
- secondary schools, 160, 320
- Secretary of Commerce, 326, 327, 329
- security, xi, 33, 34, 63, 75, 76, 77, 78, 129, 142, 147, 149, 150, 152, 154, 165, 166, 167, 172, 173, 175, 264, 265, 266, 276, 286, 297, 299, 300, 301
- security services, 166
- selecting, 3, 5, 8, 102, 329
- self-definition, 122
- self-efficacy, 118, 119, 121
- semantic search, 103
- semantics, 101, 147, 198
- semi-structured interviews, 5
- Senate, 315, 323, 324, 326, 327, 328, 330, 331, 341, 342
- sensing, xii, 229, 231, 233
- sensitivity, 128, 152, 174, 275
- sensors, xiii, 65, 229, 231, 264, 265
- separation, 180, 255
- sequencing, 106
- service provider, 2, 115, 145, 256, 324, 339
- service quality, 29
- severity, 123
- sex, 160
- shape, xiv, 88, 231, 269, 282, 310, 317
- shaping, 267, 269
- sharing, 11, 12, 34, 39, 47, 48, 49, 113, 134, 137, 144, 166
- shortage, 209
- signal quality, 266, 268
- signal transduction, 186
- signaling pathway, 182, 184, 185, 186
- signalling, xi, 179
- signals, xi, xiii, 41, 45, 58, 165, 175, 263, 264, 265, 267, 268, 270, 273, 287, 288, 306, 307
- simulation, 50, 265, 286, 287, 300, 301, 302
- Singapore, 189, 213, 218, 220
- skills, 85, 86, 92, 121, 123, 126
- sleep apnea, 99
- smoking, 128, 131, 138
- smoking cessation, 128
- smoothing, 247
- SMS, 7, 15, 17
- sociability, 115, 116, 161
- social activities, 113
- social anxiety, 124, 159, 162
- social behaviour, 93
- social capital, ix, 111, 112, 113, 114, 115, 116, 117, 119, 120, 121, 122, 123, 124
- social circle, 121
- social consequences, xiv, 114, 309, 310
- social environment, 129
- social network, 82, 85, 113, 121, 123, 124, 229, 236
- social relations, 111, 112, 121
- social relationships, 111, 112, 121
- social resources, 112
- social responsibility, 93
- social skills, 114
- social structure, 112, 114
- social support, 112, 113, 114, 121
- social workers, 129
- socioeconomic status, 159
- sodium, 223
- software, vii, viii, x, xii, 9, 10, 19, 28, 29, 30, 31, 33, 34, 35, 36, 37, 40, 41, 42, 46, 47, 49, 51, 57, 58, 64, 65, 73, 75, 77, 93, 105, 107, 125, 137, 143, 149, 151, 152, 161, 172, 180, 182, 186, 229, 231, 232, 233, 250, 252, 253, 259, 262, 275, 299
- space, xii, 34, 66, 74, 131, 173, 182, 192, 229, 231, 232, 233, 282, 286, 300
- spam, 239, 240, 247
- spatial processing, 175
- spectrum, 66, 115, 127, 267, 269, 270, 283, 306, 316, 317, 331
- speech, xiii, 30, 107, 127, 197, 212, 263, 264, 265, 267, 268, 269, 270, 271, 273, 274, 275, 276, 278, 286, 287, 288, 292, 297, 298, 300, 301, 302, 305, 306, 307, 308, 319, 341
- speed, xiii, 66, 75, 88, 236, 264, 265, 310, 311, 312, 313, 314, 315, 325, 327, 328, 329, 330, 338
- spelling, 88, 100
- spinal cord, 203
- sports, 104
- stability, 342
- stakeholders, 3, 27, 40
- standardization, 26, 146
- standards, 115, 146, 298, 315, 316, 317, 322, 324
- state planning, 325
- statistics, 136, 163, 193, 194, 208, 313
- stimulus, xiv, xv, 309, 324, 331, 337, 342
- storage, x, xi, xiii, 9, 13, 40, 65, 76, 77, 125, 146, 165, 166, 205, 225, 237, 238, 250, 263, 264, 265, 266, 305
- storage media, 125
- strategic planning, 336
- strategies, 4, 35, 92, 99, 106, 107, 109, 191, 204, 211, 214, 252, 269
- strength, 19, 38, 42, 45, 196, 211, 313
- stressors, 119, 121
- structural equation modeling, 161
- structuring, 206
- students, viii, 81, 82, 83, 84, 85, 86, 87, 88, 89, 90, 91, 92, 93, 94, 130, 134, 153, 159
- subgroups, 161
- subscribers, xiv, 10, 13, 35, 309, 310, 312, 315, 319, 338
- substitutes, 300
- substitution, 173, 268, 303
- subtraction, 50
- superiority, 202

- supervision, 65, 93
 Supreme Court, 338
 surveillance, 35, 126, 130, 137
 Sweden, 84, 93, 179, 313
 Switzerland, 217
 symbiosis, 221
 symbols, 88
 symmetry, 275
 symptoms, 118, 123, 127
 synchronization, xiii, 43, 68, 69, 70, 172, 263, 264, 265, 269, 274, 275, 276, 277, 278, 289, 290, 294, 305
 syndrome, 190, 203, 210
 synthesis, xiii, 263, 264, 270, 271, 272, 273, 274, 282, 284, 285
- T**
- tangible benefits, 83
 tanks, 324
 targets, 91, 303
 task difficulty, 154
 tax credit, 317, 327, 328, 330
 tax incentive, 328, 331
 taxonomy, 58, 107, 127, 193, 211
 teachers, viii, 81, 83, 85, 86, 89, 90, 93, 94
 teaching, 86, 89, 333
 technical assistance, 321
 technology gap, 335
 technology transfer, 209
 teenagers, 85, 103, 104, 142
 telecommunications, xiii, xiv, 30, 309, 310, 312, 314, 315, 318, 319, 320, 321, 322, 324, 332, 335, 337, 338, 341, 342
 Telecommunications Act, 314, 315, 316, 319, 320, 321
 telecommunications services, xiii, 309, 318, 319, 320, 321, 322, 338
 telephone, xiv, 84, 90, 115, 120, 153, 307, 309, 310, 318, 319, 325, 329, 332, 335, 338, 341, 342
 television, 59, 89, 236
 terminals, 66
 territory, 286, 301, 330
 test-retest reliability, 118
 text messaging, 84, 106, 123
 text mining, vii, xii, 189, 190, 191, 192, 194, 195, 199, 202, 203, 206, 207, 212, 215, 217, 218, 219, 220, 221, 227, 228
 theft, 34
 thesaurus, 195
 threats, 33, 47, 142, 149, 150, 152, 166
 threshold, 38, 43, 182, 184, 199, 269, 279, 292, 303, 311
 thresholds, 66, 330
 time frame, 40, 41, 45, 46, 57, 59, 248, 301
 timing, 83, 127
 tissue, 180
 Title I, 319, 324, 326, 338, 339, 340
 Title II, 324, 338, 339
 Title V, 323, 325, 329
 tobacco, 224
 tones, 17
 topology, 65, 147, 297, 302
 tracking, ix, 97, 101, 274
 tracks, 190
 trade, 131, 268
 trade-off, xiii, 18, 54, 263, 264, 269, 279, 294, 297, 298
 tradition, 209
 traffic, 151, 166, 167, 236, 286, 287, 300, 305, 308, 339, 340
 training, 86, 104, 129, 134, 202, 310, 317, 321, 326
 transactions, 10, 71, 76, 77, 78
 transcription, 127, 186, 201, 205, 219, 222, 224, 305
 transcription factors, 186, 201, 205, 222, 224
 transcriptomics, 189
 transcripts, 224, 236
 transformation, 77, 93, 166, 209, 268
 transition, 83, 317, 324
 translation, 175, 237
 transmission, xi, xiii, 68, 150, 165, 166, 176, 263, 264, 265, 266, 270, 276, 277, 278, 279, 281, 287, 289, 290, 293, 296, 299, 300, 301, 302, 305
 transmits, 286, 300, 312
 transport, viii, 61, 62, 64, 65, 66, 67, 68, 70, 71, 72, 75, 76, 78, 79, 150, 265, 284, 297, 301, 305
 trial, 84, 85
 triangulation, 6
 triggers, 72
 trust, 11, 113, 114, 124, 160, 161, 240
 trypsin, 181
 tuberculosis, 127
 type 2 diabetes, 138
- U**
- uncertainty, 94, 130, 182, 204, 220, 314, 322
 unions, 324
 unique features, 137
 United Kingdom, 89, 93, 95
 United States, x, xiv, 93, 125, 154, 159, 161, 309, 310, 311, 313, 314, 316, 319, 330, 335, 341
 updating, 57, 75
 urban areas, 312, 320, 322
 USDA, xiv, 309, 312, 319, 323, 326, 329, 330
- V**
- Valencia, 210, 214, 215, 223
 validation, 78, 124, 180, 184, 185
 variability, 126, 134, 187
 variables, vii, 1, 3, 4, 5, 6, 16, 21, 22, 27, 118, 160, 284
 variance, 118, 119

vector, 172, 176, 192, 193, 199, 203, 212, 215, 284, 285
vehicles, viii, 61, 62, 64, 65, 66, 69, 71, 72, 75, 78
venture capital, 321
Verizon, 339
versatility, 265, 296
victims, 91, 92
village, 92
viral infection, 130
viruses, 299
vision, 29, 63, 64, 68, 69, 231
visions, 330
visual system, 174
visualization, 204, 209, 211, 215, 262
vitamin B6, 221
voice mail, 152, 153
voicing, xiii, 263, 264, 278
voluntary organizations, 124
voting, 195
vulnerability, 116

W

web, vii, ix, xii, 1, 26, 35, 37, 38, 47, 84, 85, 86, 87, 89, 97, 99, 100, 101, 103, 105, 106, 107, 151, 225, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 247, 254, 257, 299
web pages, ix, xii, 97, 100, 106, 151, 235, 236, 237, 238, 239, 240, 241, 242, 243, 244, 245, 246, 257
web sites, 6, 35, 37, 89, 101, 103, 106, 144
White House, 316
wireless networks, 66
wireless systems, 311
workflow, 10, 13, 183
working groups, 146
World Wide Web, 89, 107, 108, 109, 114, 236, 237, 243
worms, 151, 299

X

XML, 19, 147, 148, 150, 226, 261