

Volume 31 Number 2 June 2007

ISSN 0350-5596

# *Informatica*

**An International Journal of Computing  
and Informatics**

Special Issue:

**e-Society**

Guest Editor:

**Maggie McPherson**

**Pedro Isaias**



1977

## EDITORIAL BOARDS, PUBLISHING COUNCIL

Informatica is a journal primarily covering the European computer science and informatics community; scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor from the Editorial Board can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the list of referees. Each paper bears the name of the editor who appointed the referees. Each editor can propose new members for the Editorial Board or referees. Editors and referees inactive for a longer period can be automatically replaced. Changes in the Editorial Board are confirmed by the Executive Editors.

The coordination necessary is made through the Executive Editors who examine the reviews, sort the accepted articles and maintain appropriate international distribution. The Executive Board is appointed by the Society Informatika. Informatica is partially supported by the Slovenian Ministry of Higher Education, Science and Technology.

Each author is guaranteed to receive the reviews of his article. When accepted, publication in Informatica is guaranteed in less than one year after the Executive Editors receive the corrected version of the article.

### Executive Editor – Editor in Chief

Anton P. Železnikar  
Volaričeva 8, Ljubljana, Slovenia  
s51em@lea.hamradio.si  
<http://lea.hamradio.si/~s51em/>

### Executive Associate Editor - Managing Editor

Matjaž Gams, Jožef Stefan Institute  
Jamova 39, 1000 Ljubljana, Slovenia  
Phone: +386 1 4773 900, Fax: +386 1 251 93 85  
matjaz.gams@ijs.si  
<http://dis.ijs.si/mezi/matjaz.html>

### Executive Associate Editor - Deputy Managing Editor

Mitja Luštrek, Jožef Stefan Institute  
mitja.lustrek@ijs.si

### Executive Associate Editor - Technical Editor

Drago Torkar, Jožef Stefan Institute  
Jamova 39, 1000 Ljubljana, Slovenia  
Phone: +386 1 4773 900, Fax: +386 1 251 93 85  
drago.torkar@ijs.si

### Board of Advisors:

Ivan Bratko, Marko Jagodič,  
Tomaž Pisanski, Stanko Strmčnik

### Editorial Board

Anders Ardo (Sweden)  
Juan Carlos Augusto (Argentina)  
Costin Badica (Romania)  
Vladimir Batagelj (Slovenia)  
Francesco Bergadano (Italy)  
Ranjit Biswas (India)  
Marco Botta (Italy)  
Pavel Brazdil (Portugal)  
Andrej Brodnik (Slovenia)  
Ivan Bruha (Canada)  
Wray Buntine (Finland)  
Hubert L. Dreyfus (USA)  
Jozo Dujmović (USA)  
Johann Eder (Austria)  
Vladimir A. Fomichov (Russia)  
Maria Ganzha (Poland)  
Janez Grad (Slovenia)  
Marjan Gušev (Macedonia)  
Dimitris Kanellopoulos (Greece)  
Hiroaki Kitano (Japan)  
Igor Kononenko (Slovenia)  
Miroslav Kubat (USA)  
Ante Lauc (Croatia)  
Jadran Lenarčič (Slovenia)  
Huan Liu (USA)  
Suzana Loskovska (Macedonia)  
Ramon L. de Mantras (Spain)  
Angelo Montanari (Italy)  
Pavol Návrat (Slovakia)  
Jerzy R. Nawrocki (Poland)  
Nadja Nedjah (Brasil)  
Franc Novak (Slovenia)  
Alberto Paoluzzi (Italy)  
Marcin Paprzycki (USA/Poland)  
Gert S. Pedersen (Denmark)  
Ivana Podnar Žarko (Croatia)  
Karl H. Pribram (USA)  
Luc De Raedt (Germany)  
Dejan Raković (Serbia and Montenegro)  
Jean Ramaekers (Belgium)  
Wilhelm Rossak (Germany)  
Ivan Rozman (Slovenia)  
Sugata Sanyal (India)  
Walter Schempp (Germany)  
Johannes Schwinn (Germany)  
Zhongzhi Shi (China)  
Oliviero Stock (Italy)  
Robert Trapp (Austria)  
Terry Winograd (USA)  
Stefan Wrobel (Germany)  
Konrad Wrona (France)  
Xindong Wu (USA)

### Publishing Council:

Ciril Baškovič, Cene Bavec, Jožko Čuk,  
Matjan Krisper, Vladislav Rajkovič, Tatjana Welzer

# Editor's Introduction to the Special Issue

## e-Society

### 1 Introduction

This special issue of the Informatica Journal has extended versions of best papers from the e-Society 2006 conference (<http://www.esociety-conf.org>) co-organised by IADIS – International Association for Development of the Information Society and Trinity College. The e-Society 2006 conference was held in Dublin, Ireland and was a great success.

The IADIS e-Society 2006 conference aimed to address the main issues of concern within the Information Society. This conference covered both the technical as well as the nontechnical aspects of the Information Society. Broad areas of interest were eGovernment /eGovernance, eBusiness / eCommerce, eLearning, eHealth, Information Systems, and Information Management. The IADIS e-Society 2006 Conference had 273 submissions from more than 36 countries in seventy-two topics in all. The papers in this issue represent some of the outstanding papers received during the conference and have been extended for this special issue.

### 2 Overview of the Issue

The e-Society is comprised of several areas. The areas mentioned in this special issue make a contribution for the so called e-Society and the papers mark advancements and progress in its respective fields leading way to a more and more elaborate e-Society. Current papers have been selected among more than 273 submissions from the referred conference and especially from the Information Systems and Information Management areas.

The special issue comprises nine papers in several areas that contribute to the e-Society:

- Enterprise architectures and integration;
- Computational trust and content quality;
- Ubiquitous Computing and RFID;
- Collaborative learning and web communities;
- Electronic Business and CRM;
- Mobility and location-based services (LBS);
- Collaborative learning and intelligent tutoring systems;
- Web security;
- Online security.

The first paper from Amjad Umar, named “Intelligent Decision Support for Architecture and Integration of Next Generation Enterprises” describes an intelligent decision support environment that uses patterns, best practices, inferences, and collaboration for enterprise architecture and integration projects. The referred environment consists of a set of intelligent advisors that

collaborate with each other in a fashion similar to a team of consultants who are working on an integration project.

The second paper from Pierpaolo Dondio and Stephen Barrett, titled “Computational Trust in Web content quality: a comparative evaluation on the Wikipedia Project” presents a method to predict Wikipedia articles trustworthiness based on computational trust techniques and a deep domain-specific analysis. Authors’ assumption is that a deeper understanding of what in general defines high-standard and expertise in domains related to Wikipedia – i.e. content quality in a collaborative environment – mapped onto Wikipedia elements would lead to a complete set of mechanisms to sustain trust in Wikipedia context. Several experiments are presented to exemplify the concept.

The third paper from Hans Weghorn, Hans Peter Großman, Dieter Hellwig, Cahya Kusuma Ratih, Andreas Schmeiser and Heiko Hutschenreiter, titled “Mobile Ticket Control System with RFID Cards for Administering Annual Secret Elections of University Committees” presents a technical solution using a prototype mobile phone, which is equipped with a communication module for contact-less information exchange with the student ID in order to provide a voting administration service. The presented system assures confidentiality, with this mobile vote administration system of Ulm University.

The fourth paper from Stephan Lukosch, titled “Facilitating shared knowledge construction in collaborative learning” presents the web-based collaborative learning platform CURE, used by the German distance learning university to support different collaborative learning scenarios. The paper reports on extensions to CURE that were designed to foster shared knowledge construction and allow learning in a entertaining way. These extensions were developed in a participatory process with the students.

The fifth paper from Andreas Meier and Nicolas Werro, titled “A Fuzzy Classification Model for Online Customers” proposes a fuzzy classification model, in which, customers with similar behaviour and qualifying attributes have similar membership functions and therefore similar customer values. The paper illustrates how webshops can be extended by a fuzzy classification model. This allows webshop administrators to improve customer equity, launch loyalty programs, automate mass customization and personalization issues, and refine marketing campaigns to maximize the real value of the customers.

The sixth paper from Jörg Lonthoff and Erich Ortner titled “Mobile Location-Based Gaming as Driver for Location-Based Services (LBS) – Exemplified by Mobile Hunters” introduces MLBG and the adventure game “Mobile Hunters” – an implemented MLBG that uses the

currently available cellular phone network to create a virtual playing field that represents the real world. This innovative way of playing proves to be a helpful step towards context-based value-added services.

The seventh paper from Eliane Pozzebon, Janette Cardoso, Guilherme Bittencourt and Chihab Hanachi titled “A Group Learning Management Method for Intelligent Tutoring Systems” proposes a group management specification and execution method that seeks a compromise between simple course design and complex adaptive group interaction. This is achieved through an authoring method that proposes predefined scenarios to the author. These scenarios already include complex learning interaction protocols in which student and group models use and update are automatically included. The method adopts ontologies to represent domain and student models, and object Petri nets to specify the group interaction protocols. During execution, the method is supported by a multi-agent architecture.

The eighth paper from Carsten Maple, Geraint Williams and Yong Yue, titled “Reliability, Availability and Security of Wireless Networks in the Community” reviews wireless network protocols, investigates issues of reliability, availability and security when using wireless networks. The paper, by use of a case study, illustrates the issues and importance of implementing secured wireless networks, and shows the significance of the issue. The paper presents a discussion of the case study and a set of recommendations to mitigate the threat.

The ninth and last paper from Graeme Pye and Matthew J. Warren titled “A Model and Framework for Online Security Benchmarking” is also on security and proposes a benchmarking framework to guide both the development and application of security benchmarks created in the first instance, from recognized information technology and information security standards and then their application to the online security measures and policies utilized within online business. Furthermore, the benchmarking framework incorporates a continuous improvement review process to address the relevance of benchmark development over time and the changes in threat focus.

### 3 Acknowledgements

We are grateful, not only to all these authors for extending these interesting papers for this publication, but to the reviewer who have taking the time to give appropriate feedback to achieve the quality we seek for this journal.

Finally, we would like to express our gratitude to members of the e-Society 2006 Programme Committee, with particular mention to the Programme Chair, Frank Bannister, Trinity College, Dublin, Ireland. In addition, special thanks go to the following track chairs, without whom the running of the conference would not have gone as smoothly as it did:

- eBusiness/eCommerce Track  
Mohini Singh, RMIT University, Australia
- eLearning Track  
Inmaculada Arnedillo-Sánchez, Trinity College, Dublin, Ireland
- Information Systems Track  
Hans Weghorn, University of Cooperative Education, Germany
- Information Management Track  
Christian Voigt, University Of South Australia, Australia
- eHealth Track  
Jane Grimson, Trinity College Dublin, Ireland

Maggie McPherson and Pedro Isaias

# Intelligent Decision Support for Architecture and Integration of Next Generation Enterprises

Amjad Umar  
 Information and Communication Systems, Schools of Business  
 Fordham University, New York, USA  
 E-mail: umar@amjadumar.com, website: amjadumar.com

**Keywords:** enterprise architectures, enterprise integration, next generation enterprises, computer aided decision support, PISA, business patterns

**Received:** April 24, 2007

*Architectures and integration of emerging next generation enterprises (NGEs) require a series of complex decisions. This paper describes an intelligent decision support environment that uses patterns, best practices, inferences, and collaboration for enterprise architecture and integration projects. This environment consists of a set of intelligent advisors that collaborate with each other in a fashion similar to a team of consultants who are working on an integration project. It guides the user to appropriate strategic choices, architectural configurations, COTS (commercial off the shelf) packages and project plans. Instead of rushing to automatic code generation from business process models, this paper takes a more cautious approach that is based on practical experience and first concentrates on a decision support environment that will introduce more automation in later iterations.*

*Povzetek: Opisano je okolje za integracijo projektov z uporabo najboljših praks.*

## 1 Introduction

Enterprise architecture and integration projects are complex undertakings especially in the emerging next generation enterprises (NGEs) that rely on deep technology stacks on a daily basis. Specifically, NGEs rely on web-technologies, mobile services, real-time business activity monitoring, agility, self-service, and widely distributed operations to conduct business [31]. Modern architecture and integration projects (AIPs) require participation of, and information sharing between, IT staff, IT managers, consultants, customers, and even business partners. Based on lessons learned from several industrial consultation and academic/research assignments, and review of vendor products and research efforts, we have found that comprehensive decision support environments are needed to lead the participants systematically through the maze of business scenarios, strategic choices involving outsourcing and warehousing, and integration tradeoffs based on cost, performance and security issues. Although environments of this nature are urgently needed, they are virtually non-existing. To fill this gap, we have initiated research on decision support for enterprise integration projects with the following “requirements”:

- **Support different players of the integration projects.** These projects require many decisions that need to be shared, monitored and controlled by different parties. Thus it is important to capture high level business process models and enterprise ontologies for ease of communications, support different views and what-if scenarios for different project

participants, capture different architectural configurations (e.g., outsourcing and data warehousing) that impact integration solutions, and facilitate evaluation of cost, performance and security tradeoffs before implementation.

- **Adopt a breadth first approach.** Development of an environment that supports decisions in different phases of a project should be a *parallel* effort and not an afterthought once all the individual problems have been solved. It is important to provide visibility throughout an integration project as it proceeds through various stages of its life cycle. This is especially crucial for IT managers because they need a total project view.
- **Approach automation systematically.** Automation of integration projects is a desirable goal but it is best to take a cautious approach that first concentrates on capturing/managing knowledge and supporting decisions throughout the project. This will help us better understand what to automate and when, instead of quickly automating the irrelevant activities or generating code from business process models without considering architectural details.
- **Develop a set of collaborating advisors.** Instead of one expert system, it is best to develop a set of intelligent advisors that collaborate with each other in a fashion similar to a team of consultants who are working with each other on real-life integration projects. These advisors should use *patterns* to capture the common and best practices instead of every possible point in the

solution space, heavily rely on *inferences* to reach conclusions instead of asking too many irrelevant questions, and *collaborate* with other consultants to solve complex problems.

- **Educate the users.** Due to the complexity and recurring nature of integration projects, the environment must support e-learning through online tutorials, guides, explanations, and justifications.
- **Stay close to standards and industrial developments.** The environment must closely follow Web Services and results of other efforts such as the INTEROP, Integration Consortium, and OMG (Object Management Group).

We have developed PISA (Planning, Integration, Security, and Administration) environment, an intelligent decision support system that addresses these issues. PISA consists of a set of collaborating advisors that are segmented into two major modules: a) PlanIT (Planner for IT), discussed elsewhere [38], that concentrates on IT planning projects and develops a plan at enterprise level, and b) Architecture and Integration Module (AIM), the focus of this paper, that deals with the architecture and integration issues. Section 2 describes the conceptual model at the core of AIM and Section 3 illustrates AIM through a simple example. PISA is discussed in [40].

## 2 Decision Support Model for Enterprise Architecture and Integration

Figure 1 displays the proposed decision support model for architecture and integration projects (AIPs) of NGEs. This simple model concentrates on four broad stages (enterprise modeling, requirements development, architecture selection, and solution evaluation) and has several important features. First, it specifically addresses

the decision support issues raised previously and also clearly captures the key decisions such as opportunity evaluation and BPO (Business Process Outsourcing), selection of applications and requirements for integration, integrated architecture choices (strategic, technology selection, architectural configuration), and integrated solution evaluation based on cost, performance and security. Second, the model is asynchronous – it captures the real life situation where different activities are typically invoked whenever enough input is available. Third, this model is intelligent because inferences are used heavily between the stages, and the planning knowledgebase provides extensive patterns and COTS (Commercial-Off-The-Shelf) information (explained later). In fact, the user inputs are optional in most cases. Thus this model can produce a quick sketch based on patterns and accumulated knowledge without any user interaction. Fourth, the plans are developed gradually in different stages and captured in the knowledgebase, thus later stages can learn from previous decisions. Fifth, this model conforms to the latest thinking in architectures, especially Model Driven Architecture ([www.omg.org](http://www.omg.org)), which emphasizes the separation and isolation of platform specific issues as much as possible. For example, the enterprise modeling stage is completely independent of platform considerations – these considerations are introduced gradually in later stages. Finally, this model is specifically designed for a computer aided platform where each stage can be supported by an automated consultant (“advisor”), thus different artifacts can be introduced in each advisor and the advisors can collaborate with each other as a team of consultants to solve complex problems (Section 3).

The stages of this model, discussed below, are designed to address the decision support issues raised previously. A more formal treatment of this model with theoretical foundation is presented in Appendix A.

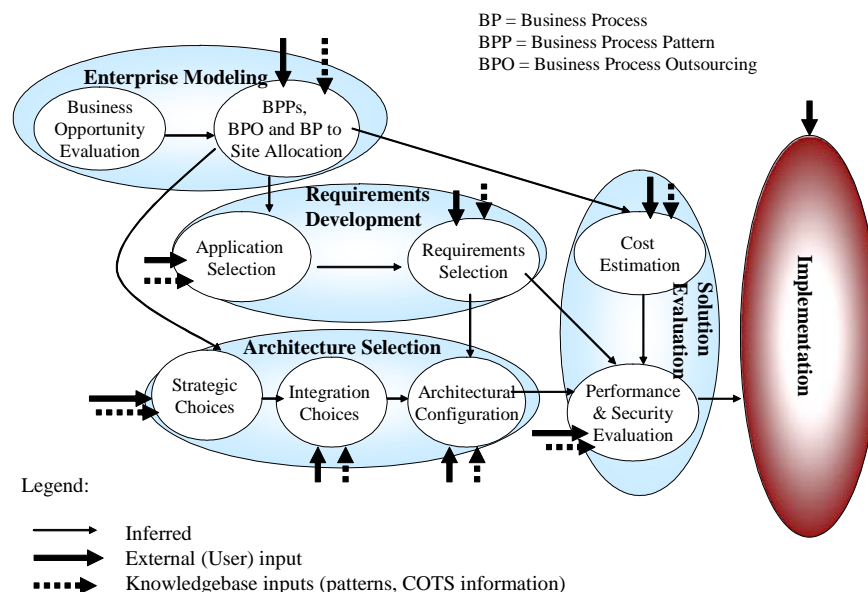


Figure 1: Decision Support Model for Enterprise Architecture and Integration

### 2.1 Stage 1: Enterprise Modelling

The main challenge in this stage is how to quickly build model of an enterprise. We have found that business process patterns (BPPs), shown in **Figure 2**, is a good starting point and of great practical value to identify business scenarios that drive integration projects and to conduct quick sensitivity analysis. For example, a business scenario may be concerned with re-engineering of one business process (e.g., purchasing), revamping of all applications in one functional area (e.g., back-office operations), a combination of the two, or an enterprise application integration (EAI) that encompasses all business processes (BPs) in all areas. **Figure 2** is based on an extensive review of enterprise ontologies [10, 21], business patterns [1, 15, 16] and industrial classifications (e.g., SAP’s Business Maps – www.sap.com). We have mapped this pattern to XML to facilitate high level sensitivity analysis of scenarios such as the following: a) if one BP is eliminated, then what other BPs will be impacted, b) if an application package that supports a BP is replaced with another application, what other applications/BPs will be impacted, c) which application, if replaced, will have the most impact in terms of integration, d) which application, if replaced, will have the least impact in terms of integration. We have created enterprise business patterns of this type for 9 industry segments that include manufacturing, healthcare, telecom, and others; and have mapped them to BPEL representation. We will investigate the use of business processing modeling languages such as BPML, UML 2.0, and others [7, 39] to represent these patterns.

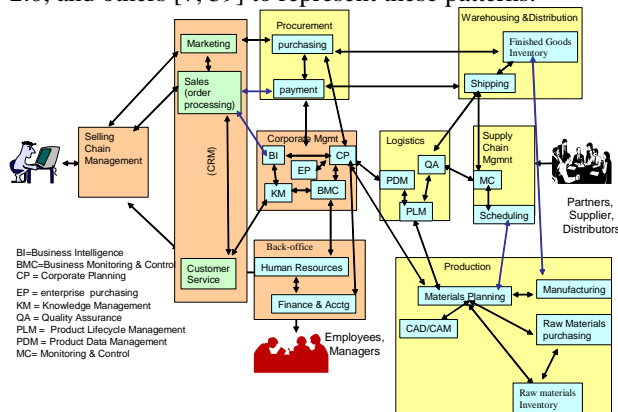


Figure 2: Business Process Pattern (BPP) for a Manufacturing Company

### 2.2 Stage 2: Requirements Development

Requirements development and analysis for integration projects can be considerably expedited through requirements templates that are based on a combination of requirements patterns suggested by Haage and Lappe [12], Ferdinandi [9], and standards

bodies [41]. This template should then be customized by considering factors such as user access devices, back-end apps, B2B apps, transaction value, transaction volume, number of partners, mobility support needed, etc. Several rules can be used to infer functional, interface and integration, mobility, performance and security requirements. For example, transaction volume can impact performance and transaction value can impact security requirements.

### 2.3 Stage 3: Architecture Selection

A three step process shown in **Figure 3** is proposed to help a user to explore various architectural strategies and then develop the integration components (front-end, back-end) of each architectural configuration. This approach generates many more architecture and integration patterns than have been reported in the literature [6, 11, 13]. The first step allows the user to choose between the following strategies for the *target applications* (applications of concern within an integration project):

- **Outsourcing (remote hosting):** decide where the target applications will reside: customer (your) site, service provider site, or a mixture.
- **Access in Place:** integrate without modifying any applications. Just access them by using adapters/mediators.
- **Data Warehouse:** build a "shadow" system to house the frequently accessed data. This is especially useful for BI (Business Intelligence) applications.
- **Migration:** re-architect and transition the target applications gradually

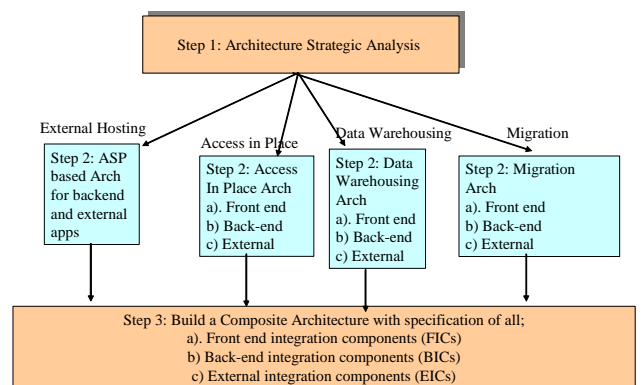


Figure 3: Architecture Steps

The architectures and the technologies needed to integrate these architectures depend on the strategic options of hosting, data warehouses, access in place, and migration. To illustrate these options and their impact on integration, let us consider a situation where a company XCorp chooses to outsource (rent) an online purchasing (OP) system from an ASP (e.g., use Amazon.com’s purchasing system through XCorp) but the inventory and shipping reside at XCorp site. In this



case, remote integration between XCorp applications and ASP is needed. In addition, the OP at the ASP site needs to integrate with the shipping and inventory apps at XCorp. This type of architecture raises interesting questions about security, performance, and interoperability. These questions are not raised if OP is not outsourced. Let us now consider the situation when OP is hosted at XCorp site. In this case, you need to investigate architectures for access-in-place, data warehouses and migration at XCorp site and determine the appropriate Front-end Integration Components (FICs) and Back-end Integration (BICs) for each configuration. For example, data warehouses require extraction, transfer and load (ETL) of back-end data; access-in-place requires adapters/mediators, and migrations typically need a migration gateway. Additional rules are needed to suggest appropriate middleware technologies. For example, if the XCorp order processing application needs to access data from three inventory management systems then remote data access middleware can be used but for many systems an EAI platform is more appropriate.

#### 2.4 Stage 4: Solution Evaluation

Solution evaluation goes into further details by translating the architecture  $A$  into plausible integrated solutions ( $S_1, S_2, \dots, S_n$ ) with appropriate commercial-off-the-shelf (COTS) packages and performance/security/cost evaluations. The first step, COTS selection, allows the user to search the COTS database, part of the knowledgebase, and select the most appropriate solution based on cost constraints, the services needed and the technical interdependencies (for example, a .NET application does not work well in a Linux environment). Due to the complex interrelationships and interdependencies between different layers of technologies, the model shown in Figure 4 is used to analyze the process to process, process to app, app to app, app to middleware, and middleware to middleware interdependencies. Middleware to network and network to network integration is mainly considered in wireless integration projects where seamless services across Wi-Fi, cellular, and Bluetooth are needed [35]. We have specified several COTS selection rules that use this model and heavily rely on the current commercially available application and integration middleware (e.g., EAI software) and their interdependencies [22, 26, 32, 33, 34].

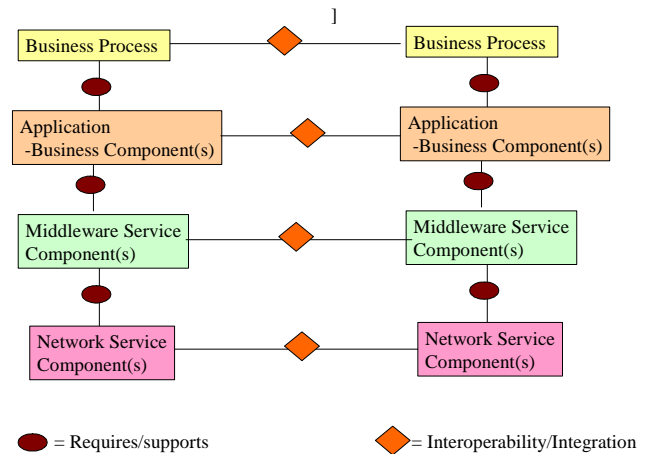


Figure 4: Interdependencies Model

In the next step, the solution  $S_i$ , as a result of COTS selection, is evaluated for performance and security issues. A performance analysis of FICs/BICS is conducted through an analytical model based on Little's formula [18]. For a security analysis of the FICs/BICS, the security issues due are investigated by using attack trees and security patterns [17, 36]. For more detailed security analysis, we are investigating the attack trees developed by the Amenaza's SecureITrees ([www.amenaza.com](http://www.amenaza.com)). For cost estimation, the costs and benefits of all strategies are analyzed before reaching a final decision. The effort needed to integrate systems depends on the number and nature of integration components (FICs, BICs) identified by the IAA. From this, rough estimates of effort and cost can be obtained by using techniques similar to function point analysis.

### 3 PISA-AIM: The Decision Support Tool

The advisors of these modules, illustrated in Figure 1, collaborate with each other to develop IT plans and then analyze the architecture and integration aspects of the plan. Specifically, the PlanIT advisors do the following: the Enterprise Modeler develops a model of an enterprise, the Application Advisor develops an Application plan, the Platform Advisor develops a computing platform plan, the Network Advisor builds a network plan, and the Security Advisor develops a security plan. PlanIT is described elsewhere [38, 40]. The AIM advisors, described in this paper, help the user through life cycle of an integration project and develop an architecture and integration approach. PISA is supported through an extensive knowledgebase that contains a pattern repository, object models, and a COTS database. This knowledgebase also serves as a means of knowledge management by supporting queries and reports and consists of numerous strategies, patterns, COTS tools, project plans, and links to information sources. Common components (COTS Advisor, Project Planner, and Diagram Generator), support the overall system. Let us illustrate the operations of the AIM



modules through example of a manufacturing company (XCorp) introduced previously.

The **Business Problem Explorer (BPE)** helps the user to quickly build business process models and identify business scenarios that drive the specific integration project. After conducting high level analysis of different scenarios, the user proceeds by selecting BPs of interest for further exploration. At this stage, BPE heavily relies on the pattern repository (PR), part of the knowledgebase, to allow the users to review different aspects of the chosen BPs (e.g., process models, use cases, class diagrams, sequence diagrams, service descriptions, etc). These “knowledge chunks” serve as patterns that allow the users to quickly build more detailed and customized models by using tools such as UML.

The **Intelligent Requirements Generator (IRG)** conducts an interview, shown in **Figure 6**, which generates the company specific information for requirements document. The interview also captures the business scenario that drives the requirements. The outputs of this interview plus the knowledge chunks retrieved from the pattern repository are used to populate the various sections of the requirements document. The generated document is a requirements sketch in MS Word format that can be customized by the user.

The **Integrated Architecture Advisor (IAA)** suggests a technical architecture based on the requirements and walks the user through strategic decisions and scenarios of outsourcing, migrations, and data warehousing. A short interview helps the user analyze these strategies and suggest actions based on

the answers to questions such as the following: type of target application (queries, updates), types of resources needed from the back-end applications (data, processes), number of applications you need to integrate with (few, many), application type(s) you need to integrate with (legacy, new, mixture), and data currency requirements (low, medium, high). **Figure 7** shows a sample interview and a strategy recommended by IAA. Given an integration strategy, IAA also suggests a technical solution to support the strategy

The **Integrated Solution Advisor (ISA)** maps the technical architecture produced by IAA to COTS (commercial-off-the-shelf) solutions and evaluates the what-if scenarios for performance, security, and cost tradeoffs. The main results of these analysis show the estimated cost/effort, security and performance for different solutions (*S1, S2,..*) evaluated so far. **Figure 8** shows a partial output produced by ISA that shows performance information. The user can review this information and re-iterate to choose another solution if needed. At conclusion, ISA produces a detailed report that includes a project plan for the selected solution.

As stated previously, the AIM advisors are supported by an extensive knowledgebase of numerous strategies, hundreds of patterns, hundreds of COTS packages, dozens of customizable project plans, and the results of numerous user developed plans, solutions, and interviews stored as object models (OMs). The knowledgebase is organized in three parts: pattern repository, object models, and COTS database. Details of the knowledgebase are not given here due to space limitations.

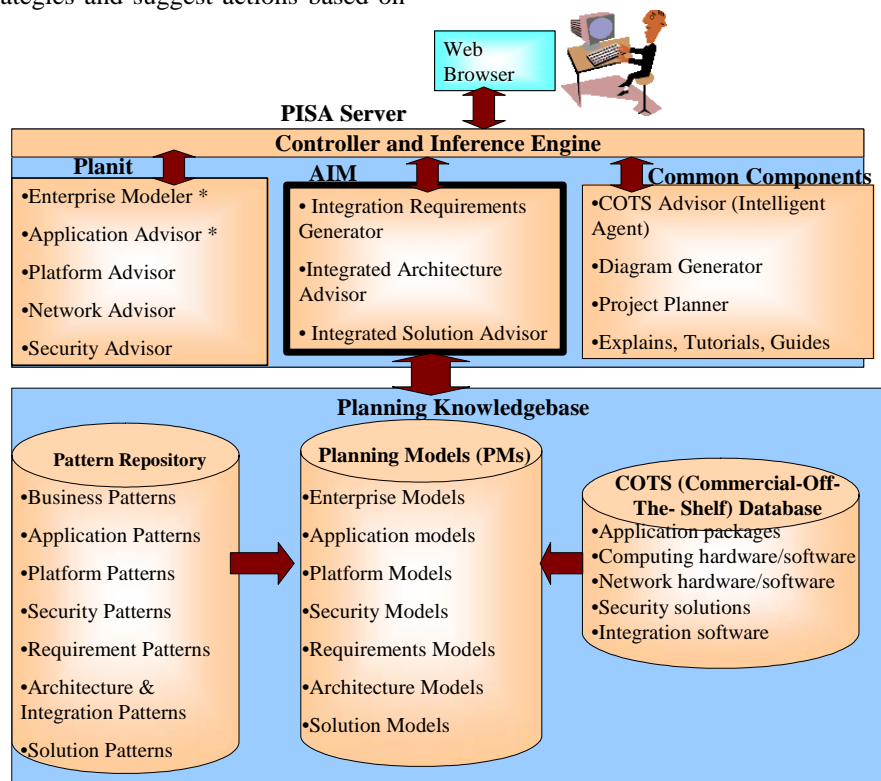


Figure 5: Conceptual View of PISA

Application chosen	Order Processing
User Access	<input type="radio"/> Web Access; <input type="radio"/> Wireless Handheld-Device Access; <input type="radio"/> Voice Access; <input type="radio"/> Application Access;
Number of back-end Applications Interfaced	None
Coupling with back-end Applications	Loose (Asynchronous)
Type of back-end Applications	Monolithic (legacy)
Services needed from back-end Applications	<input checked="" type="radio"/> Data <input type="radio"/> Function <input type="radio"/> Both
Data Translation for Back-end Application	No translation
Number of External Applications Interfaced for B2B	None
Interaction with External Applications	<input checked="" type="radio"/> Direct <input type="radio"/> Indirect (Intermediary)
Coupling with external Applications	Loose (Asynchronous)
Type of external Applications	Monolithic (legacy)
Services needed from External Application	<input checked="" type="radio"/> Data <input type="radio"/> Function <input type="radio"/> Both
Data Translation for External Application	No translation
Value of Transactions Handled by Application	None (\$0)
Volume of Transactions Handled by Application	Small (less than 10 per day)
Boundaries crossed in trade	<input checked="" type="radio"/> Local <input type="radio"/> National <input type="radio"/> International

Figure 6: Sample Interview for Requirement Generation

Application/BC Name	Order Processing
FACTORS TO BE CONSIDERED	CHOICES
Type of target application	Queries (Business Intelligence)
Type of resources needed from the back-end applications	<input checked="" type="radio"/> Data <input type="radio"/> Process <input type="radio"/> Both
Business value of back-end applications	<input checked="" type="radio"/> Low and Medium <input type="radio"/> High
Data currency requirements	Not strict (Daily or Weekly)
Flexibility needed in back-end systems	<input checked="" type="radio"/> Low <input type="radio"/> High
Number of back-end applications	<input checked="" type="radio"/> Few (less than 5) <input type="radio"/> Many (more than 5)
Type of back-end applications	Well structured (well defined interfaces)
Attitude towards external hosting	Not favored
<input type="button" value="Show Recommendations"/>	

Figure 7: Sample Interview for Strategic Analysis

Important Parameters				
Application:	Order Processing	Incoming Message size (assumed) =	100 bytes	
Local Datarate:	10 Mbps Ethernet	Outgoing Message size (assumed) =	500 bytes	
Internetwork (campus-campus) datarate:	T1	Local Server Processing rate =	20 queries / sec (assumed)	
Local Users:	10	Corporate Main Frame server processing rate =	100 queries / sec (assumed)	
Users at other Sites:	10	Utilization = Arrival Rate at server * Service time per message		
<input type="button" value="Calculate Performance Indicators"/>				
Performance Calculations				
Application	Backend Applications	Solution	Response Time (seconds per message)	Server Utilization (single application on server)
Order Processing		Allocated to local site server	0.0006	0.0060
		Allocated to Corporate MainFrame Server	0.0040	0.0460

Figure 8: Sample Output- Performance Evaluation from Integrated Solution Advisor

## 4 Comparison With Related Work and Concluding Remarks

A great deal of literature on enterprise architecture and integration projects (AIPs) has been published in the recent years under headings such as EAI (Enterprise Application Integration), MDA (Model Driven Architectures) and SOA (Service Oriented Architectures). This includes textbooks [19, 32], standards and consortia [42, 43], research papers [44, 45, 46, 47] and industry tools (e.g., IBM eBusiness Framework [15]). To locate an approach and/or a decision support tool that is somewhat comparable to PISA-AIM, we reviewed 11 textbooks, 83 research articles, 52 research and industrial tools, and numerous websites of industry consortia and standards bodies. The results were disappointing because we found no DSS tools for AIPs. Library and general Web searches with keywords like ‘DSS for architecture and integration’ yield articles that discuss how to integrate DSS *with* EAI (e.g., Lee [45]) but nothing about DSS *for* EAI. Specifically, a few models, under the heading of EAI have been proposed in the research literature to highlight different aspects of AIPs. For example, research challenges in EAI (e.g., how integration fits in ebusiness and ecommerce) are presented in [47], different EAI frameworks (e.g., Zachman, ISO Open Data Processing) are compared in [46], and a reference architecture for EAI is developed in [44] to improve understanding of the EAI concepts. However, these models basically illustrate the structural components of EAI systems and do not discuss how to develop a DSS for AIPs.

To conclude, this paper describes an intelligent decision support environment (PISA-AIM) that uses patterns, best practices, inferences, and collaboration for enterprise architecture and integration projects. A system of this nature has not been reported previously in the literature. The PISA-AIM system, operational as a beta version at present, has been used in four consulting assignments (others are in progress) and found it to be a very useful tool in exploring different business scenarios and evaluating the tradeoffs between integration strategies of outsourcing, data warehouses, and access-in-place. The IT managers especially appreciated the opportunity to develop some understanding of the options before embarking on an integration project. In addition, AIM has been used to teach six systems design, enterprise architecture and integration courses so far with very encouraging results. In each course, the students were assigned three projects: 1) manually develop an integrated architecture for an SMB that is going through a major re-engineering effort, 2) use AIM to solve the same problem, and 3) use AIM for a project of their own interest. Most students had a great deal of fun with the third project -- they built models of different businesses and developed integrated architectures by using AIM for “what-if” analysis of different scenarios. We are currently negotiating with

several universities and businesses for additional experiments.

We are using our current experiences and lessons learned to guide future research and development directions. To be realistic, we are following an iterative approach where each iteration adds more depth and intelligence to the advisors, expands the current knowledgebase through improved inference and learning (e.g., case-based reasoning) techniques, and increases automation by generation of sample code where possible (e.g., WSDL for service definitions and front-end/back-end adapters for integration). We intend to extend the business process model to capture more business intelligence based on developments in business ontologies and business process definition languages and environments such as ARIS, BPML, WS-BPEL, XPDL, and UML 2.0 [7, 39]. We are also accumulating more patterns, refining the existing ones, and using them in more situations. An interesting research area is to automatically generate “optimal” solutions by taking advantage of the extensive knowledgebase. For example, a solution could be generated that minimizes cost, performance delays, or certain security threats based on some constraints. Naturally, we want to expand our current focus from SMBs to large businesses.

### Acknowledgement

Members of the NGE Solutions team (Dolores Umar, Kamran Khalid, Adnan Javed, Adalberto Zordan, and Nauman Javed) have significantly contributed to the development of this project by building and testing the PISA-AIM system..

### References

- [1] Adams, J., et al, "Patterns for e-Business: A Strategy for Reuse", IBM Press, October 2001.
- [2] Alexander, C., "The Timeless Way of Building", Oxford University Press, 1979
- [3] Alexander, C. et al , "A Pattern Language", Oxford University Press, 1977
- [4] Brodie, M. and Stonebraker, M., "Migrating Legacy Systems", Morgan Kaufman, 1995
- [5] Boehm, B., and Abts, C., "COTS Integration: Plug and Pray". Computer, vol. 32, no. 1, 1999, pp. 135-138.
- [6] Buschmann, E., et al, "Pattern-Oriented Software Architecture, Vol. 1: A System of Patterns", John Wiley, 1996.
- [7] Business Process Definition Languages -- An Analysis (<http://www.ebpml.org/status.htm>)
- [8] Ericksson, H. and Penker, M., "Business Modeling with UML – Business Patterns at Work", John Wiley, 2000.
- [9] Ferdinandi, P., "A Requirements Pattern: Succeeding in the Internet Economy", Addison-Wesley, January 2002.
- [10] Fox, M. and Gruninger, M., "On Ontologies and Enterprise Modeling", International Conference on Enterprise Integration Modeling, 1997
- [11] Gamma, E., et al, "Design Patterns", Addison Wesley, 1994.
- [12] Hagge, L. and Lappe, K., "Sharing Requirements Engineering Experience Using Patterns", IEEE Software, January-February 2005, pp. 24-31.
- [13] Hohpe, G. and Woolf, B., "Enterprise Integration Patterns : Designing, Building, and Deploying Messaging Solutions", Addison-Wesley, 2003.
- [14] Hull, E., "Requirements Engineering", Springer 2004
- [15] IBM e-Business Framework website <http://www-106.ibm.com/developerworks/patterns/>
- [16] Kalakotta and Robinson, "E-Business 2.0", Wiley, 2002
- [17] Kienzle, D., and Elder, M., "Security Patterns for Web Development", DARPA Contract No: F30602-01-C-0164, June 2001. Weblink: <http://www.scrpt.net/~celer/securitypatterns/final%20report.pdf>.
- [18] Kleinrock, L., "Queuing Systems - Vol. 2", John Wiley, 1976.
- [19] Linticum, D., "Enterprise Application Integration", Addison-Wesley Information Technology Series, 1999
- [20] Maiden, N.A., and Neube, C., "Acquiring COTS Software Selection Requirements", IEEE Software, vol. 15, No. 2, 1998, pp. 46-56.
- [21] Maedche, A., et al, "Ontologies for Enterprise Knowledge Management", IEEE Intelligent Systems, 2003
- [22] Missier, P. and Umar, A., "Representing Knowledge about Modern Software Architectures", IFIP international conference on Knowledge-based systems (April 2001).
- [23] Patterns Web Site, <http://hillside.net/patterns>
- [24] Ovum Report, "Enterprise Application Integrators", Ovum Group, 1999.
- [25] Schneier, B., "Attack Trees", Dr. Dobb's Journal, Dec., 1999.
- [26] Schmidt, D., "Real Time Embedded Systems Research", Institute for Software Integrated Systems, Vanderbilt University, [www.isis.vanderbilt.edu](http://www.isis.vanderbilt.edu), Presentation: August 2004.
- [27] Sommerville, I., "Integrated Requirements Engineering", IEEE Software Magazine, Jan-Feb. 2005, pp. 16-23
- [28] Sneed, H., "Planning the Reengineering of Legacy Systems", IEEE Software, January 1995, pp. 24-34.
- [29] Umar, A., and Missier, P., "A Knowledge-based Decision Support System for Enterprise Integration", First International Workshop on Enterprise Management & Resource Planning Systems, Venice, Nov. 1999.
- [30] Umar, A., and Missier, P., "A Framework for Analyzing Virtual Enterprises", Research Issues in Data Engineering (RIDE) Workshop on Virtual Enterprises, March 1999.

- [31] Umar, A., "IT Infrastructure to Enable Next Generation Enterprises", *ISF (Information Systems Frontiers) Journal*, Volume 7, Number 3, July 2005, pp: 217 - 256.
- [32] Umar, A., "e-Business and Distributed Systems Handbook: Integration Module", NGE Solutions, May 2003, revised August 2004.
- [33] Umar, A., "e-Business and Distributed Systems Handbook: Architecture Module", NGE Solutions, May 2003, revised August 2004.
- [34] Umar, A., "e-Business and Distributed Systems Handbook: Middleware Module", NGE Solutions, May 2003, revised August 2004.
- [35] Umar, A., "Mobile Computing and Wireless Communications", NGE Solutions, July 2004.
- [36] Umar, A., "Information Security and Auditing in the Digital Age", NGE Solutions, May 2003, Revised August 2004.
- [37] Umar, A., "Optimal Program and Data Allocations in Distributed Environments", Ph.D. Dissertation, Univ. of Michigan, 1984
- [38] Umar, A., et al, "Computer Aided Consulting for SMBs", IRMA 2005 Conference, May 2005
- [39] Vemadat, F., "Enterprise Modeling: Objectives, Constructs, and Ontologies", Tutorial posted on INTEROP site (<http://www.interop-noe.org/>).
- [40] Web Link1: PISA Website [www.ngesolutions.com/pisa](http://www.ngesolutions.com/pisa)
- [41] Web Link2: ISO Software Requirements Template, <http://www.12207.com/templates.htm>
- [42] Web Link3: MDA site ([www.omg.org/mda](http://www.omg.org/mda))
- [43] Weblink4: IDEAS (Interoperability Development for Enterprise Application and Software) Roadmap: "State of the Art Summary", [www.ideas-roadmap.net](http://www.ideas-roadmap.net), deliverable D1.1, Mat 2003.
- [44] Chalmeta, R., Campos, C. and Grangel, R., "References architectures for enterprise integration", *Journal of Systems and Software*, Volume 57, Issue 3, 15 July 2001, Pages 175-191.
- [45] Lee, J., Siau, K., and Hong, S., "Enterprise Integration with ERP and EAI", *CACM*, 2003, vol. 46, no. 2, pp. 54-60.
- [46] Losavio, F., Ortega, D., and Pérez, M., "Comparison of EAI Frameworks", *Journal of Object Technology*, Vol. 4, No. 4, May-June 2005, pp.93-114.
- [47] Sharif, A., et al, "Integrating the IS with the enterprise: key EAI research challenges", *The Journal of Enterprise Information Management*, Volume 17 Number 2 2004 pp. 164-170

## Appendix A: Formal AIP Model

Figure 9 shows a formal view of the architecture and integration project (AIP) model shown in Figure 1. The main inputs of this decision model are:

- **Enterprise parameters**  $E = \{e_t, e_s, e_d, e_w, e_m, e_a\}$  where  $e_t$  = company type (e.g., manufacturing),  $e_s$  = company size in terms of number of employees,

$e_d$  = company distribution (e.g., local, regional, international),  $e_w$  = reliance on the web to conduct business,  $e_m$  = reliance on mobility to conduct business, and  $e_a$  = reliance on agility (e.g., on-demand services) to conduct business. These core parameters, as shown in Table 1, impact all stages of the AIP model. Thus by specifying or changing these few parameters, the planner can quickly investigate a new business scenario by generating application plans, requirements documents, and architectural configurations.

- **Patterns**, introduced by Alexander [2, 3], play an important role in this process. We specifically use business process patterns (BPPs) [1, 16], application and requirement patterns (APs) [9, 12], architecture patterns (HPs) [6], and integration patterns (IPs) [13] to provide generic sketches that are customized to produce a company scenario specific solution. The planning knowledgebase, shown and discussed in Section 3, contains these patterns and other models needed to support the decision model. Customization of patterns is based on the enterprise parameters  $E$ , stage specific inputs and  $PM$ s accumulated in the planning knowledgebase.
- **Local input parameters**  $W, X, Y$  and  $Z$  that are provided by the user in the four stages of the AIP model (see Figure 9). These parameters are used to provide the stage specific information, if needed. AIM automatically provides a set of reasonable defaults for these parameters based on patterns. Thus, if the user is in a hurry, a default solution sketch can be created extremely quickly (within 10 minutes). However, the user may override the default parameters for more customized analysis.

The main output produced by this process is the **planning model**  $PM$ , a set which consists of several subsets where each subset represents results of an AIP stage. In the beginning,  $PM$  is a simple sketch that is successively enriched as the user progresses through different stages of the proposed AIP model. At the conclusion of an interview, a complete company specific plan is represented in the  $PM$ , i.e.,  $PM = [M, R, A, S]$  where  $M, R, A$ , and  $S$  represent the enterprise model, the application requirements plan, the integrated architecture configuration, and the integrated solution, respectively.

Given the inputs  $E$  that represent enterprise parameters, ( $W, X, Y, Z$ ) that represent stage specific information and the patterns ( $BPP, AP, HP, IP$ ), the following relationships are used to create a planning model  $PM = [M, R, A, S]$  that represents the complete AIP:

$$\begin{aligned} M &= f(E, W, BPP) \\ R &= f(E, M, X, AP) \\ A &= f(E, R, Y, HP) \\ S &= f(E, A, Z, IP) \end{aligned}$$

The planning model  $PM$ , as shown in Figure 9, is constructed successively in various stages of the model. Note that the  $E$  impacts all outputs. Table 1 displays details of the interrelationships between  $E$



and the stages of the AIP model and presents the theoretical foundation of the proposed AIP model. Each cell of this table has been converted into a set of rules. Specifically, each function  $f$ , e.g.,  $f(E, W, BPP)$ , represents a set of rules that are discussed qualitatively in the main body of the paper. These rules represent the core inference net used by AIM.

Note that  $E$  is the *only* input required from the user, all others use defaults and/or patterns from the knowledgebase. Thus the planner can adjust a few parameters in  $E$  (e.g., introduce on-demand services, increase number of sites, etc) for new business scenarios and quickly run through all stages of an AIP.

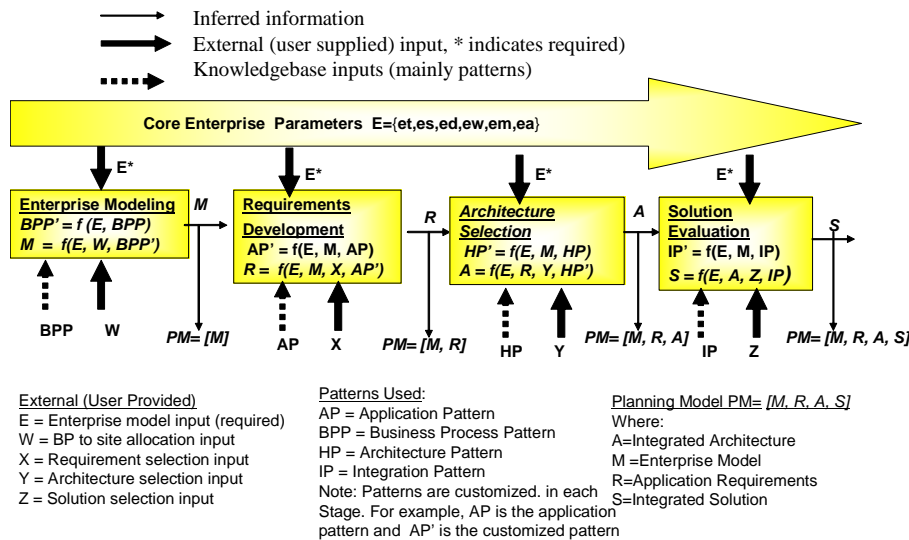


Figure 9: Formal View of the Architecture and Integration Project (AIP) Decision Model

Table 1: Core Parameters and their Impact on Integration Planning

Core Parameters $E=\{e_t, e_s, e_d, e_w, e_m, e_a\}$	Enterprise Model M Produced	Application Requirements Plan R Produced	Integrated Architecture plan A Produced	Integrated Solution Plan S Produced
$e_t$ = Company Type (e.g., manufacturing)	$e_t \rightarrow BPP \rightarrow M$ $e_t$ determines business process pattern (BPP) which is the foundation of enterprise model M	$e_t \rightarrow BPP \rightarrow AP \rightarrow R$ $e_t$ and BPP determine the application patterns (APs) that specify the type of application packages needed (e.g., payment package for payment). AP determines R	$e_t \rightarrow BPP \rightarrow R \rightarrow A$ $e_t$ and BPP influences the type of interfaces (batch, interactive) between BPs and determines requirements model R. R determines A.	$e_t \rightarrow A \rightarrow S$ $e_t$ impacts security exposure (e.g., financial BPs need higher security) but does not have direct impact on cost and performance.
$e_s$ =Company Size: Number of Employees (Low, Medium, High)	$e_s \rightarrow BPP \rightarrow M$ $e_s$ customizes business process pattern (BPP) because larger companies may have additional BPs. BPP is used to build M	$e_s \rightarrow AP \rightarrow R$ $e_s$ influences application pattern AP (e.g., small companies need a few BPs to be automated through applications). AP drives R.	$e_s \rightarrow R \rightarrow A$ $e_s$ influences requirements model R (e.g., integrated architectures for large scale systems use expensive EAI platforms).	$e_s \rightarrow S$ $e_s$ impacts security (larger companies have higher impact) and cost (more users need expensive solutions).
$e_d$ =Company Distribution: Local, Regional, International	$e_d \rightarrow BPP \rightarrow M$ $e_d$ customizes business process pattern (BPP) because highly distributed companies may have additional BPs for international trade. BPP is used to build M	$e_d \rightarrow AP \rightarrow R$ $e_d$ determines the application pattern AP (e.g., transaction processing and workflow packages are needed for compliance for international laws, taxes and standards, e.g., I18N).	$e_d \rightarrow R \rightarrow A$ $e_d$ helps determine the requirements model R to reflect sophistication of applications needed to support B2B trade (robust supply chain management for many partners).	$e_d \rightarrow S$ $e_d$ impacts security exposure (B2B, international) and could impact cost and performance (widely distributed systems can be more expensive)
$e_w$ =Reliance on Web (Low, Medium, High)	$e_w \rightarrow BPP \rightarrow M$ High value of $e_w$ customizes BPP to include more eBusiness BPs and BPP is used to build M	$e_w \rightarrow AP \rightarrow R$ High value of $e_w$ indicates strong eBusiness and possibly real-time business activity monitoring (RBAM) support (this influence AP)	$e_w \rightarrow R \rightarrow A$ High value of $e_w$ requires high degrees of Web integration (this is reflected in requirements model R). R determines A.	$e_w \rightarrow S$ High value of $e_w$ impacts security exposure, cost and performance of a solution.
$e_m$ =Reliance on Mobility (Low, Medium, High)	$e_m \rightarrow BPP \rightarrow M$ High value of $e_m$ introduces Mobile Business BPs in BPP and BPP is used to build	$e_m \rightarrow AP \rightarrow R$ High value of $e_m$ requires M-Business applications (e.g., M-CRM, M-SCM, M-Portal)	$e_m \rightarrow R \rightarrow A$ High value of $e_m$ requires front-end integration (this is reflected in	$e_m \rightarrow S$ High value of $e_m$ requires stronger security, increases cost

	enterprise model M	for high mobility and location-based services.	requirements model R). R determines A.	and raises performance issues due to wireless. .
<b><math>e_a</math> =Reliance on Agility (e.g., On-demand services)</b>	$e_a \rightarrow BPP \rightarrow M$ High value of $e_a$ introduces new BPs for agility in BPP and BPP determines M	$e_a \rightarrow AP \rightarrow R$ High value of $e_a$ requires component-based apps for flexibility(reflected in AP).	$e_a \rightarrow R \rightarrow A$ High value of $e_a$ requires SOA (reflected in R). R determines A.	$e_a \rightarrow BPP \rightarrow S$ High value of $e_a$ increases security exposure

# Computational Trust in Web Content Quality: A Comparative Evaluation on the Wikipedia Project

Pierpaolo Dondio and Stephen Barrett

Trinity College Dublin, School of Computer Science and Statistics, Dublin, Ireland

E-mail: {dondiop, stepehn.barrett}@cs.tcd.ie

**Keywords:** computational trust, Wikipedia, content-quality

**Received:** March 17, 2007

*The problem of identifying useful and trustworthy information on the World Wide Web is becoming increasingly acute as new tools such as wikis and blogs simplify and democratize publication. It is not hard to predict that in the future the direct reliance on this material will expand and the problem of evaluating the trustworthiness of this kind of content become crucial. The Wikipedia project represents the most successful and discussed example of such online resources. In this paper we present a method to predict Wikipedia articles trustworthiness based on computational trust techniques and a deep domain-specific analysis. Our assumption is that a deeper understanding of what in general defines high-standard and expertise in domains related to Wikipedia – i.e. content quality in a collaborative environment – mapped onto Wikipedia elements would lead to a complete set of mechanisms to sustain trust in Wikipedia context. We present a series of experiment. The first is a study-case over a specific category of articles; the second is an evaluation over 8 000 articles representing 65% of the overall Wikipedia editing activity. We report encouraging results on the automated evaluation of Wikipedia content using our domain-specific expertise method. Finally, in order to appraise the value added by using domain-specific expertise, we compare our results with the ones obtained with a pre-processed cluster analysis, where complex expertise is mostly replaced by training and automatic classification of common features.*

*Povzetek: Ocenjena je stopnja zaupanja v strani v Wikipediji.*

## 1 Introduction

In the famous 1996 article *Today's WWW, Tomorrow's MMM: The specter of multimedia mediocrity* [1] Cioleck predicted a seriously negative future for online content quality by describing the World Wide Web (WWW) as “*a nebulous, ever-changing multitude of computer sites that house continually changing chunks of multimedia information, the global sum of the uncoordinated activities of several hundreds of thousands of people*”. Thus, the WWW may come to be known as the MMM (MultiMedia Mediocrity). Despite this vision, it is not hard to predict that the potential and the growth of the Web as a source of information and knowledge will increase rapidly. The Wikipedia project, started in January 2001, represents one of the most successful and discussed example of such phenomenon, an example of *collective knowledge*, a concept that is often lauded as the next step toward truth in online media. Wikipedia is a global online encyclopaedia, entirely written collaboratively by an open community of users, it now supports one million registered user, delivers 900.000 articles in its English version alone, and it is one of the ten most visited web sites.

On one hand, recent exceptional cases have brought to the attention the question of Wikipedia trustworthiness. In an article published on the 29th of

November in USA Today [2], Seigenthaler, a former administrative assistant to Robert Kennedy, wrote about his anguish after learning about a false Wikipedia entry that listed him as having been briefly suspected of involvement in the assassinations of both John Kennedy and Robert Kennedy. The 78-year-old Seigenthaler got Wikipedia founder Jimmy Wales to delete the defamatory information in October. Unfortunately, that was four months after the original posting. The news was further proof that Wikipedia has no accountability and no place in the world of serious information gathering [2].

On the other hand, Wikipedia is not only being negatively discussed. In December 2005, a detailed analysis carried out by the magazine Nature [3] compared the accuracy of Wikipedia against the Encyclopaedia Britannica. Nature identified a set of 42 articles, covering a broad range of scientific disciplines, and sent them to relevant experts for peer review. The results are encouraging: the investigation suggests that Britannica's advantage may not be great, at least when it comes to science entries. The difference in accuracy was not particularly great: the average science entry in Wikipedia contained around four inaccuracies; Britannica, about three. Reviewers also found many factual errors, omissions or misleading statements: 162 and 123 in Wikipedia and Britannica respectively.



This paper seeks to face the problem of the trustworthiness of Wikipedia by using a computational trust approach; our goal is to set up an automatic and transparent mechanism able to estimate the trustworthiness of Wikipedia articles. In the next section 2 we review related work on trust and content quality issues; in section 3 we argue that, due to the fast changing nature of articles, it is difficult to apply the trust approaches proposed in related work. In section 4 this discussion will lead us to introduce our domain-specific approach, that starts from an in-depth analysis of content quality and collaborative editing domains to give us a better understanding of what can support trust in these two Wikipedia related fields. In section 5 we map conclusions of the previous section onto elements extracted directly from Wikipedia in order to define a new set of sources of trust evidence. In section 6 we present our evaluation conducted through three different experiments. The first is a study case over 250 articles from the single category “country of the world”; the second is an extension conducted over almost 8,000 Wikipedia. In the third experiment we perform a cluster analysis to isolate article of low and great quality and we compare the results obtained with this implicit approach to the previous one based explicitly on domain expertise. Section 7 will collect our conclusions and future work.

## 2 Related Works

There are many definitions of the human notion trust in a wide range of domains from sociology, psychology to political and business science, and these definitions may even change when the application domains change. For example, Romano’s definition tries to encompass the previous work in all these domains: “trust is a subjective assessment of another’s influence in terms of the extent of one’s perceptions about the quality and significance of another’s impact over one’s outcomes in a given situation, such that one’s expectation of, openness to, and inclination toward such influence provide a sense of control over the potential outcomes of the situation.”[4].

However, the terms trust/trusted/trustworthy, which appear in the traditional computer security literature, are not grounded on social science and often correspond to an implicit element of trust. Blaze et al [5] first introduced “decentralized trust management” to separate trust management from applications. PolicyMaker [6] introduced the fundamental concepts of policy, credential, and trust relationship. Terzis et al. [7] have argued that the model of trust management [5,6] still relies on an implicit notion of trust because it only describes “a way of exploiting established trust relationships for distributed security policy management without determining how these relationships are formed”.

Computational trust was first defined by S. Marsh [8], as a new technique able to make agents less vulnerable in their behaviour in a computing world that appears to be malicious rather than cooperative, and thus to allow interaction and cooperation where previously there could be none. A computed trust value in an entity may be seen as the digital representation of the trustworthiness or

level of trust in the entity under consideration. The EU project SECURE [9] represents an example of a trust engine that uses evidence to compute trust values in entities and corresponds to evidence-based trust management systems. Evidence encompasses outcome observations, recommendations and reputation. Depending on the application domain, a few types of evidence may be more weighted in the computation than other types. When recommendations are used, a social network can be reconstructed. Golbeck [10] studied the problem of propagating trust value in social networks, by proposing an extension of the FOAF vocabulary [11] and algorithms to propagate trust values estimated by users rather than computed based on a clear count of pieces of evidence. Recently, even new types of evidence have been proposed to compute trust values. For example, Ziegler and Golbeck [12] studied interesting correlation between similarity and trust among social network users: there is indication that similarity may be evidence of trust. In SECURE, evidence is used to select which trust profile should be given to an entity. Thus similar evidence should lead to similar profile selection. However, once again, as for human set trust value, it is difficult to clearly estimate people similarity based on a clear count of pieces of evidence. However, the whole SECURE framework may not be generic enough to be used with abstract or complex new types of trust evidence. In fact, in this paper, we extracted a few types of evidence present in Wikipedia (detailed in the next sections) that did not fit well with the SECURE framework and we had to build our own computational engine.

We think that our approach to deeply study the domain of application and then extract the types of trust evidence from the domain is related to the approach done in expert systems where the knowledge engineer interacts with an expert in the domain to acquire the needed knowledge to build the expert system for the application domain. In this paper, we focus on trust computation for content quality and Bucher [14] clearly motivates our contribution in this paper because he argues that on the Internet “we no longer have an expert system to which we can assign management of information quality”.

We finish this section by two last computational projects related to content quality in a decentralised publishing system. Huang and Fox in [15] propose a metadata-based approach to determine the origin and validity of information on the Web.

## 3 The problem of Wikipedia Articles Trustworthiness and our method

Wikipedia shows intrinsic characteristics that make the utilization of trust solutions challenging. The main feature of Wikipedia, appointed as one of its strongest attribute, is the speed at which it can be updated. The most visited and edited articles reach an average editing rate of 50 modifications per day, while articles related to recent news can reach the number of hundreds of modifications. This aspect affects the validity of several trust techniques.

Human-based trust tools like feedback and recommendation systems require time to work properly, suffering from a well know ramp-up problem [16]. This is a hypothesis that clashes with Wikipedia, where pages change rapidly and recommendations could dramatically lose meaning. Moreover, the growing numbers of articles and their increasing fragmentation require an increasing number of ratings to keep recommendations significant.

Past-evidence trust paradigm relies on the hypothesis that the trustor entity has enough past interactions with the trustee to collect significant evidence. In Wikipedia the fact that past versions of a page are not relevant for assessing present trustworthiness and the changing nature of articles makes it difficult to compute trust values based on past evidences. In general, user past-experience with a Web site is only at 14<sup>th</sup> position among the criteria used to assess the quality of a Web site with an incidence of 4.6% [17]. We conclude that a mechanism to evaluate articles trustworthiness relying exclusively on their present state is required.

Our method starts from the assumption that a deeper understanding of the domains involved in Wikipedia, namely the content quality domain and the collaborative editing domain, will help us to identify trust evidence we required to set up an automatic trust computation. The procedure we followed can be summarized in a 4-stage process.

We begin by modelling the application under analysis (i.e. Wikipedia). The output of the *modelling Phase* should be a complete model showing the entities involved, their relationships, the properties and methods for interacting: here we will find out trust dynamics. It is also necessary to produce a valid theory of domain-compatible trust, which is a set of assertions about what behaviours should be considered trustworthy in that domain. This phase, referred as *theories analyser*, is concerned with the preparation of a theoretical trust model reasonable for that domain. To reach this goal a knowledge-based analysis is done to incorporate general theories of Trust, whose applicability in that domain must be studied, joined with peculiar domain-theories that are considered a good description of high-quality and trustworthy output in that domain.

The output is a domain compatible trust theory that acts like a sieve we apply to the application model in order to extract elements useful to support trust computations. This mapping between application model and domain-specific trust theory is referred as *trust identifier*. These elements, opportunely combined, will be the evidence used for the next phase, our trust computation. The more an entity (a Wikipedia page) shows properties linked to these proven domain-specific theories, the more is trustworthy. In this sense, our method is an evidence-based methodology where evidences are gathered using domain related theories.

In other words, after understanding what brings trust in those domains, we mapped these sources of evidence into Wikipedia elements that we previously isolated by defining a detailed model of the application. This resulting new set of pieces of evidence, extracted directly from Wikipedia, allow us to compute trust, since it relies

on proven domains' expertise. In the next three paragraphs we will apply our method: *theories analyzer* phase (section 4), *modelling phase* and *trust identifier* (section 5) and our expertise-based *trust computation* in the evaluation section.

## 4 Wikipedia Domain Analysis

In this section we identify a trust theory derived from domain-specific expertise relevant to Wikipedia, the *theories analyzer* phase of our method. Wikipedia is a combination of two relevant areas involved in Wikipedia: the *content quality* domain and *collaborative editing* domains. In this section, we analyse what can bring high quality in these two domains. The quality of online content is a critical problem faced by many institutions. Alexander [18] underlines how information quality is a slippery subject, but it proposes hallmark of what is consistently good information. He identified three basic requirements: objectivity, completeness and pluralism. The first requirement guarantees that the information is unbiased, the second assesses that the information should not be incomplete, the third stresses the importance of avoiding situations in which information is restricted to a particular viewpoint. University of Berkeley proposes a practical evaluation method [19] that stresses the importance of considering authorship, timeliness, accuracy, permanence and presentation. Authorship stresses the importance of collecting information on the authors of the information, accuracy deals with how the information can be considered good, reviewed, well referenced and if it is comparable to similar other Web content, in order to check if it is compliant to a standard. Timeliness considers how the information has changed during time: its date of creation, its currency and the rate of its update; permanence stresses how the information is transitory or stable.

In a study already cited [17], presentation resulted in the most important evaluation criterion with an incidence of 46%. The Persuasive Technology Lab has been running the Stanford Web Credibility Research since 1997 to identify which are the sources of credibility and expertise in Web content. Among the most well-known results are the ten guidelines for Web credibility [20], compiled to summarize what brings credibility and trust in a Web site. The guidelines confirm what we described so far and again they emphasize the importance of the non anonymity of the authors, the presence of references, the importance of the layout, the constant updating and they underline how typographical errors and broken links, no matter how small they could be, strongly decrease trust and represent evidence of lack of accuracy.

Beside content quality domain, Wikipedia cannot be understood if we do not take into consideration that it is done entirely in a collaborative way. Researches in collaborative working [21] help us to define a particular behaviour strongly involved in Wikipedia dynamics, the balance in the editing process. A collaborative environment is more effective when there is a kind of emerging leadership among the group; the leadership is able to give a direction to the editing process and avoid

fragmentation of the information provided. Anyway, this leadership should not be represented by one or two single users to avoid the risk of lack of pluralism and the loss of collaborative benefits like merging different expertises and points of view. We summarize our analysis with the prepositions shown in table 1: in the first column are theoretical propositions affecting trust, second column lists the domains from which each preposition was taken.

Proposition 1 covers the authorship problem. Proposition 2 derives from the accuracy issues. Proposition 3, 4 and 5 underline the importance that the article should have a sense of unity, even if written by more than one author. Proposition 7 underlines the fact that a good article is constantly controlled and reviewed by a reasonable high number of authors. Proposition 8 stresses the stability of the article: a stable text means that it is well accepted, it reached a consensus among the authors and its content is almost complete. Proposition 9 emphasizes the risk, especially for historical or political issues, that different authors may express personal opinions instead of facts, leading to a subjective article or controversial disputes among users. In order to have meaning, these prepositions need to be considered together with their interrelationships along with some conditions. For example, the length of an article needs to be evaluated in relation to the popularity and importance of its subjects, to understand if the article is too short, superficial or too detailed; the stability of an article has no meaning if the article is rarely edited, since it could be stable because it is not taken in consideration rather than because it is complete.

**Table 1.** A Trust domain-compatible theory. CQ is Content Quality domain and CE is Collaborative Editing domain.

Propositions about Trustworthiness of articles (T). T increases if the article...	Domain of origin
1 was written by expert and identifiable authors	CQ
2 has similar features or it is compliant to a standard in its category	CQ
3 there is a clear leadership/direction in the group directing the editing process and acting like a reference	CE
4 there is no dictatorship effect, which means that most of the editing reflects one person's view.	CQ/CE
5 the fragmentation of the contributions is limited: there is more cohesion than dissonance among authors	CE
6 has good balance among its sections, the right degree of details, it contains images if needed, it has a varied sentence structure, rhythm and length	CQ
7 is constantly visited and reviewed by authors	CQ
8 is stable	CQ
9 use a neutral point of view	CQ
10 the article is well referenced	CQ

### 5 Mapping Theories onto Wikipedia

In this section we produce a model of Wikipedia and

we map over the model the domain-specific trust we identified in the previous section. We first need a model of Wikipedia in order to extract elements useful for our purpose. Wikipedia has been designed so that any past modification, along with information about the editor, is accessible. This transparency, that by itself gives an implicit sense of trust, allows us to collect all the information and elements needed.

Our Wikipedia model is composed of two principal objects (Wiki Article and Wiki User) and a number of supporting objects, as depicted in fig. 1. Since each user has a personal page, user can be treated as an article with some editing methods like creating, modifying and deleting article or uploading images. An article contains the main text page (class wiki page) and the talk page, where users can add comments and judgments on the article. Wiki pages include properties such as its length, a count of the number of sections, images, external links, notes, and references. Each page has a history page associated, containing a complete list of all modifications. A modification contains information on User, date and time and article text version.

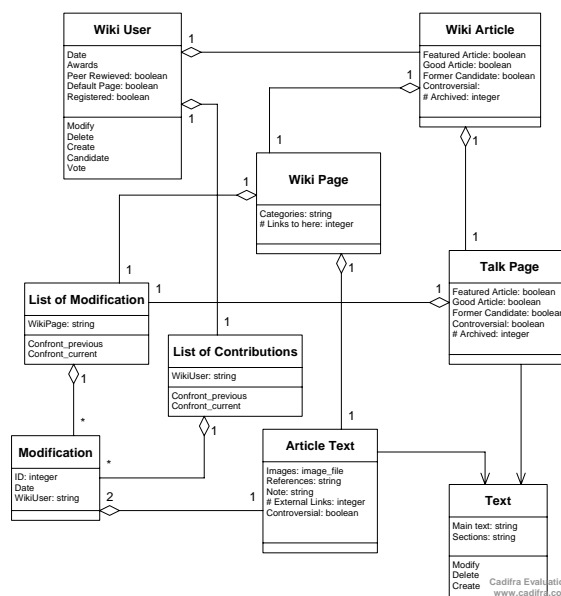


Figure 1 The Wikipedia UML model

The community of users can modify articles or adding discussion on article's topic (the talk page for that article).

We are now ready to map the proposition listed in table 1 onto elements of our Wikipedia model. We remind that the output of this phase will be a set of trust evidence to be used in our trust computation. In general, computing trust using a domain-specific analysis means to aggregate some elements of the application into formulae, that in general maybe be not intuitive and elaborated, in order to model more accurately as possible expert conclusions.

By mapping the conclusions achieved in section 4 - the ten propositions - over our model we identified about 50 sources of trust evidence classified in 6 macro-areas:

Quality of User, User Distribution and Leadership, Stability, Controllability, Quality of editing and Importance of an article. We now analyses as an example two of the six macro-areas.

### 5.1 User’s Distribution/Leadership (p. 3,9)

Given an article  $w$  in the set  $W$  of all Wikipedia articles we define:

$$U(w)$$

as the ordered set of all users  $u$  that contributed to the article  $w$ . Thus, the set  $U$  is a property of a single article. Then we define a set of formulas that are properties of a single user  $u$ .

$$E(u, w) : U \otimes W \rightarrow \mathfrak{S}.$$

Or only  $E(u)$ , the number of edits for user  $u$  for article  $w$ . We define:

$$T(w) : w \rightarrow \mathfrak{S},$$

the total number of edits for article  $w$ . We then define

$$P(n) : [0..1] \rightarrow \mathfrak{S} \quad P(n) = \sum_{U_a} E(u)$$

Table 2. Users Distribution factors

Trust Factors	Comments
Average of $E$	Average number of edits per user.
Standard Deviation of $E$	Standard deviation of edits
$\frac{P(n)}{T}$	% of edits produced by the most active users
$\frac{Pe(t)}{T}$	% of edits produced by users with more than $n$ edit for that article
Number of discussions (talk edit)	It represents how much an article is discussed
Blocked (the article cannot be edited)	The article is blocked due to vandalism. Provided by Wikipedia
Controversial	Article’s topic is controversial. Provided by Wikipedia

Where  $U_a$  is the set of  $n\%$  most active users in  $U(w)$   $P(n)$ , given a normalized percentage  $n$ , returns the number of edits done by the top  $n\%$  most active users among the set  $U(w)$ . Similar to  $P(n)$  is

$$Pe(n) : \mathfrak{S} \rightarrow \mathfrak{S}$$

$$Pe(n) = \sum_{U_n} E(u)$$

$$U_n = \{u \in U \mid E(u) > n\}$$

that, given a number of edits  $n$ , represent the number of edits done by users with more than  $n$  edits. The different between  $P$  and  $Pe$  is that  $P$  considers the most active users in relation to the set of users contributing to the article, while  $Pe$  considers the most active users in relation to an absolute number of edit  $n$ .

We explain the meaning of the functions defined:  $P(n)/T$  tells us how much of the article has been done by a subset of users. If we pose  $n=5$  and we obtain:

$$\frac{P(5)}{T} = 0.45$$

this means that the 45% of the edits have been done by the top 5% most active users. If the value is low the article leadership is low, if it is high it means that a relatively small group of users is responsible for most of

the editing. We introduced the function  $Pe(n)/T$  to evaluate leadership from a complementary point of view.  $Pe(n)/T$  is the percentage of edits done by users that did more than  $n$  edits for the article. If we pose  $n=3$  and we obtain:

$$\frac{Pe(3)}{T} = 0.78$$

This means that 78% of the edits were done by users with more than 3 edits and only 22% by users that did 1,2 or 3 edits. Thus,  $1-Pe(n)/T$  with  $n$  small (typically 3) indicates how much of the editing’s process was done by occasional users, with a few edits. Thus, it can represent a measurement of the fragmentation of the editing process. The average and standard deviation of the function  $E(u)$  (total edits per user  $u$ ) reinforces the leadership as well: average close to 1 means high fragmentation, high standard deviation means high leadership. The last three factors are a clue of how much an article is discussed and controversial.

### 5.2 Stability (propositions 8)

We define the function

$$N(t) : t \rightarrow \mathfrak{S}$$

That gives the number of edits done at time  $t$ . Then we define:

$$Et(t) = \sum_t^P N(t)$$

that, given time  $t$  it gives the number of edits done from time  $t$  to the present time  $P$ . We than define

$$Txt(t) : t - > \mathfrak{S}$$

that gives the number of words that are different form the version at time  $t$  and the current one. We define  $U$  as the largest period of time for that article, i.e. its age. We define  $L$  as the number of words in the current version.

Table 3. Article’s stability factors.

Trust Factors	Comments
$\frac{Et(t)}{U}$	Percentage of edits from time $t$
$\frac{Txt(t)}{U}$	Percentage of text different between version at time $t$ and current version

We evaluate the stability of an article looking at the values of these two functions. If an article is stable it means that  $Et$ , from a certain point of time  $t$ , should decrease or be almost a constant that means that the number of editing is stable or decreasing: the article is not being to be modified. The meaning of  $Txt(t)$  is an estimation of how different was the version at time  $t$  compared to the current version. When  $t$  is close to the current time point,  $Txt$  goes to 0, and it is obviously 0 when  $t$  is the current time.

An article is stable if  $Txt$ , from a certain point of time  $t$  not very close to the current time is almost a constant value. This means that the text is almost the same in that period of time. As mentioned above, an article can be stable because it is rarely edited, but this may mean it is not taken in consideration rather than it is complete. To avoid this, the degree of activity of the article and its text

quality are used as a logic condition for stability: only active and articles with good text can be considered stable.

## 6 Evaluation

We developed a working prototype in C able to calculate our trust factors. A diagram of the prototype is depicted in figure 3. The system, using the *factors updater* module, is continuously fed by the Wikipedia DB and it stores the results in the factor DB. The Wikipedia database is completely available for download. When we want to estimate the trustworthiness of an article, the Data Retrieval module query the Wikipedia DB (it could retrieve information directly from the web site as well), and it collects the needed data: article page, talk page, modification list, user’s list, article category and old versions. Then, the *factors calculator* module calculates each of the trust factors, merging them into the defined macro-areas. Using the values contained in the *Factors DB* about pages of the same or comparable category, it computes a ranking of the page for each macro-area. Finally, the *trust evaluator module* is in charge for estimating a numeric trust value and a natural language explanation of the value. The output is achieved by merging the partial trust value of each macro-area using constraints taken from the *Logic Conditions* module. This contains logic conditions that control the meaning of each trust factor in relationship to the others:

- IF leadership is high AND dictatorship is high THEN warning
- IF length is high AND importance is low THEN warning
- IF stability is high AND (length is short OR edit is low OR importance is low) THEN warning

By looking at the page rank in each macro-area and considering the warnings coming from the logic condition module, explanations like the following can be provided:

*“The article has a strong editing leadership. The very high standard deviation of the edits suggests that it could be an article written mainly by few people. The quality of editing is good but its length is the highest in its category and the topic has average importance. The number of discussions is below average.”*

We present now three different experiments we conducted. Two experiments were performed using our trust factors identified using domain-specific expertise.

The third and last experiment was preformed with a radical different approach. We performed a cluster analysis to isolate featured and standard articles. The experiment was performed on the same set of data used for the second article. The radical difference of the approaches is that in the first two experiments we exploit explicit rules and factors deducted from expertise, while the last approach is obviously implicit. The comparison of the results will show the added value, if any, of using domain-specific expertise in the Wikipedia context.

In all our three experiment, in order to test our predictions we should know if the quality of an article is

actually good. Wikipedia gives its best articles some awards that guarantee that these articles represent the highest standard of the encyclopaedia. There are two levels of awards. The first is the *featured article* status, which means that it has been identified as one of the best articles produced by the Wikipedia community, particularly well written and complete. Only 0.1% of the articles are featured articles. The second level is the *good article status*: articles contain excellent content but are unlikely in their current state to become featured; they may be too short, or about too an extensive or specific topic, or on a topic about which not much is known. We focused on featured articles: they should represent the trustworthiest ones, and the evaluation phase will succeed if our trust computation indicates these articles among the most trustworthy.

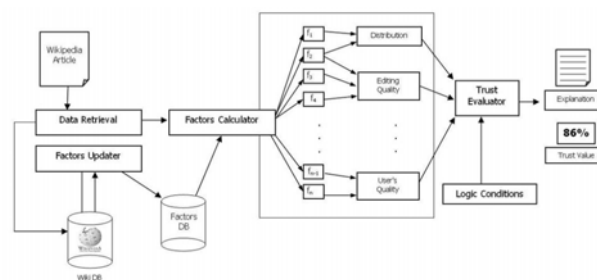


Figure 2 Trust Calculator for Wikipedia

### 6.1 Study-Case 1: A category of articles

We consider a subset of Wikipedia pages, the articles that describe geographical countries. We analyzed 250 countries. We decided to use these articles because they are among the more visited and edited pages; their topic is multidisciplinary, inter-cultural, they interest almost the whole community of wikipedians and they tend to have a standard that lets us meaningfully compare them to each other. For each page, we calculated a trust value in [0.1], where 1 is defined to be the most trustworthiness. The experiment was done on the 30<sup>th</sup> January 2006. On this date, there were 8 featured articles: *Australia, Belgium, Cambodia, Bhutan, Hong Kong, India, Nepal* and *South Africa*.

Table 5. User distribution ranking.

R	Article	T. V.	R	Article	T. V.
1	Portugal	1	9	Pakistan	0.921
2	Cuba	0.994	10	Trinidad and T.	0.914
3	Australia	0.974			
4	Cambodia	0.971			
5	India	0.961	16	Hong Kong	0.907
6	Canada	0.960	36	Bhutan	0.776
7	Belarus	0.953	49	Nepal	0.682
8	Belgium	0.924	55	S. Africa	0.633

Table 6. Dictatorship effect ranking.

R	Article	T. V.	R	Article	T. V.
1	Portugal	1	9	Venezuela	0.455
2	Cuba	0.787	10	Trinidad and T.	0.433
3	India	0.676			

4	Chile	0.606			
5	Australia	0.576	20	Hong Kong	0.339
6	Belarus	0.526	29	Bhutan	0.285
7	Cambodia	0.506	47	Nepal	0.24
8	Canada	0.480	70	S. Africa	0.199

Table 5 and 6 estimate *users distribution* and the *dictatorships effect*. The article *Portugal* seems to have a high possibility of suffering from the “dictatorship effect”, while the same trust value decreases rapidly for the other articles. Our hypothesis is proven by reading a discussion on the *Portugal Talk page* on Wikipedia, where users complained about an author that did 35% of the edits, writing that “*Wikipedia is not a personal web page*”.

The *quality of users* macro-area seems not to be important in the trust computation. Among the featured articles, only *Nepal* (7), *Australia* (10) and *Bhutan* (41) seem to have a good rank. Since the quality of the users writing the article should be a strong factor for its trustworthiness, the evaluation phase suggests that our formulas extracted from the application model failed and we need to go deeper in the analysis.

Table 7. Quality of Editing ranking.

R	Article	T. V.	R	Article	T. V.
1	Australia	1	9	U.K.	0.902
2	U.S.A.	0.987	10	Israel	0.891
3	Portugal	0.97			
4	S. Africa	0.968	34	India	0.842
5	Germany	0.967	37	Nepal	0.751
6	Singapore	0.951	47	Bhutan	0.714
7	Turkey	0.936	52	Cambodia	0.685
8	Belgium	0.922			

Table 7 shows the *quality of editing*. This factor is very effective: featured articles are among the more referenced, they have the right length (which is about 5000-6000 words), balanced sections and images. It is interesting that many articles, good in the others factors, cannot survive the quality of editing analysis.

We observed that *Portugal* is the longest article, with double the text of *France* and 30% more than *United Stated*. The only comparable one is *Cuba*. If we look at the *dictatorship effect* (table 6) we can think that this is the result of lack of control on single users’ edits.

Table 8. Article’s stability ranking.

R	Article	T. V.	R	Article	T. V.
1	Belgium	1	9	Paraguay	0.877
2	Saudi Arabia	0.99	10	Austria	0.86
3	Liberia	0.954			
4	Fiji	0.95	23	Cambodia	0.782
5	Honk Kong	0.914	73	Bhutan	0.488
6	Australia	0.904	97	S. Africa	0.357
7	China	0.895	106	India	0.281
8	Madagascar	0.886	146	Nepal	0.104

Table 8 shows the *stability* ranking. The two most stable articles are *Belgium* and *Saudi Arabia*. Featured articles like *S. Africa*, *India* and *Nepal* show a bad degree

of stability. Regarding *Article’s activity ranking*, as expected the most important and influent countries appear at the top of the table. This factor should be considered as a condition to test stability: stable articles with less than 0.5 degree of activity are considered not edited rather than stable; the instability of an article is more dangerous if it has a high degree of activity and controllability.

Table 9. Overall Ranking.

R	Article	T. V.	R	Article	T. V.
1	Australia	1	9	Portugal	0.87
2	Belgium	0.93	10	U.S.A.	0.86
3	Singapore	0.91			
4	China	0.90			
5	Germany	0.89	16	S. Africa	0.84
6	H. Kong	0.89	20	Cambodia	0.83
7	India	0.88	37	Bhutan	0.76
8	Japan	0.88	52	Nepal	0.71

Table 10. Warning among Table 12 articles

R	Article	Warning reason
1	Portugal	Dictatorship effect. Article too long
2	India	Instability
3	Nepal	Instability

Joining all the previous factors, we can estimate our trust value. *Australia* is the more trustworthy article, 4 out of 8 featured article are in the top 10 position, 6 out of 8 with a trust value higher than 83%. *Nepal*, the worst among them, scored 71.3%. *Nepal* was a featured article in a previous version that was almost 20% different from the current one, situation that is underlined by the warning on stability. *Belgium* had a warning on its activity rate but, due to the quality of the editing and its higher stability, the warning could be interpreted as an evidence that the article has reached a reasonably complete and satisfying state. Regarding the non-featured articles in the top-ten list, *U.S.A.* and *Japan* have the *good article* status, while *Singapore* is a *former featured article*.

## 6.2 Study-Case 2: 8 000 articles

The experiment was conducted on the 17th of March 2006 on 7 718 Wikipedia articles.

These articles include all 846 featured articles plus the most visited pages with at least 25 edits. These articles represent 65% of the editing activity of Wikipedia and the vast majority of its access, making it a significant set. The results are summarized in figure 3. The graph represents the distribution of the articles on the base of their trust values. We have isolated the featured articles (grey line) from standard articles (black line): if our calculation is valid, featured articles should show higher trust values than standard articles. Results obtained are positive and encouraging: the graph clearly shows the difference between standard articles distribution, mainly around a trust value of 45-50%, and featured articles distribution, around 75%.

Among the featured articles, 77.8% are distributed in the region with trust values > 70%, meaning that they are all considered good articles, while only 13% of standard articles are considered good. Furthermore, 42.3% of standard articles are distributed in the region with trust values < 50%, where there are no featured articles, demonstrating the selection operated by the computation. Only 23 standard articles are in the region >85%, where there are 93 featured ones. The experiment, covering articles from different categories, was conducted on an absolute scale, and it shows a minimal imprecision if compared with a previous experiment conducted on a set of 200 articles taken all from the category “nations” [22], where we could rely on relative comparisons of similar articles. This shows that the method has a promising general validity.

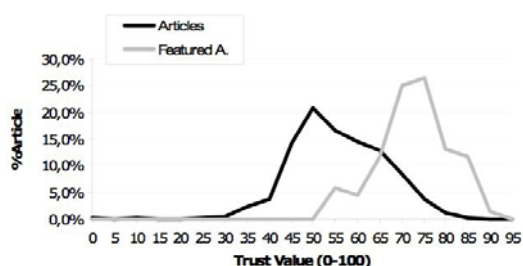


Figure 3 Expertise-based computation

Table 11: expertise-based computation. TV = trust value; SA = standard articles, FA = featured articles

Correlation	18.8 %		
	% of FA	% of SA	GAP
Bad: TV < 50	0	42.3%	42.3%
Average: 50 < TV < 70	22.2 %	54.7 %	32.5%
Good: TV > 70	77.8 %	13 %	64.8 %
Very Good: TV > 85	13.2 %	23 articles	13.2 %

### 6.3 Study-Case 3: Cluster Analysis

In this experiment we performed a pre-processed cluster analysis over Wikipedia articles after identifying a subset of principal articles characteristics. The scope of this experiment is to verify the value added by the expertise by comparing the results obtained in the two cases.

In previous experiments we exploited some aggregated and non-intuitive trust factors, justified and derived by expertise in areas relevant to Wikipedia. We defined some formulae that in general were not intuitive but achieved relying on domain specific expertise.

In this experiment we perform a cluster analysis to automatically divide featured and standard articles based on common features among articles. The comparison of the two results will show if the application of expertise has added value to the quantitative value of the predictions or has only a negligible effect.

Cluster analysis is an unsupervised learning technique, but in our experiment before applying data clustering we trained the system in order to identify, among a set of basic article characteristics, the most important one for a classification of articles.

The key difference with previous experiments is that we now need limited expertise, because we have replaced it by training the system (using a subset of articles of known quality) and by relying on the common featured identification of the clustering algorithm.

The hypothesis is that featured articles are recognizable by simple characteristics that do not require the application of complex expertise. Of course, some kind of knowledge is needed in order to identify a set of article components to be used by the training and by the cluster algorithm, but we avoided complex or derived trust evidence. Note that all of the elements used in this experiment have been considered also by the expertise-based computation (in their simple form or as part of a formula); this is perfectly in line with the aim of the experiment, that is to test the added value of those characteristics that are clearly expertise-derived. In other words we test if, by using only loosely expertise-dependent elements, we will still have valid results. We remind that, in any case, implicit approaches cannot give justifications like our expertise-based method does.

We performed the experiment in two phases. We started by selecting 13 base characteristics of a Wikipedia article. Elements cover both text and editing characteristics. We performed a pre-processing of the data in order to increase the validity of the experiment: we discarded trivial and out-of-standard articles, and we normalized some of the characteristics (like article length, number of images...) using the number of ingoing links of an article, representing a good estimation of its importance (note that we used the same mechanism for the expertise-based computation). These are common procedures and not domain-specific expertise.

Article characteristics are listed in table 15. All the 13 characteristics have low correlation to each other. In the first phase we simplified the list of elements by identifying the principal components. We used 30% of the featured and standard articles to train the system and the rest to perform the experimentation. Using this sample of articles we computed how each characteristic is correlated to featured article and standard article status, i.e. how effective it is in separating the two types of articles. The results are listed in the second column of table 12, and 5 principal components were identified.

In the second phase we performed a cluster analysis of the articles using the 5 principal components identified. We used the well-know k-mean clustering algorithm [13] to identify 2 clusters (featured and standard articles).

Table 12: Components of an Article

N.	Components	Importance
1	Average number of edits per author	Low
2	<b>Variance of edits length</b>	<b>High</b>
3	<b>Percentage of Registered Authors</b>	<b>High</b>
4	Percentage of Contributions by Registered Authors	Medium
5	Percentage of Reverted edits	Low
6	Average Length of editing	Medium
7	<b>Variance of sections</b>	<b>High</b>
8	Average length of a section	Medium



9	Number of discussions	Low
10	Number of Images	High
11	Length of the article	High
12	Number of Section	Medium
13	Number of references, external link	Medium

S. Articles	52.30%	37.00%	10.80%
F. Articles	6.10%	45.60%	48.30%

A graphical representation of the results is displayed in Figure 4. The graph represents the distribution of the *featured articles* (grey line) and *standard articles* (black line) according to the normalized difference of the distances between the article and the two centroids (centres of each cluster). A value of 0 on the horizontal axis means that the article has the same distance from the two centroids; a negative value means that the article is closer to centroid 1 – standard article cluster – while a positive value means that the article is closer to centroid 2 – featured article cluster. Articles whose values are less than -1 or greater than 1 are accumulated at the border of the graph.

If we compare these results with the expert-based computation, we see that by using expertise-derived trust evidence the main added value is the reduction of uncertainty: 77.2% of featured articles have a clear high trust value against only 48.3% in the cluster computation. Moreover, in the expertise case not a single featured article was placed in the region with trust value <50%. This means that, if an article is in that region, it is almost certainly a standard one. On the contrary, 6.1% of featured articles in the cluster computation fall into the region of standard articles (cluster 1). The value added is due especially to the analysis of user distribution and stability. Thanks to these aggregated expertise-justified functions, more certain predictions can be done in borderline cases, and it is possible to capture characteristics that an implicit approach fails to identify.

Referring to table 16a, the two clusters have a recognizable separation: 76.4% of standard articles are in one cluster (region of negative values) while 78% of featured articles are in the other cluster. 23.6% of standard articles fall into featured articles cluster, while in the expertise-based computation they were only 13.2%. A portion of 2.5% standard articles is very close to the featured articles centroid, while only 23 standard articles (less than 0.03%) had a trust value >85 in the expert computation. In general, the results still show an interesting validity partially comparable with previous results. The value added by the expertise results more evident if we consider the uncertainty of the predictions. We divided the algorithm results in 3 zones: featured articles (-1,-0.33), standard articles (0.33,1) and an intermediate zone (-0.33,0.33) where a decision cannot be taken. Referring to table 13b, 45.6% of featured articles and 39.6% of standard articles fall into the intermediate region. This means that in almost half of the cases the algorithm predictions are highly uncertain. Only 10.8% of standard articles are in the featured cluster (slightly better than expertise computation), but 6.1% of articles are in the standard articles cluster compared to none in the expertise case.

### 7 Conclusions

In this paper we have proposed a transparent, non-invasive and automatic method to evaluate the trustworthiness of Wikipedia articles. The method was able to estimate the trustworthiness of articles relying only on their present state, a characteristic needed in order to cope with the changing nature of Wikipedia. After having analyzed what brings credibility and expertise in the domains composing Wikipedia, i.e. content quality and collaborative working, we identified a set of new trust sources, trust evidence, to support our trust computation. The experimental evidence that we collected from almost 8 000 pages covering the majority of the encyclopaedia activity leads to promising results. This suggests a role for such a method in the identification of trustworthy material on the Web. The detailed study case, conducted by comparing a set of articles belonging to the category of “national country” shows how the accuracy of the computation can benefit from a deeper analysis of the article content. In our final experiment we compared our results with the results obtained using a pre-processed cluster analysis to isolate featured and standard articles. The comparison has shown the value added by explicitly using domain-specific expertise in a trust computation: a better isolation of articles of great or low quality and the possibility to offer understandable justifications for the outcomes obtained.

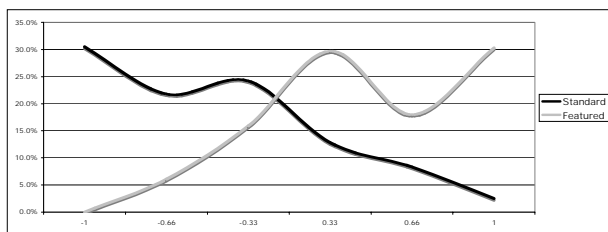


Figure 4 Graphical representation of Cluster Analysis.

Table 13: Cluster Divisions

CASE A			
	Cluster 1	Cluster 2	
S. Articles	76.40%	23.60%	
F. Articles	22%	78%	
CASE B			
	Cluster 1	Intermediate	Cluster 2

### References

- [1] Ciolek, T., Today's WWW, Tomorrow's MMM: The specter of multi-media mediocrity, IEEE COMPUTER, Vol 29(1) pp. 106-108, January 1996
- [2] How much do you trust wikipedia? (March 2006) [http://news.com.com/20091025\\_3-5984535.html](http://news.com.com/20091025_3-5984535.html)
- [3] Gales, J. Encyclopaedias goes head a head, Nature Magazine, issue N. 438, 15, 2005

- [4] Romano, D. M., *The Nature of Trust: Conceptual and Operational Clarification*, Louisiana State University. PhD Thesis, 2003
- [5] Blaze, M., Feigenbaum, J. and Lacy, J., Decentralized Trust Management. Proceedings of IEEE Conference on Security and Privacy, 1996
- [6] Chu, Y., Trust Management for the World Wide Web, Master Thesis, MIT, 1996
- [7] Terzis, S., Wagealla W. *The SECURE Collaboration Model*, SECURE Deliverables 2.1, 2.2, 2.3, 2004
- [8] Marsh, S. Formalizing Trust as a Computational Concept. PhD thesis, University of Stirling, Scotland, 1994
- [9] V. Cahill, et al. Using Trust for Secure Collaboration in Uncertain Environments. IEEE Pervasive Computing Magazine, 2003
- [10] Golbeck, J., Hendler, J., Parsia, B., *Trust Networks on the Semantic Web*, University of Maryland, USA, 2002
- [11] [www.foaf-project.com](http://www.foaf-project.com), FOAF project website
- [12] Ziegler, C., Golbeck, J., *Investigating Correlations of Trust and Interest Similarity*, Decision Support Services, to appear.
- [13] MacQueen J., *Methods for classification and Analysis of Multivariate Observations*. Proceedings of 5-th Berkeley Symposium on Mathematical Statistics, Berkeley, University of California Press, USA, 1967
- [14] Bucher, H., Crisis Communication and the Internet: Risk and Trust in a Global Media, First Monday, Volume 7, Number 4 2002)
- [15] Huang, J., Fox, S., *Uncertainty in knowledge provenance*, Proceedings of the first European Semantic Web Symposium, Heraklio, Greece, 2004.
- [16] Burke, R., *Knowledge-based Recommender Systems*. Encyclopaedia of Library and Information Systems. Vol. 69, Supplement 32, New York 2000
- [17] Fogg, B. J., *How Do Users Evaluate The Credibility of Web Sites?* Proceedings of the conference on Designing for user experiences, ACM Press, USA, 2003
- [18] Alexander, J., Tate, M., *Web Wisdom: How to Evaluate and Create Information Quality on the Web*, Lawrence Erlbaum Associates Inc, New Jersey, USA, 1995
- [19] Cassel. R., *Selection Criteria for Internet Resources*, College and Research Library News, N. 56, pagg. 92-93, 1995
- [20] Stanford Web Credibility Guidelines web site, <http://credibility.stanford.edu/guidelines>
- [21] Roberts, T., *Online Collaborative Learning, Theory and Practice*, Idea Group Pub, USA, 2004
- [22] <http://download.wikimedia.org>, download site of the Wikipedia project

# Mobile Ticket Control System with RFID Cards for Administering Annual Secret Elections of University Committees

Hans Weghorn

BA-University of Cooperative Education

Rotebühlplatz 41, 70178 Stuttgart, Germany

E-mail: weghorn@ba-stuttgart.de, <http://hansweghorn.org>

Hans Peter Großmann, Dieter Hellwig, Cahya Kusuma Ratih, Andreas Schmeiser and Heiko Hutschenreiter

University of Ulm, Albert-Einstein-Allee 43, 89081 Ulm, Germany

E-mail: hans-peter.grossmann@uni-ulm.de, <http://omi.e-technik.uni-ulm.de>

**Keywords:** ubiquitous computing, RFID, NFC, mobile applications, wireless JAVA

**Received:** April 23, 2007

*RFID technology often is suspected to provide means of undesired surveillance and observation of people in their roles, e.g., as private persons, customers of any business or employees. Here, an opposite scope of application sample shows how such technology based on intelligent smart cards can help improving secrecy in a sensitive environment. The annual election runs for the representatives of the studentship of Ulm University have to be operated in a way that it cannot be tracked, who is electing how often. On the other hand, it has to be ensured that voting is not abused, hence some entrance check is obligatory. The technical solution presented here is using a prototype mobile phone, which is equipped with a communication module for contact-less information exchange with the student ID. In addition to ensuring confidentiality, with this mobile vote administration system the studentship of Ulm University has the opportunity of enhancing the operation of the election, because the election office is also mobile now and not fixed to any location in the University buildings. This enables that the election assistants go out for the students instead of waiting for them to come, and by that, the poll is likely to be increased.*

*Povzetek: Razvita je metoda za preverjanje pri volitvah, temelječa na RFID.*

## 1 Introduction

Today, RFID tags are applied in many fields of daily life. For instance, clothes in department stores may be marked with such electronic tags intending different aims: On the one hand side, these tools may simplify purchase and payment, on the other hand stealing of goods may be prevented by electronic supervision of exit doors. The latter aspect leads to the part of RFID technology, which is in general feasible for tracing and observing people. And of course, there exist well-funded concerns of people against such scenarios which remind us of George Orwell's famous novel "1984".

Especially considering the technical weaknesses of radio monitoring systems [1], these concerns appear quite reasonable. Overall, the question arises whether everything which is technically possible should be implemented, and what kind of impacts radio monitoring constructions will have on societies [2].

Despite this all, it also has to be respected that in many environments access control is required and unavoidable. For instance, protecting buildings, offices, or private property against unauthorized access can be performed with electronic systems in a quite convenient and efficient manner. Such a sample system is an electronic ID card of our University students: In Ulm

University environment, a smart card is used as student ID. It is a MIFARE® classic card that is distributed by the company Philips [3]; its capabilities are an extension of the RF communication standard ISO 14443A. This communication standard often also is called Near Field Communication (NFC), because the RFID reader and the card have to be in maximum distance of 10 cm. Our specific Mifare® card contains 1 kbyte of E<sup>2</sup>PROM memory that is sectorized, and for each data sector there can be different access rights and keys defined, which are independent of the other data sectors.

In its storage system, the card holds the matriculation number of the student, the access number for library services, and other information in terms of read-write byte arrays. Another possibility is using a data block on the card as counter value. The counter can be configured for an access with two distinct secret keys: One key enables setting the counter value, while the other access key is good only for reading and decrementing the counter. With this, electronic payment functions [4] will be realized in near future: The loading of the counter – i.e. booking money onto the card – is done with the "better" key, and this is used only in a specially protected environment. Discharging the account is not as sensitive,

and there is no risk of imposture by leaving the other – weaker – key in electronic cash boxes, or in public terminals for inspecting the account balance and last transactions in this electronic wallet.



Figure 1: Student ID card of Ulm University, and prototype phone equipped with RFID communication hardware.

This specially shielded counter system on the student ID cards was also already used since several years as entry ticket for the elections of the studentship's representatives, because this application is equivalent to money payment scenarios: With each re-matriculation (Fig. 2), the students earn the right to vote for their representatives during the new semester. Since there may be several elections in the same semester, an initial high counter value is decremented with each vote. The start counter value is increased with each studying year by a big stepping count, so there is a clear distinction between the different years (there have to be never more than this count of elections in the same year). The counter is loaded on the student ID in an electronic administration terminal accordingly. Up to now, the counter was read and decremented in the election office with a desktop computer with special RF communication equipment and software.

Overall, this construction represents a closed system, and the students can rely on the fact that it is not traced in any database, how often they vote, or whether they vote at all. Ensuring confidentiality is a very important issue for making the students to contribute their vote. The stationary computers in the election office may feed the suspect that the voting could be traceable, because these are easily connected to the Intranet and by that to the administration database of the University. Therefore, a mobile control system should help diminishing this concern. For this, a prototype cellular phone was supplied in the frame of an industrial cooperation (Fig. 1) [5], which is equipped with an additional RF hardware for communicating with the student ID. The following sections discuss the implementation and operational aspects, which are required for constructing mobile software for the election office.

Another important positive effect of using a mobile system for the vote control is that the students do not have to come any more actively to an office, which is located in one single building of a widely distributed campus. The election office is made movable itself, and

it can be placed to efficiently meet many students, e.g., after courses in big lecturing halls. Due to this improvement, the turnout at the election, which is traditionally low for different reasons, is presumed to increase in future.

## 2 Data representation and data handling on the student ID card

The student ID card contains one kilobyte of EEPROM memory, which is split into 64 blocks [3]. Four blocks are always clustered in one sector, while the last block in each cluster defines the access rights and cryptographic access keys for the entire sector. The data blocks can be used either as flat memory of 16 bytes size, or as 32 bit counter value [6]. In the latter case, the counter value is replicated within its block two times for security reasons and for enabling consistency checking. For controlling access to the elections of the studentship's representatives, one dedicate block on our student ID card was defined as containing a counter value. On base of this data entry, it can be decided whether a student is allowed to apply a vote or not. For better understanding of the functionality and its requirements, it has to be noted that the number of election runs are varying each year, and in general, there will be several elections. This control system was introduced in year  $y_b$ , and the counter is reloaded each year with the first re-matriculation of the student, which is obligatory in each semester, with a new value  $c_r$ :

$$c_r = (y - y_b + 1) \cdot s$$

$y$  denotes the actual year

$y_b$  denotes the year of the introduction of this system

$s$  denotes a stepping factor that is larger than the maximum number of elections in each year

When the student applies a vote during the election this counter value on his/her ID card is in first approximation decremented by one. After the  $n$ -th election the counter will carry the following target value  $t$ :

$$t = (y - y_b + 1) \cdot s - n$$

An entry control system for the votes now has to compare the target value  $t$  calculated for the current run with the counter value  $c$  on the ID card. From this, a decision can be applied:

$$c > t \Rightarrow \text{Vote is allowed}$$

$$c = t \Rightarrow \text{Already voted, i.e. vote is not allowed}$$

$$c < t \Rightarrow \text{Re-matriculation not yet performed – has to be completed first}$$

During each election, the same person may only apply one single vote, even if there was no participation in elections that took place earlier in the same year. Therefore, in the true control system the counter value is not simply decremented during each vote, but it is always set to the target value as defined above (performed by iterative decrementing). This ensures that only one single

vote can be applied in each run. The reloading of the counter is tracked in the administration system, so that a second re-matriculation will not reset the counter. On the other hand, the counter on the card itself is not tracked and also not stored in any central database or administration system of the University. Hence, this construction represents a closed system, and it cannot be observed from outside, how often a student attends the elections, or whether the student votes at all.



Figure 2: One of the public terminals at Ulm University, where students can re-matriculate with their ID and perform other administration tasks, e.g., updating their home address or printing out certificates of exams.

For security reasons, precise values of the above-defined parameters are not documented here. The ID card also contains couple of additional information, like registration number for the studying course and lending code for the University library. These additional data contents are not relevant here, and the different access keys for the different data sectors on the ID card prevents interference between data handling of the different instances of administration that are indeed required for operating a University system. The different access keys shield the different application against each other, and make these all perpendicular.

Up to now, the election entry control was performed with stationary desktop computers that are equipped with NFC readers / writers. This control system shall be miniaturized and made moveable by using mobile terminals instead of desktop computers. For NFC experiments the company Siemens / BenQ Mobile supplied few prototype mobile phones, which are equipped with NFC communication hardware in addition to their normal functionality (Fig. 1). These devices were used for the implementation of the application described in the following sections.

### 3 System construction and its operational behaviour

#### 3.1 Start up of election control tool

For simplicity and convenience of handling, the election tool was automated as much as possible. After displaying a welcome screen, which disappears automatically after a short delay, the user has to enter a PIN code (Fig. 3).

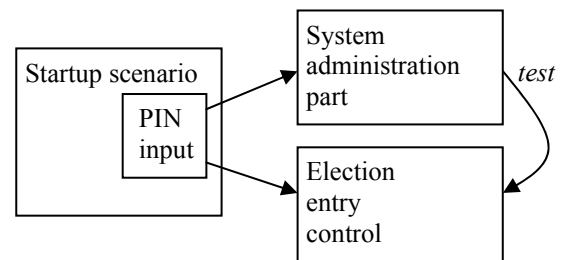


Figure 3: The election tool is structured into three parts, each of which is realized as finite state machine (FSM). Depending on the entered PIN code, the SW branches from the start-up scenario either to the system administrator part, or to the regular operational mode, which is using the system as entry check for the elections.

The system contains two valid PIN codes, one for the system administrator, and another one for the vote assistant. Hence, after entering the PIN code, the application automatically can branch to the required operational scenario and by that, one selection menu is saved. Of course, all error cases – at this stage entering an invalid or no code – are treated also separately. After installation of the vote software on the device, a default PIN for the system administrator is generated internally. The election control, which is considered being the regular use case, cannot be launched until all required system parameters are entered by the system administrator.

#### 3.2 System administration

In the system administration scenario, the administrator of the tool has to enter the following information into the system:

- System administrator PIN
- Vote administrator PIN
- Date period for election run
- Running number for the election in the actual year
- Secret access key for the ID card
- Timeout values for card scans and sensitive menus

For testing purposes, the system administrator can read the election counter on the card. With this, it can be verified that the configured secret key is valid. Also for system testing, the vote assistant software can be launched directly from the system administration menu.



### 3.3 Vote control

After displaying a timed welcome screen for the vote administrator, the system switches to an input screen, from which a card scan can be launched by one single key press (Fig. 4). After detecting a card, and reading the counter value (and if appropriate decrementing it) one of the possible results is displayed. The student, who wishes to vote, may be granted or rejected, or this student may

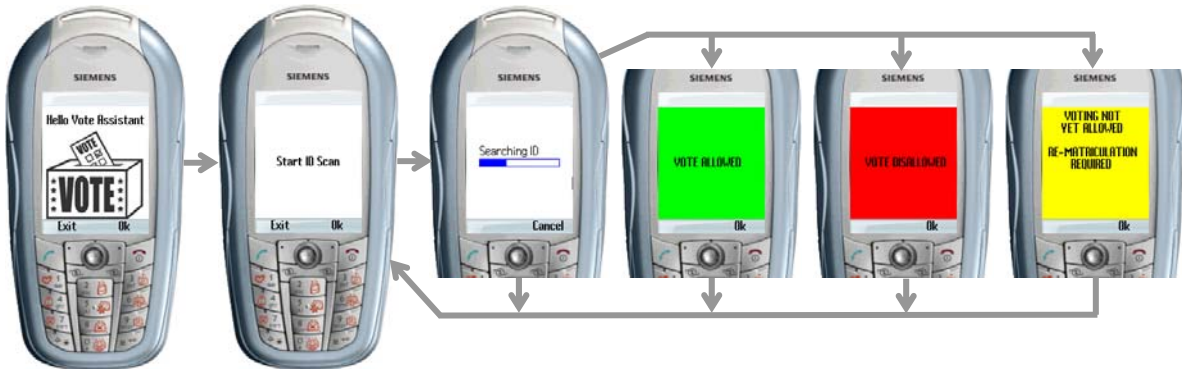


Figure 4: UI chart for the vote control system without the error screens. After scanning the card, the result is displayed with indicative coloured screens.

be required to re-matriculate first. A card scanning problem or time-outs represent the possible irregular cases. The vote administrator has to acknowledge the result screen, and then the system returns to the screen, from where a card scan can be initiated again. In the regular case, the vote administrator has to apply for each ID control two manual button presses – one for starting the scan, and a second one for acknowledging the result. The scanning of the ID card could theoretically be started automatically, but this is not recommendable because the card search consumes high power. Hence, the manual start of the card scanning will help to extend the operational stand-by time of the device.

### 3.4 Aspects of technical implementation

The software was realized in wireless JAVA [7], which offers convenient and adequate programming functionality for SW that shall be executed on mobile phones. In wireless JAVA, the standard UI provides the possibility to constrain input in text fields with certain properties. In particular, the input field for passwords can be masked automatically, and the char set for input can be limited, e.g., to decimal digits only. With this kind of control, the extend of error checking code for wrong input is reduced considerably. Another special standard UI element is the class DateField, which allows the user to comfortably enter the required information (Fig. 5). Overall, wireless JAVA offers a UI concept that is fully appropriate for small devices [8], which are constraint in terms of screen size and keyboard char set.

On the prototype phone, the internal communication with the NFC hardware is performed through a serial connection, which also can be handled by standard library functions of wireless JAVA. The election tool is structured in two parts: The application layer, which takes care of user inputs and treatment of execution parameters, and a connection library that handles the

communication with the ID card. Both layers are active in general simultaneously, and hence multi-threading was used to serve this requirement of parallel processing.

Multi-threading [9] represents one important standard language feature of Java that provides advantage over alternative programming languages like C or C++. For a communication with smart ID cards there exist already a library specification for wireless Java [10], which has the

identification JSR-257, and a preliminary library implementation was available for the prototype phone. Unfortunately, this library specification completely disregards the most advantageous features like using cryptographic keys for data exchange or like accessing the card's built-in counter functionality. Instead, JSR-257 treats the ID card as streaming medium, and since this is inappropriate for the given data structures on our student ID card, an own card access library had to be developed.



Figure 5: Wireless Java provides standard UI elements for simplified application programming. The samples here show how password input can be masked automatically, and how comfortable the input of date information can be.

The software metrics of the developed system shall here be described in terms of number of code lines, UI screens, and total count of states of the FSMs. Approximately 20 different UI screens are contained in the complete application, while 20% of them represent error messages and confirmation screens. The display language can be switched by an internal constant between English and German, and hence the string resources had to be implemented for all the screens and messages for both cases. The application layer consists of 1200 lines of code (without comments), and handles 30

different system states. The number of states is higher than the number of UI screens, because the same UI screens can be reused in different states. Two threads are required in the application layer – one for handling UI events, and the second one for cycling through the FSMs –, while the connection library, which consists of 600 lines of code, requires only one background thread that handles the communication via serial interface with the NFC hardware module. Since the latter communication is

For didactic reasons, the above sections show only the relevant core part of the ticket control application, which consists of a few UI screens only. For a complete fault-tolerant tool, which is robust against wrong handling – which is usually unintended – and error cases like mal-functional cards, an extended software construction is required. This can be seen from the numbers of the before discussed software metrics in contradiction to the limited functionality of the inner

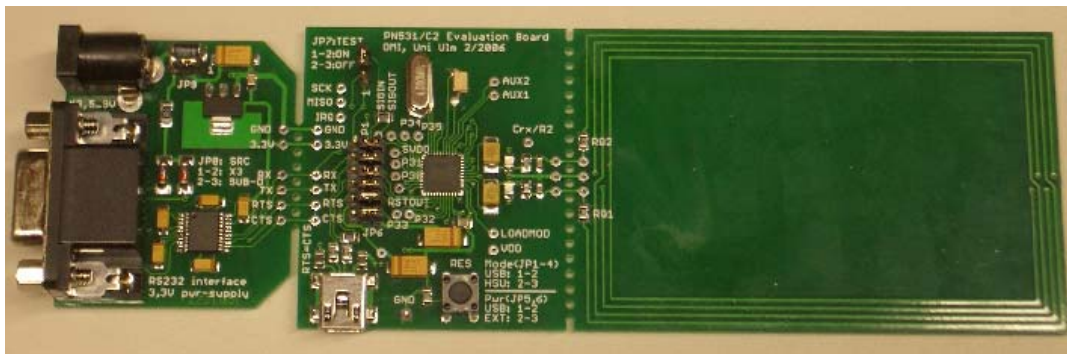


Figure 6: PCB with electronic RF circuit and on-board antenna that was developed at Ulm University. This first prototype design is equipped with USB and RS232 communication interface, so that it can be connected to different host systems (e.g., recent PDAs have today micro USB interfaces). Further designs of this will have to be adapted to dedicate handheld devices in use, and it will also be considerably more miniaturized for this purpose.

time-critical, this connection library thread runs with increased priority

“good” branch (Fig. 4).

#### 4 Data security and protection against abuse of system

For preventing abuse of the system, a variety of security features were built into the software:

❶ After installation, the tool can only be launched with a default PIN, and before the software can be used for scanning an ID card, all parameters defined above have to be entered by the administrator.

❷ One important aspect is that the secret access key for the ID card is crypted in persistent storage area on the device (part of the phone’s FLASH memory). The concept is to use the vote assistant PIN as starting value for a binary noise polynomial [11], which is used for ciphering the secret card key.

❸ The PIN value itself is not preserved on the system; instead a hash value [12] derived from the PIN is saved in the FLASH memory, so that the entry to the voting software can be verified, but without entering the correct PIN it is impossible to derive the value of the secret card access even at infinite efforts of HW and SW debugging.

❹ An implicit protection is already available with the construction of the ID card itself, because the used key is only feasible for reading and decrementing the election counter on the card. Hence, even if the secret access key could be extracted, the counter value could not be modified in a way that the ID owner is allowed to vote more than once in each election run.

❺ Another approach of protection is that sensitive menus of the application are equipped with timer

functionality: If the administration software or the vote assistant software is started successfully, but it is not used for certain defined time (this duration can be configured by the administrator), the application exits automatically. This ensures that even if the system is stolen, when an authorized person has logged on already, the tool will terminate after a while.

❻ Furthermore, for preventing any abuse, the vote assistant software works only during the specified date frame of the election run. Before and after these dates, the software cannot be launched with the vote assistant’s PIN code at all.

❼ To preserve secrecy, the actual counter value on the ID cannot be displayed in the voting control scenario. Only the system administrator can inspect from his menu the vote counter, but this possibility is required for validation of the cryptographic key and error detection in the voting system itself. During the election runs, the system administrator is not acting as vote administrator, but s/he is only in charge of technical support during this time.

#### 5 Actual and future developments

In parallel to the plain software work with the RFID phone, we have also developed an own communication hardware, which is software-compatible to the NFC interface used in the prototype phone. This hardware (Fig. 6) is more versatile, because it can be connected via either USB or RS232 interface. Hence, this communication device can be attached electrically to standard desktop computers as well as to handheld computers, or smart phones. Several samples of this own



hardware have been manufactured manually, and the design could successfully be verified already.

Comparative software investigations with this new PCB already unveiled a severe firmware bug in the prototype phones. Unfortunately, exactly this firmware bug in the RFID communication module prevents that at the moment, that the outlined security concepts can be used to their full extend.

Future use of this PCB design will be in connection to very simple handheld computers, which do not have any else communication interface like wireless phone network access or WLAN. This shall yield higher confidence in the secrecy of the election entry control. This hardware design will add benefits to the project in different ways: At first, it will allow us becoming manufacturer-independent, and secondly it will enable using cheaper consumer devices for this application. Furthermore, at the moment this RFID circuit is adapted to replace the faulty RFID communication module in the prototype phones. The software written so far is adapted specifically to the properties of the used prototype phone. For using in future alternative devices, the software will have to be extended in terms of dynamically handling screen sizes, and possibilities for persistent data storage.

## 6 Conclusion

A concept for an appropriate ticket system for the elections of representatives of the studentship in Ulm University was developed. The system provides secrecy for the voting action, and the concept prevents an abuse of voting. With the development and implementation of software for a prototype mobile phone that is equipped with an additional NFC communication module besides its standard hardware, the vote ticket control itself is made mobile. The advantages of this approach are obvious: By enabling easily movable control tools, the election can be operated in decentralized manner, and by that the poll is presumed to be increased.

As outlined, the vote control represents an embedded closed system. Due to RFID communication with the intelligent smart card, the application can run entirely in local mode. It is not required for the application to interact with University's administration system or database, and hence this application shows how RFID technology can help improving privacy and secrecy.

Meanwhile two dozen NFC-enabled phones were provided by our industrial partner for a use in the project. Unfortunately, it turned out that the NFC communication subsystem in these phones contain firmware errors, so these devices cannot be used in the safest mode like described above. Therefore, at the moment we are developing our own communication hardware module for these phones to fix this problem by replacing the faulty hardware. After achieving this, the next election run shall take advantage of this new technology.

Furthermore, with the own hardware the number of devices and the handling in the election scenario can be improved additionally. This NFC technology can be applied for wider purposes during the operation of our studying courses, for instance using the NFC-enabled

devices as presence control in examinations, or entering marks and results in laboratory exercises and oral examinations through the handheld scanning device [5].

## Acknowledgement

We want to thank the companies Siemens AG, Munich, and BenQ Mobile GmbH, Munich, for their kind hardware and software support, and for providing the set of prototype RFID phones.

## References

- [1] Molnar, D., and Wagner, D. (2004) Privacy and security in library RFID: issues, practices, and architectures. *Proceedings of the 11th ACM conference on computer and communications security*. Washington DC, USA, pp. 210--219.
- [2] Hugl, U. (2005) Employment of upcoming technologies and aspects of privacy. *Proceedings of the IADIS conference on e-Society*. Qwara, Malta, pp. 333--339.
- [3] Philips (2006) *MIFARE Classic - contactless Smart Card ICs*, <http://www.semiconductors.philips.com/products/identification/mifare/classic>, last access: March 2007.
- [4] Stoklosa, J. (1998) Cryptography and electronic payment systems, *Informatica. An International Journal of Computing and Informatics*, vol. 22, no. 1, pp. 29--33.
- [5] Weghorn, H., Schmeiser, A., Großmann, H. P., Pirker, M., and Haubold, S. (2005) ISA4G - Integrated Student Access ID Card of Fourth Generation, *Proceedings of the IADIS Conference on WWW/Internet*, Isaias, P., and Nunes, M. B. (Eds.), Lisbon, pp. 333--337
- [6] Philips, 2001, *MIFARE Standard Card IC MF1 IC S50 Functional Specification*, Revision 5.1, [http://www.semiconductors.philips.com/acrobat\\_download/other/identification/m001051.pdf](http://www.semiconductors.philips.com/acrobat_download/other/identification/m001051.pdf), last access: March 2007.
- [7] Knudsen, J., and Li, S. (2005) *Beginning J2ME: From Novice to Professional*. Apress, Berkeley.
- [8] Mahmoud, Q. H. (2002) *Learning Wireless JAVA*. O'Reilly. Sebastopol, 1st edition.
- [9] Bell, D., and Parr, M. (2002) *JAVA for Students*. Prentice Hall. Dorchester, 3rd edition.
- [10] JCP (Java Community Process) (2005) *JSR-00025 Contactless Communication API*, Version 2.6, [http://www.jcp.org/aboutJava/communityprocess/ed\\_r/jsr257](http://www.jcp.org/aboutJava/communityprocess/ed_r/jsr257), last access: March 2006.
- [11] Rorabaugh, C. B. (2004) *Simulating Wireless Communication Systems*. Prentice Hall PTR, Indianapolis.
- [12] Reeds, J. A., and Weinberger, P. J. (1984) File Security and the UNIX Crypt Command. *AT&T Bell Laboratories Technical Journal*, Vol. 63, No. 8, pp 1673-1684.

# Facilitating Shared Knowledge Construction in Collaborative Learning

Stephan Lukosch  
 Fern Universität in Hagen  
 Department of Mathematics and Computer Science  
 E-mail: stephan.lukosch@fernuni-hagen.de,  
<http://kalu.fernuni-hagen.de>

**Keywords:** collaborative learning, shared knowledge construction, game-based learning, web-based learning communities

**Received:** March 16, 2007

*The German distance learning university uses the web-based collaborative learning platform CURE to support different collaborative learning scenarios, e.g. collaborative exercises or virtual seminars. During these scenarios students form learning groups upon teacher's request and collaborate on a common task. But as soon as the given tasks are accomplished, in most times the collaboration stagnates and finally stops. In this paper, we report on extensions to CURE that were designed to foster shared knowledge construction and allow learning in a entertaining way. These extensions were developed in a participatory process with the students. To enable the students to design and develop these extensions, we used patterns for computer-mediated interaction. In our opinion, these entertaining extensions will lead to a higher degree of interaction between the students which leads to a shared knowledge and finally a learning community.*

*Povzetek: Opisan je dodatek pri platformi CURE za sodelovalno učenje.*

## 1 Introduction

The FernUniversität in Hagen is the German distance learning university. Teaching at the FernUniversität includes different forms of learning: courses, seminars, and different forms of practical problem solving in lab courses. Course material and accompanying individual exercises are sent to distributed students via surface mail or the Internet. As the students are distributed all over Germany, it is difficult for them to find appropriate co-learners and to learn together. As a result, students at the FernUniversität primarily learn individually, feel isolated, lack practice of collaboration, and miss the motivation that teamwork and team members may provide.

A survey including students and different faculties showed a major interest in collaborative learning scenarios and the FernUniversität built the collaborative learning platform CURE [9]. Up to now, students mainly use CURE to form learning groups upon teacher's request. In these groups, they discuss course content, solve assignments, or collaboratively write a seminar thesis. This cooperation and discussion works well as long as there is a group task given by the teachers. If the given tasks are accomplished, the collaboration in most cases stagnates or even finally stops.

In our opinion, a learning community could help to increase collaboration among the students. However, communities cannot be designed. Instead, learning communities evolve through the collective building of shared knowledge and the shifting participation of their members [15] and only the software that supports the

community is designed [22]. There are some key factors for a successful online community. Wenger [27] and Haythornthwaite et al. [11] are of the opinion that participation and practice are the key factors for developing a learning community. Palloff and Pratt [20] emphasize that community members must have possibilities for shared knowledge construction.

CURE can serve as a basis for a learning community if students could be motivated to a higher degree of collaborative interaction and shared knowledge construction. Our approach to reach this goal consists of combining entertaining learning, e.g. based on learning games [23], and participatory design. In our opinion, entertaining learning approaches can be used to increase the motivation for more frequent collaborative interaction which may result in the construction of shared knowledge. Due to a higher degree of collaborative interaction, the students will get to know each other much better and develop responsibilities for their peer students.

Therefore, we extended CURE with learning gadgets, i.e. entertaining tools for collaborative learning. To ensure that these learning gadgets will be accepted by the students, we employed a participatory approach to these new forms of educational interaction. In a lab course, we let our students suggest learning gadgets which in their opinion would foster collaborative interaction and support shared knowledge construction.

To enable our students to design such learning gadgets, we equipped them with a pattern language for

computer-mediated interaction [24]. Patterns can serve as a *Lingua Franca* for design [6] that help end-users and developers in communication and as an educational and communicative vehicle [4; 14].

In the following sections, we first describe CURE in more detail before we focus on the participatory process, the patterns, the resulting learning gadgets, and finally give an outlook on future work.

## 2 CURE in a nutshell

CURE [9] is a web-based collaboration space. It was developed to support the initial scenarios of collaborative exercises, seminars, lab courses, and the preparation of theses.

Students can structure their interaction in groups that inhabit virtual rooms. Room metaphors [7; 21] have been widely used to structure collaboration. Figure 1 shows the abstractions that are offered by CURE. Users enter the cooperative learning environment via an entry room that is called Hall. Rooms can contain pages, communication channels, e.g. chat, threaded mail, and users. Users, who are in the same room at the same time, can communicate by means of a synchronous communication channel, i.e. by using the chat that is automatically established between all users in the room. They can also access all pages that are contained in the room. Changes of these pages are visible to all members in the room. The concept of a virtual key [10] is used to express access permissions of the key holder on rooms. The access permissions distinguish rights to enter a room, create sub rooms, edit pages, or to communicate within the room. Rooms with public keys are accessible by all registered users of the system.

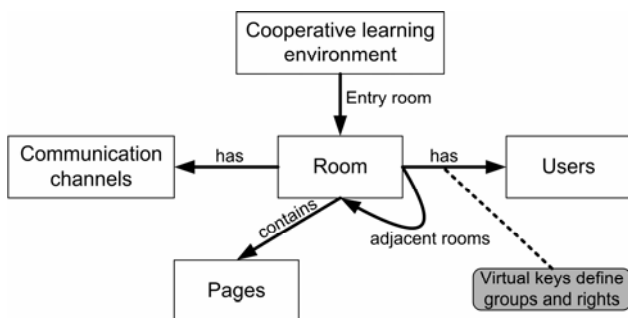


Figure 1: CURE abstractions.

Users can enter a room to access the room’s communication channels and participate in collaborative activities. Users can also create and edit pages in the room. Pages may either be directly edited using a simple Wiki-like syntax [16], or they may contain binary documents or artefacts. In particular, the syntax supports links to other pages, other rooms, external URLs or mail addresses. The server stores all artefacts to support collaborative access. When users leave the room, the content stays there to allow users to come back later and continue their work on the room’s pages.

Figure 2 shows a typical room in CURE. The numbers in the figure refer to details explained in the following paragraphs. A room contains documents (1)

that can be edited by those users, who have sufficient edit rights (2). CURE stores all versions of a page. Users can browse different versions (3) to understand their colleagues' changes. Communication is supported by two room-based communication channels, i.e. a mail box (4) and a persistent chat (5). Users can use the room-based e-mail to send a mail to the room. Users of the room that have sufficient communication rights will receive this message.

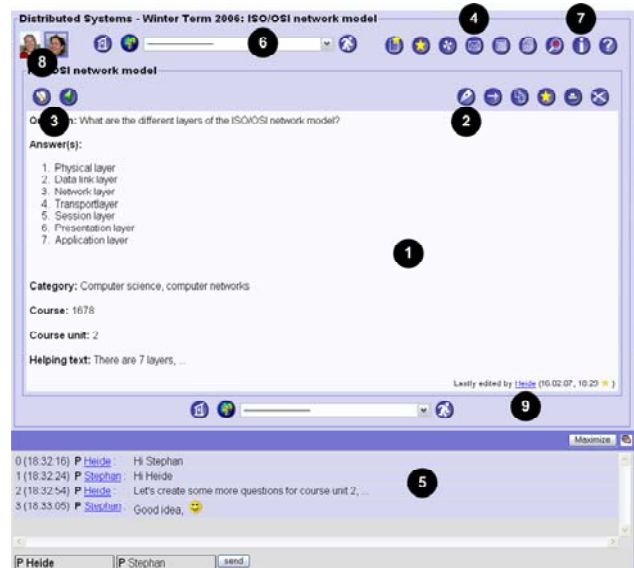


Figure 2: A room in CURE

By providing a plenary room, sharing and communication in a whole class or organization can be supported. By creating new rooms for sub-groups and connecting those to the classes' or organization's room, work and collaboration can be flexibly structured. Starting from the plenary room users can navigate to the connected sub-rooms (6).

For user coordination, CURE supports various types of awareness information:

- Users can see in the room's properties who else has access to this room (7).
- Users can see which users are currently in this room (8).
- If the chat is enabled in the room, users can directly start chatting to each other (5).
- Users can see who has lastly edited the current page (9).
- Daily reports automatically posted to all users of a room include all changes made since the last report was sent.

## 3 Pattern-based participatory design of the learning gadgets

Each year our department conducts a practical lab course in which groups of up to 8 students collaboratively develop a collaborative application. At the beginning of one of these lab courses, we asked the students to suggest learning gadgets for the CURE environment that

- assist them in collaborative learning,

- help to build a shared knowledge,
- motivate them to become an active member of a learning community, and
- foster learning in an entertaining way.

As starting point for further investigations or thoughts, we suggested game-based learning approaches [23] to the students, as these are well-known to increase the motivation for more frequent collaborative interaction. As result of our basic requirements, the students suggested 23 learning gadgets that in their opinion fulfil our requirements. Most of the suggestions focused on the game-based learning approach. After a discussion among teachers and students, some of the proposed learning gadgets were selected and developed in the lab course.

To enable our students to design and develop such learning gadgets, we equipped them with a pattern language for computer-mediated interaction [24] that evolved in our group over the last years. The idea of patterns originates from Christopher Alexander's work [1; 2] in urban architecture. According to Alexander, 'patterns describe a problem which occurs over and over again and the core of the solution to that problem'. Each pattern includes a problem description, which highlights a set of conflicting forces and a proven solution, which helps to resolve the forces. An interconnected set of patterns is called a pattern language. Patterns of a pattern language are intended to be used together in a specific problem domain for which the pattern language guides the design decisions in the specific problem domain.

Especially, when developing applications that support collaborative interaction, end-user involvement is a crucial issue [17]. To foster communication between developers and end-users, they need a common language and understanding of the problem space to determine the requirements. Pattern languages are an educational and communicative vehicle for reaching this goal. Once the requirements are identified, patterns support developers in implementing the collaborative application by teaching them on how to design and develop groupware applications and reuse proven solutions.

Applications that support collaborative interaction are often considered as *socio-technical systems*, as the technical system has to support the social process of collaboration. Thus, our patterns have to describe the *technology* that supports the *group process* and therefore include a technical and a social aspect.

Compared to software patterns, our patterns need a special form that can be understood by end-users as well as software developers. Our patterns follow the pattern structure outlined in the *Oregon Software Development Process* [25]. A pattern starts with the *pattern name* followed by other possible names for the pattern (*AKA*), the *intent*, and the *context* of the pattern. These first descriptions help readers to decide whether or not the pattern may fit in their current situation.

The problem section of a pattern contains the *problem* statement in bold font, followed by a *scenario* and typical *symptoms*. The *scenario* is a concrete description of a situation where the pattern could be used, which makes the tension of the *problem* statement

(the conflicting forces) tangible. The *symptoms* describe typical observations that indicate the problem exists.

The solution section of the pattern explains the actual *solution* to the problem. The *dynamics* section states the main components or actors that interact in the pattern and explains how they relate to each other. The *rationale* section explains the impact of the solution on the various (conflicting) forces involved. Unfortunately, applying a pattern may introduce new unbalanced forces. These counter forces are described in the section labelled *danger spots*.

The solution presented in a pattern represents a proven solution to a recurring problem, so the *known uses* section provides well-known examples where this pattern is applied. Finally, the *related patterns* section states what patterns are closely related to this one, and with which other patterns this one should be used.

In the following sections, we will describe two of the selected learning gadgets. All learning gadgets can be added to the interaction possibilities of a room in CURE by simply creating a new page. Thus, each user can initiate collaboration. The learning gadget *Fountain of Wisdom* focuses on learning based on a question and answer paradigm. The questions and answers can be either provided by the teacher or they can be defined by the students themselves. In either case, the question and answers are collected in a common question repository which serves as shared learning knowledge. The learning gadget *One for all and all for one* focuses on collaborative exam preparation by giving distributed slide presentations [18]. The slide presentations can be shared in the learning community and thus again help to construct a shared learning knowledge. For space reasons, we will not go into details of the implementation, but give thumbnails, i.e. the problem and solution statement of the pattern, of the patterns (pattern names are set in SMALL CAPS) that can be identified in the learning gadgets.

### 3.1 One for all and all for one

This learning gadget allows student groups to give distributed slide presentations. These presentations are available to all members of the room in CURE in which the presentations take place. Each room member can join the COLLABORATIVE SESSION at any time by entering the presentation page which offers all functionality for the distributed presentation.

#### COLLABORATIVE SESSION

**Problem:** Users need a shared context for synchronous collaboration. Computer-mediated environments are neither concrete nor visible, however. This makes it difficult to define a shared context and thereby plan synchronous collaboration.

**Solution:** Model the context for synchronous collaboration as a shared session object. Visualize the session state and support users in starting, joining, leaving, and terminating the session. When users join a session, automatically start the necessary collaboration tools.

Before a presentation can be started, one group member has to be selected as presenter. For that purpose, *One for all and all for one* lets group members apply as presenter for a specific topic. Then, the group decides in a VOTE whether the application is accepted or not (see Figure 3) and thereby selects the presenter for a topic.

#### VOTE

**Problem:** It is hard to work out the distribution of opinions in the community. However, good understanding of other users' attitudes can be important when making decisions.

**Solution:** Provide an easy means of setting up and running a poll. Show a virtual ballot in a prominent place in the community. After the vote is over, present the result.

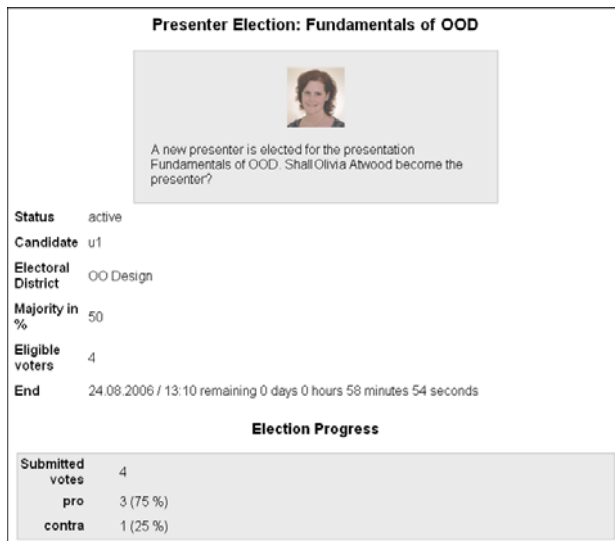


Figure 3: Group VOTE in *One for all and all for one*

In the positive case, the presenter prepares slides and gives a synchronous slide presentation in CURE. For preparing the slides, the presenter can make use of the CURE Wiki syntax. At the beginning of the presentation, the presenter switches the prepared slides to the presentation mode.

Other students who visit the presentation's page will also join the presentation but use a dedicated listener's user interface. This interface includes different means for interacting with the presenter. Most important, it creates a shared focus on a common slide. Whenever the presenter switches the slide, all listeners will automatically follow the presenter, thereby implementing the SHARED BROWSING pattern.

#### SHARED BROWSING

**Problem:** Users have problems finding relevant information in a collaboration space. They often get lost.

**Solution:** Browse through the information space together. Provide a means for communication, and collaborative browsers that show the same information at each client's site.

*One for all and all for one* makes use of the EMBEDDED CHAT in CURE which allows the audience to communicate with each other or with the presenter (see bottom of Figure 4).

#### EMBEDDED CHAT

**Problem:** Users need to communicate. They are used to sending electronic mail. But since e-mail is asynchronous by nature, it is often too slow to resolve issues that arise in synchronous collaboration.

**Solution:** Integrate a tool for quick synchronous interaction into your cooperative application. Let users send short text messages, distribute these messages to all other group members immediately, and display these messages at each group member's site.

Users can decide whether they post a message to the EMBEDDED CHAT or send the message directly to the presenter to ask a question. Questions for the presenter are collected at the presenter's view and can be answered by the presenter as soon as it fits into the presentation. Thereby, *One for all and all for one* implements the FEEDBACK LOOP pattern.

#### FEEDBACK LOOP

**Problem:** In any communication, the recipient of a message can only refer to the message in order to understand it. However, most messages are ambiguous.

**Solution:** Provide an easy means for readers to contact the author. Create a user interface element close to the content that opens input fields for the reader's questions and feedback.

To give the presenter even more feedback, group members can indicate their current degree of understanding in a barometer of opinion. The individual scores for each student are shown in the USER LIST on the left side of the screenshot in Figure 4.

#### USER LIST

**Problem:** Users do not know with whom they are interacting or could interact. Consequently, they do not have a feeling of participating in a group.

**Solution:** Provide awareness in context. Show who is currently accessing an artifact or participating in a COLLABORATIVE SESSION. Ensure that the information is always valid.

The average understanding of the audience is shown to all users in the upper part of the presentation view (see top of Figure 4). The challenge for the presenter is to keep all members of the audience on a high level of understanding, as it is the case for Olivia Atwood in Figure 4.

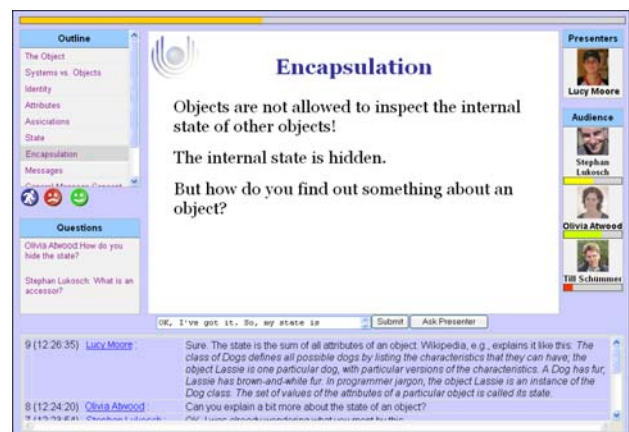


Figure 4: Presentation in *One for all and all for one*



The average understanding is tracked during the whole presentation and will be displayed together with the slides after the presentation is over (see Figure 5).



Figure 5: Average understanding in *One for all and all for one*

Once the presentation is over, the presenter can release the slides for asynchronous access which allows the users to either review the slides only or review the whole COLLABORATIVE SESSION. In the latter case, *One for all and all for one* implements the Persistent Session pattern and not only shows the slides but also the questions and the chat log.

PERSISTENT SESSION

**Problem:** After interacting in a COLLABORATIVE SESSION, users want to resume their collaboration with the results achieved, or want to review them, but the results are not available.

**Solution:** Persistently store the results of a synchronous COLLABORATIVE SESSION on a central server. Keep a master copy of the shared data and track all changes that are applied to it. Let users access the master copy at the central server for review or session resumption purposes.

In summary, *One for all and all for one*

- assists students in learning, as it allows students to collaboratively discuss learning material,
- facilitates shared knowledge construction, as the slides together with the students' feedback can be shared in the learning community and can be used by other students as learning material,
- motivates students to become an active member of the community by giving direct feedback to the presenter, and
- fosters learning in an entertaining way by challenging the presenter to create good feedback curves and the students in the audience to have an understanding that is better than the rest of the audience.

3.2 Fountain of Wisdom

*Fountain of Wisdom* is based on a 3D virtual maze in which two teams compete with each other by answering questions. Additional teams can play in parallel in the same maze. Users can meet on a so-called 3D marketplace and use the EMBEDDED CHAT to socialize with co-learners and propose a game on a specific topic.

Figure 6 shows a screenshot of the marketplace. On the left you can see a list of users that are currently on the marketplace. The bottom of the screenshot shows the chat. In the 3D view, you can see some one additional player, i.e. the snowman, and that the local user is currently proposing a game. The local user has chosen the topic 1678, which is the number for a course on distributed systems, and is going to send an INVITATION to the user Stephan which asks him to participate in the COLLABORATIVE SESSION.

INVITATION

**Problem:** One user wants to interact with another. The other user may be unavailable or busy in another context so that an immediate collaboration would disturb them.

**Solution:** Send and track invitations to the intended participants. Include meta-information on the intended COLLABORATIVE SESSION. Automatically add all users who accept the invitation to the COLLABORATIVE SESSION.

Finally, the local user has chosen to limit the duration of the game to 10 minutes. Another opportunity would have been to limit the game to a number of questions. At the end, the team that correctly answered the most questions wins the game.

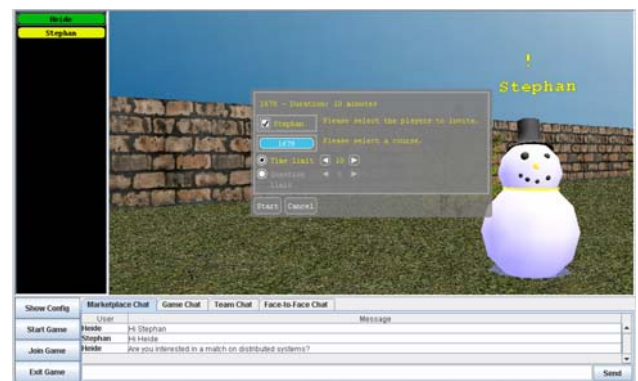


Figure 6: *Fountain of Wisdom* marketplace

After the teams formed, the 3D game maze is initiated. Users that later on visit the marketplace can join the game and decide in which of the two possible teams they want to participate. The underlying system then performs a STATE TRANSFER.

STATE TRANSFER

**Problem:** Users are collaborating in a COLLABORATIVE SESSION but not all of them participate from the beginning. Due to this, some do not know the intermediate results of the COLLABORATIVE SESSION which makes it difficult for them to collaborate.

**Solution:** Transmit the current state of shared objects to latecomers when they join a COLLABORATIVE SESSION. Since

all current participants have the most recent state of the session's shared objects, the system can ask any of the existing clients to perform the state transfer. Ensure the consistency of the state.

The maze provides fountains from which team members have to obtain a question. Complementary to the fountains, there are several sinks in the maze where possible answers to the questions can be found. Each sink contains answers to a number of questions. When users step onto a sink, their view changes and displays the answers to all questions associated with the sink. From this list of possible answers the user has to choose the correct ones knowing only the number of correct answers. Each correct answer counts for individual and team points. Figure 7 shows the view for local users when they step onto a sink. The current question is shown in the header of the 3D view. Left to the 3D view, you can see the USER LIST, which highlights the teams competing with each other. The bottom of the screenshot again shows the EMBEDDED CHAT.

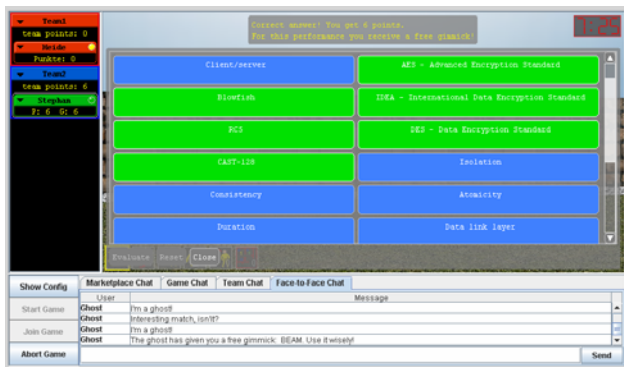


Figure 7: Fountain of Wisdom sink view

Figure 8 shows the normal maze view of *Fountain of Wisdom*. From the view of the local user, you can see another user, i.e. the snowman on the right side, and one bad ghost on the left side. The upper right corner of the 3D view shows a small ACTIVE MAP of the maze with the positions of the other users in the maze and the time that is left in this match.

#### ACTIVE MAP

**Problem:** To orient themselves and interact in space, users have to create a mental model that represents the space and the artifacts and users it contains. This is a difficult task.

**Solution:** Create a reduced visual representation of the spatial domain mode lby means of a map. Show other users' locations on the map. Ensure that the map is dynamic for artifacts and users, but static with respect to landmarks.

At the bottom of the 3D view you can see different chats, i.e. one for the marketplace, one for all participants of the current game, one for the team, and one for the user next to oneself. The chats can be used to gain information about where to find the correct sink for a question and to discuss possible answers with team members.

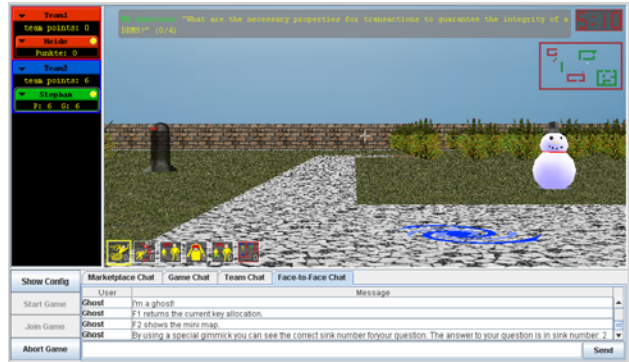


Figure 8: Fountain of Wisdom maze view

Apart from bad ghosts there are also good ghosts in the maze. Good ghosts help the players by e.g. giving tips on the correct answer to a question or giving away gimmicks that allow the user to perform special actions, e.g. to move faster or beam from one place to another in the maze. If a player comes too close to a bad ghost, this may steal the current question and the player has to get a new one.

Prior to a game, students can define questions and answers for a specific topic as CURE pages. CURE supports so-called Wiki-templates [8], which allow end-users to define form-based pages. These Wiki-templates can be used to structure shared knowledge and simplify its construction. Figure 9 shows a screenshot of a typical question and answer page in CURE. For all question and answer pages the same Wiki-template is used to enable re-use in different learning gadgets. Apart from the question, such a page consists of a number of correct answers, a question category, the specification of the course, and a helping text that is used by the good ghosts to provide hints about the correct answer of the question.

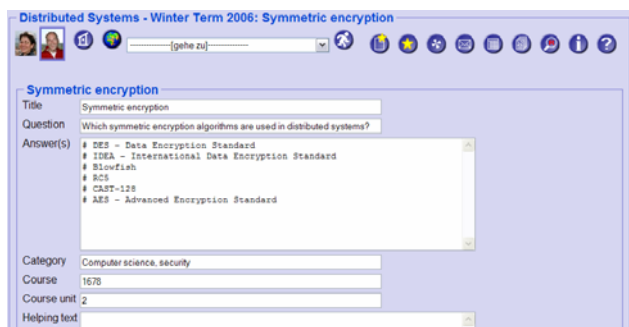


Figure 9: Question and answer page in CURE

The question and answer pages are elements of a shared question repository in CURE and define the topics that are available for a game. To ensure the quality of the questions, CURE, implements a combination of the patterns QUALITY INSPECTION and VOTE.

#### QUALITY INSPECTION

**Problem:** Members participate in a community to enjoy high-quality contributions from fellow members. However, not every contribution has the same quality. Low-quality contributions can annoy community members and distract their attention from high-quality gems.



**Solution:** *Select users as moderators and let them release only relevant contributions into the community's interaction space. Give moderators the right to remove any contribution and to expel users from the community.*

Due to the implementation of the patterns QUALITY INSPECTION and VOTE, students can rate the quality of a question and the corresponding answers using a special page CURE provides (see Figure 9). The ratings of all students are accumulated and shown as stars in the last column of the table in Figure 9. Additionally, students can act as quality inspectors by removing questions from the repository that are of low quality.

Course	Course unit	Question	Rating
1678	2	What kinds of group awarene...	☆☆☆
1678	2	What asymmetric encryption ...	☆☆☆
1678	2	What are the different laye...	☆☆☆
1678	2	What kind of network archit...	☆☆☆
1678	2	What are the necessary prop...	☆☆☆

Figure 9: Rating in the shared question repository

#### Summarizing, *Fountain of Wisdom*

- increases social interaction, as students can meet on the marketplace to form teams and during a game the different chats allow cross-group interaction,
- strengthens the group feeling, as the students cooperate in groups to gain as much group points as possible, and
- allows students to self-organize their learning and their interaction, as each user can create rooms in CURE and define the interaction possibilities of a room, and
- supports collaborative learning by constructing a shared question repository that can be used by all students of the CURE environment.

## 4 Related Work

Most web-based learning platforms focus more on the distribution of learning material than social interaction or possibilities to construct shared knowledge. *Blackboard* [3] offers a virtual classroom, in which interaction can take place. Students and teachers can collaborate on a shared whiteboard and communicate via chat. To support peer collaboration teachers can offer group projects in which each group can be given its own file exchange area, discussion board, etc. But the results of these group projects cannot become part of a shared knowledge.

*Centra Live for Virtual Classes* [5] organizes collaborative learning activities as events. Students can enrol in events by browsing a catalogue of upcoming events. For these events, teachers can plan a variety of synchronous interaction using the provided functionality, e.g. polls, surveys, chat, whiteboards, cooperative Web browsing, etc. Though these tools might foster interaction, the collaboration is not self-initiated and does not focus on the construction of a shared knowledge.

*Moodle* [19] offers a lot of activity modules that can be associated with learning material. Among these activity modules are, e.g., a Wiki, a glossary, or a quiz. The quiz mainly focuses on assessment. The glossary and the Wiki allow interaction and collaboration among the students but must be enabled by the teacher. As the teacher has to enable all these activity modules self-initiated interaction and collaboration among the students is not possible.

*BSCW* [26] is based on *BSCW* and offers a group of students a web-based shared workspace. Compared to other learning platforms, *BSCW* offers special tools for collaborative knowledge building, but again, all interaction and collaboration possibilities are defined by the teacher.

*ILIAS* [13] or *KOLUMBUS* [12] are further learning environments that offer similar functionalities, but like the above platforms do not support entertaining interaction.

Compared to other learning environments, CURE in combination with the described learning gadgets offers a much higher degree of collaborative interaction and participation possibilities. In our opinion, this provides a good means to motivate students for continuous participation in a learning community. The question repository and the shared slide presentation allow students to construct a persistent shared knowledge which might serve as glue for forming a learning community. Students will not feel isolated anymore as they on the one hand can benefit from the existing repository and on the other hand can become an active member of the learning community by participating in the construction of the shared repository.

## 5 Conclusion

Up to now there is no long-time study of CURE in combination with the learning gadgets. At the end of the lab course, we asked our students if they would use the learning gadgets for collaborative learning. Almost all students indicated their interest. Additionally, the participatory design of the learning gadgets, which is based on the re-use of proven solutions in the form of patterns for computer-mediated interaction [24], suggests that the learning gadgets will be accepted and used by the students. For that reason, we are currently planning long-time studies of the learning gadgets and add the learning gadgets to the deployed version of CURE. We plan to evaluate the usage and impact of the learning gadgets on how they foster collaborative interaction, shared knowledge construction, and the building of a learning community.

## Acknowledgement

Special thanks are due to Mohamed Bourimi and Till Schümmer for their engagement in supervising two of the lab groups and to all members of the lab groups (in alphabetical order): Yves Albrecht, Lukas Beyer, Marco Blum, Christian Brandtner, Kathrin Dentler, Volker Engels, Thomas Grasse, Matthias Hellweg, Knut Linke,

Irini Ntokoutsi, Frank Plieninger, Martin Rasel, Franz Schinerl, and Julia Schmeisser.

## References

- [1] Alexander, C., Ishikawa, S., Silverstein, M., Jacobson, M., Fiksdahl-King, I., Angel, S., (1977). A pattern language. Oxford University Press, New York.
- [2] Alexander, C., (1979). The timeless way of building. Oxford University Press, New York.
- [3] Blackboard Inc. (2006). Product homepage. <http://www.blackboard.com/>, last visited March 2007.
- [4] Brugali, D., Menga, G., Aarsten, A. (1997). The framework life span. *Communications of the ACM* 40 (10), pages 65–68.
- [5] Centra Live Suite (2007). Product homepage. <http://www.saba.com/products/centra/live/>, last visited March 2007.
- [6] Erickson, T. (2000). Lingua francas for design: sacred places and pattern languages. In: *Proceedings of the conference on Designing interactive systems*. ACM Press, pages 357–368.
- [7] Greenberg, Saul and Mark Roseman (2003). Using a Room Metaphor to Ease Transitions in Groupware. In M. Ackermann, V. Pipek, and V. Wulf, editors, *Sharing Expertise: Beyond Knowledge Management*, pages 203–256. MIT Press, Cambridge, MA, USA.
- [8] Haake, Anja, Stephan Lukosch, and Till Schümmer (2005). Wiki-Templates: Adding Structure Support to Wikis On Demand. In: *WikiSym 2005 – Conference Proceedings of the 2005 International Symposium on Wikis*, Seiten 41-51. ACM Press.
- [9] Haake, Jörg. M, Till Schümmer, Anja Haake, Mohamed Bourimi, and Britta Landgraf (2004). Supporting flexible collaborative distance learning in the CURE platform. In *Proceedings of the Hawaii International Conference On System Sciences (HICSS-37)*. IEEE Press.
- [10] Haake, Jörg M., Anja Haake, Till Schümmer, Mohamed Bourimi, and Britta Landgraf (2004). End-User Controlled Group Formation and Access Rights Management in a Shared Workspace System. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 554–563, ACM Press.
- [11] Haythornthwaite, Caroline, Michelle M. Kazmer, Jennifer Robins, and Susan Shoemaker (2000). Community Development Among Distance Learners: Temporal and Technological Dimensions. *Journal of Computer-Mediated Communication*, 6(1).
- [12] Herrmann, Thomas and Andrea Kienle (2003). KOLUMBUS: Context-oriented communication support in a collaborative learning environment. *Informatics and the Digital Society. Social, Ethical, and Cognitive Issues*, Kluwer, pages 251-260.
- [13] ILIAS (2007). ILIAS open source. <http://www.ilias.de/>, last visited March 2007.
- [14] Johnson, R. E. (1997). Frameworks = (components + patterns). *Communications of the ACM* 40 (10), 39–42
- [15] Lave, J. and Wenger, E. (1991). *Situated Learning Legitimate Peripheral Participation* Cambridge University Press.
- [16] Leuf, Bo and Ward Cunningham (2001). *The WIKI way*. Addison-Wesley, Boston, MA, USA.
- [17] Lukosch, Stephan and Till Schümmer (2006). Groupware Development Support with Technology Patterns. *International Journal of Human Computer Studies*, Special Issue on 'Theoretical and Empirical Advances in Groupware Research', 64(7):599-610.
- [18] Lukosch, Stephan and Till Schümmer (2006): Making exam preparation an enjoyable experience. *International Journal of Interactive Technology and Smart Education*, Special Issue on 'Computer Game-based Learning', 3(4):259-274.
- [19] Moodle, (2007). Moodle - A Free, Open Source Course Management System for Online Learning. <http://moodle.org/>, last visited March 2007.
- [20] Palloff, Rena M. and Keith Pratt (1999). *Building Learning Communities in Cyberspace – Effective Strategies for the Online Classroom*. Jossey Bass Wiley.
- [21] Pfister, Hans-Rüdiger, Christian Schuckmann, Jennifer Beck-Wilson, and Martin Wessner (1998). The Metaphor of Virtual Rooms in the Cooperative Learning Environment CLear. In *Cooperative Buildings - Integrating Information, Organization and Architecture. Proceedings of CoBuild'98*, LNCS 1370, pages 107–113, Springer-Verlag Berlin Heidelberg.
- [22] Preece, Jenny (2000). *Online Communities*. Wiley, Chichester, UK.
- [23] Prensky, Marc (2001). *Digital Game-Based Learning*. McGraw-Hill Education.
- [24] Schümmer, Till and Stephan Lukosch (2007). Patterns for Computer-Mediated Interaction. John Wiley and Sons, Chichester, UK.
- [25] Schümmer, Till, Stephan Lukosch, and Robert Slagter (2006). Using Patterns to empower End-users The Oregon Software Development Process for Groupware. *International Journal of Cooperative Information Systems, Special Issue on '11th International Workshop on Groupware (CRIWG'05)'*, 15(2):259-288.
- [26] Stahl, Gerry (2002). Groupware Goes to School. In *Groupware: Design, Implementation, and Use, 8th International Workshop, CRIWG 2002*, LNCS 2440, pages 7-24, Springer-Verlag Berlin Heidelberg.
- [27] Wenger, Etienne, (1998). *Communities of Practice: Learning, Meaning, and Identity*. Cambridge University Press.

# A Fuzzy Classification Model for Online Customers

Andreas Meier and Nicolas Werro  
 University of Fribourg, Switzerland  
 E-mail: Andreas.Meier@unifr.ch, Nicolas.Werro@unifr.ch

**Keywords:** electronic business, webshop, fuzzy classification, online customer, customer relationship management

**Received:** March 1, 2007

*Building and maintaining customer loyalty are important issues in electronic business. By providing customer services, sharing cost benefits with online customers, and rewarding the most valued customers, customer loyalty and customer equity can be improved. With conventional marketing programs, groups or segments of customers are typically constituted according to a small number of attributes. Although corresponding data values may be similar for two customers, they may fall into different classes and be treated differently. With the proposed fuzzy classification model, however, customers with similar behavior and qualifying attributes have similar membership functions and therefore similar customer values. The paper illustrates how webshops can be extended by a fuzzy classification model. This allows webshop administrators to improve customer equity, launch loyalty programs, automate mass customization and personalization issues, and refine marketing campaigns to maximize the real value of the customers.*

*Povzetek: Razvit je model za določanje lojalnosti internetnih kupcev.*

## 1 Motivation

Within what is now a global market, the attention span of the customer has decreased. The customer behaves more individually, and customer loyalty has become difficult to maintain. A successful company needs, therefore, to increase the value it provides to its customers. Furthermore, if a company cannot react quickly to changing customer needs, the customer will find someone else who can.

The World Wide Web has created a challenging arena for e-commerce: with a webshop, products and services can be offered to online customers. In this context, two specific strategic goals must be addressed. First, new online customers, or lost customers, have to be acquired; these customers should have attractive market and resource potential. The second strategic goal is to maintain and improve customer equity; this can be achieved by cross-selling and up-selling, and through programs aimed at lifetime customer retention (Blattberg et al. 2001).

Managing online customers as an asset requires measuring them and treating them according to their true value. With the sharp customer classes of conventional marketing methods this is not possible (see example in Section 3.1). Here a fuzzy model is proposed for the classification of online customers. With fuzzy classification, an online customer can be treated as a member of a number of different classes at the same time. Based on these membership functions, the webshop owner can devise appropriate marketing programs for acquisition, retention, and add-on selling.

A number of fuzzy classification approaches have been proposed in the marketing literature. Twenty years

ago, Hruschka (1986) proposed a segmentation of customers using fuzzy clustering methods. A clusterwise regression model for simultaneous fuzzy market structuring was discussed by Wedel and Steenkamp (1991). Hsu's Fuzzy Grouping Positioning Model (2000) allows an understanding of the relationship between consumer consumption patterns, and a company's competitive situation and strategic positioning. The modeling of fuzzy data in qualitative marketing research was also described by Varki et al. (2000). Finally, a fuzzy Classification Query Language (fCQL) for customer relationship management was proposed by Meier et al. (2005). Most of the cited research literature applies fuzzy control to classical marketing issues. Up to now, fuzziness has not yet been adapted for e-business, e-commerce, and/or e-government. In our research work, the power of a fuzzy classification model is used for an electronic shop. Online customers will no longer be assigned to classical customer segments but to fuzzy classes. This leads to differentiated online marketing concepts and helps to improve the customer equity of webshop users.

This paper describes an extension of the webshop eSarine (see Werro et al. 2004) with a fuzzy classification model for online customers. The remainder of the paper is structured as follows: Section 2 presents the main processes and repositories of a webshop, introduces the fuzzy classification concept, model and query language, and briefly describes the architecture of the fCQL toolkit. In Section 3, aspects of customer equity, mass customization and online marketing campaigns based on fuzzy customer classes are

illustrated with examples. Section 4 depicts a generic hierarchy of fuzzy classification and proposes a controlling loop for online customers. Finally Section 5 gives a conclusion.

## 2 Webshop with Fuzzy Classification

### 2.1 Business Processes and Repositories

A webshop (often called electronic shop or online shop) is a web-based software system that offers goods and services, generates bids/offers, accepts orders and carries out delivery and modes of payment. In principle, each webshop consists of a storefront and a backfront. The online customers only have access to the storefront and can seek information on products and services, order as required, pay and receive their product. Access to the backfront is reserved to the webshop administrator. Here, products and services are inserted into the product catalog and the different procedures for ordering, paying, and purchasing are specified.

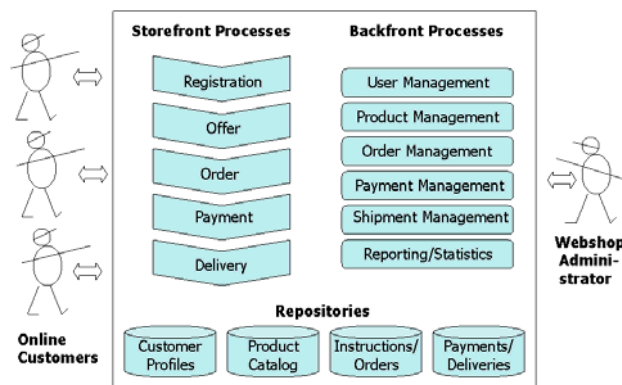


Figure 1: Logical Components of a Webshop

The most important processes and repositories of a webshop are presented in Fig. 1:

- Registration of online customers: a visitor to the electronic shop can find out about the products and services. Those intending to buy will communicate minimum data about themselves and establish user profiles along with payment and delivery arrangements.
- Customer profiles and customer administration: the data on customers is put into a database. In addition, an attempt is made to put together specific profiles based on customer behavior. This allows new, but relevant offers to be presented to the individual customer. However, the rules of communication and information desired by the user must be respected (e.g. customized push for online advertising).
- Product catalog: the products and services are listed in the catalog, grouped into categories so that the webshop can be clearly organized. Products may be listed with or without prices. With individual customer pricing, a quotation is computed and specified during the drawing up of the offer, which will also reflect the discount system selected.
- Offering and ordering: offers can be generated and goods and services bought as needed. The electronic shopping basket or cart is used by online customers to reserve the goods and services selected for possible purchase and show the total price with discount.
- Shipment options: where digital product categories are offered by webshops, goods and services can be delivered online.
- Measures for customer relationship management: online customer contact is maintained after a purchase by offering important after-sales information and services. These measures make customer contact possible when these goods and services are used, thus enhancing the customer connection.

To attract potential online customers of high quality and to retain and extend their customer value, a fuzzy classification model is helpful. The next sections present a model which allows companies to derive customer equity and treat online customers according to their real value.

### 2.2 Fuzzy vs. Sharp Classification

Fuzzy logic aims to capture the imprecision of human perception and to express it with appropriate mathematical tools. With the fuzzy classification model proposed in the next Section 2.3, marketers are able to use linguistic variables, such as ‘loyalty’, and linguistic terms like ‘high’ or ‘low’.

There are a number of advantages in using fuzzy classification for relationship management:

- Fuzzy logic, unlike statistical data mining, enables the use of non-numerical attributes. As a result, both qualitative and quantitative attributes can be used for marketing acquisition, retention, and add-on selling.
- With the help of linguistic variables and terms, marketers may describe equivalence classes more intuitively (excellent loyalty, medium loyalty, weak loyalty). The definition of linguistic variables and terms and the naming of fuzzy classes can be derived directly from the terminology of marketing and sales departments.
- Customer databases can be queried on a linguistic level. For example, the fuzzy Classification Query Language (Meier et al. 2005) allows marketers to classify single customers or customer groups by classification predicates such as ‘loyalty is high and turnover is large’.

An important difference between a fuzzy classification and a sharp one is the fact that a customer can belong to more than one fuzzy class. In conventional marketing programs, groups or segments of customers are typically constituted by a small number of qualifying attributes. If corresponding data values are similar for two customers, their membership functions are similar too. In the conventional case however, they may fall into different

classes and be treated differently (see customers Brown and Ford in Fig. 4).

With fuzzy classification it is possible to treat each customer individually. This allows managers to allocate marketing budgets more precisely. In addition, cost savings can be achieved. For instance, when offering a discount (see Section 3.2), discount rates can be chosen according to the individual customer value. Companies can try to retain the more profitable customers by giving them individualized privileges.

Needless to say there are also drawbacks when applying fuzzy classification. The definition process of a fuzzy classification remains a challenging task. In our experience, the design of fuzzy classes requires marketing specialists as well as data architects and webshop administrators. Beyond this, a methodology is needed for the entire planning, designing, and testing process for appropriate fuzzy classes.

### 2.3 Fuzzy Classification Model

The relational database of online customers (see Customer Profiles in Fig. 1) is extended by a context model in order to obtain a classification space. To every attribute  $A_j$  defined by a domain  $D(A_j)$  there is added a context  $C(A_j)$ . A context  $C(A_j)$  of an attribute is a partition of  $D(A_j)$  into equivalence classes (see Shenoj 1995). In other words, a relational database schema with contexts  $R(A,C)$  consists of a set  $A=(A_1,\dots,A_n)$  of attributes and the set  $C=(C_1(A_1),\dots,C_n(A_n))$  of associated contexts.

Throughout this paper, an illustrative example from relationship management is used. For simplicity, online customers will be evaluated by only two attributes, turnover and loyalty. In addition, these two qualifying attributes for customer equity will be partitioned into only two equivalence classes. The pertinent attributes and contexts for relationship management are:

- Turnover in Euro per month: the attribute domain is defined by  $[0..1000]$  and divided into the equivalence classes  $[0..499]$  for small and  $[500..1000]$  for large turnover.
- Loyalty: the domain {excellent, good, mediocre, bad} with its equivalence classes {excellent, good} for high and {mediocre, bad} for low loyalty behavior.

To derive fuzzy classes from sharp contexts, the qualifying attributes are considered as linguistic variables, and verbal terms are assigned to each equivalence class. With linguistic variables, the equivalence classes can be described more intuitively. In addition, every term of a linguistic variable represents a fuzzy set. Membership functions  $\mu$  (see Zimmermann 1992) are defined for the domains of the equivalence classes.

As turnover is a numeric (sharp) attribute, its membership functions  $\mu_{large}$  and  $\mu_{small}$  are continuous functions defined on the whole domain of the attribute. For qualitative attributes like loyalty, step functions are used; the membership functions  $\mu_{high}$  and  $\mu_{low}$  define a

membership grade for every term of the attribute's domain.

The selection of the two attributes 'turnover' and 'loyalty' and the corresponding equivalence classes determine a two-dimensional classification space (see Fig. 2). The four resulting classes C1 to C4 could be characterized by marketing strategies such as 'Commit Customer' (C1), 'Improve Loyalty' (C2), 'Augment Turnover' (C3), and 'Don't Invest' (C4).

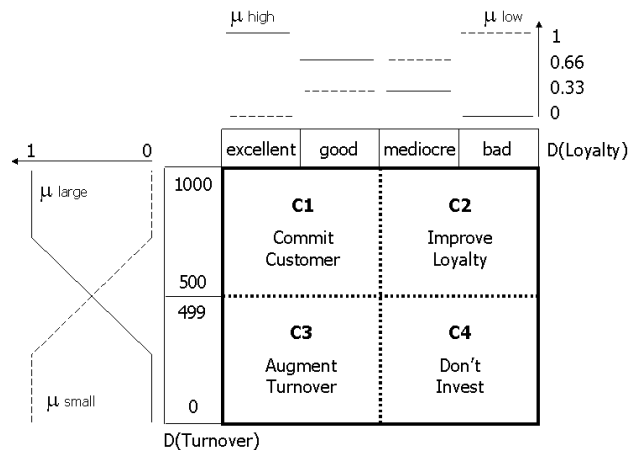


Figure 2: Fuzzy Classification Space defined by Turnover and Loyalty

The selection of qualifying attributes, the introduction of equivalence classes and the choice of appropriate membership functions are important design issues (Meier et al. 2001). Database architects and marketing specialists have to work together in order to make the right decisions.

With the proposed context model, the use of linguistic variables and membership functions, the classification space becomes fuzzy. A fuzzy classification of online customers has many advantages compared with common sharp classification approaches (see Section 2.2 and Section 3). Most importantly, with fuzzy classification a customer can belong to more than one class at the same time. This leads to differentiated marketing concepts and helps to improve customer equity.

### 2.4 Fuzzy Classification Query Language

The classification language fCQL is designed in the spirit of SQL (Schindler 1998). Instead of specifying the attribute list in the select clause, the name of the object column to be classified is given in the **classify** clause. The **from** clause specifies the considered relation, just as in SQL. Finally, the where clause is changed into a **with** clause which specifies a classification predicate.

An example in customer relationship management could be given as follows:

```

classify Customer
from CustomerRelation
with Turnover is large and
      Loyalty is high
    
```

This classification query would return the class C1 (Commit Customer) defined as the aggregation of the terms ‘large’ turnover and ‘high’ loyalty. The aggregation operator is the  $\gamma$ -operator which was suggested as compensatory and was empirically tested by Zimmermann and Zysno (1980).

In this simple example, specifying linguistic variables in the **with** clause is straightforward. However, if customers are classified on three or more attributes, the capability of fCQL for a multi-dimensional classification space is increased. This can be seen as an extension of the classical slicing and dicing operators on a multidimensional data cube.

### 2.5 Architecture of the fCQL Toolkit

The architecture of the fCQL toolkit shown in Fig. 3 illustrates the interactions between the user (resp. the webshop), the fCQL toolkit and the relational database management system containing the different repositories of the webshop.

The fCQL toolkit is an additional layer above the relational database system; this particularity makes fCQL independent of underlying database systems and thus enables fCQL to operate with every commercial product. It also implies that the user (resp. the webshop application) can always query the database with standard SQL commands (see case 1).

Before querying the fCQL toolkit, the data architect has to define the fuzzy classification (see case 2).

The user can now formulate unsharp queries to the fCQL toolkit (see case 3). Those queries are analyzed and translated into corresponding SQL statements for the database system. Then the classification results are displayed to the user or returned to the webshop software.

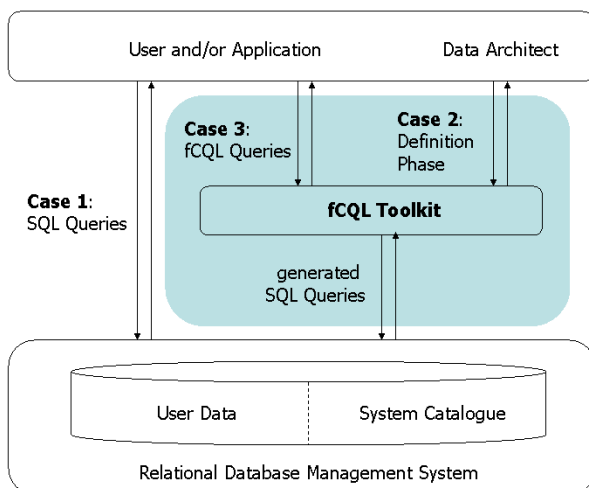


Figure 3: Architecture of the fCQL Toolkit

## 3 Fuzzy Classes for Online Customers

### 3.1 Customer Equity

Managing online customers as an asset requires measuring them and treating them according to their true value. With sharp classes, i.e. traditional customer segments, this is not possible. In Fig. 4 for instance, customers Brown and Ford have similar turnover as well as similar loyalty behavior. However, Brown belongs to the winner class C1 (Commit Customer) and Ford to the loser class C4 (Don’t Invest). In addition, a traditional customer segment strategy treats the top rating customer Smith in the same way as Brown, who is close to the loser Ford.

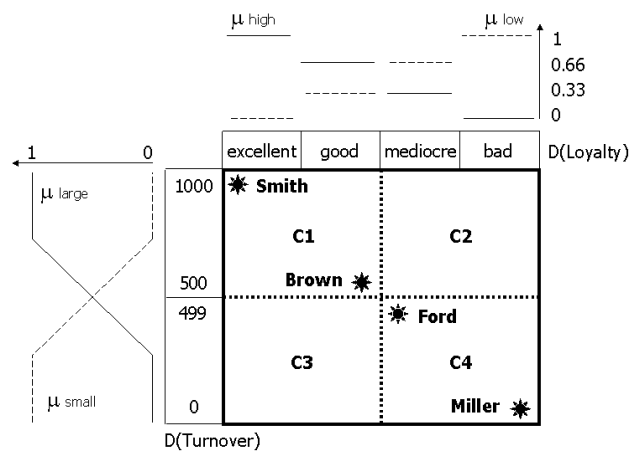


Figure 4: Customer Equity Examples based on Turnover and Loyalty

With a sharp classification, the following drawbacks can be observed:

- Customer Brown has no advantage from improving his turnover or his loyalty behavior as he already receives all the privileges of the premium class C1.
- Brown will be surprised and disappointed if his turnover or loyalty decreases slightly and he therefore falls into another class. He may even fall from the premium class C1 directly into the loser class C4.
- Customer Ford, potentially a good customer, may find opportunities elsewhere. As he belongs to the loser class C4, he is treated in the same way as Miller although he has higher turnover and better loyalty.
- The most profitable online customer with excellent loyalty is Smith. Sooner or later he will become confused. Although he belongs to the premium class C1, he is not treated according to his real value. In comparison with Brown, he might be disappointed by webshop offers or services.

The dilemmas described can be solved by applying a fuzzy classification as shown in Fig. 4: The main difference between a traditional classification and a fuzzy one is that in the fuzzy classification an online customer



can belong to more than one class. Belonging to a fuzzy class implies a degree of membership. The notion of membership functions results in the disappearance of sharp borders between customer segments. Fuzzy customer classes reflect reality better and allow the webshop administrators to treat online customers according to their real value.

### 3.2 Issues of Mass Customization and Personalization

Customization and low cost is often mutually exclusive. Mass production provides low cost but at the expense of uniformity. Mass customization is defined as customization and personalization of products and services for individual customers at a mass production price (Pine and Davis 1999).

Digital goods and services are costly to produce but cheap to reproduce. In addition, versioning of products and services can easily be achieved. Another advantage of fuzzy classification, therefore, is its potential for personalized privileges. For instance, the membership degree of online customers can determine the privileges they receive, such as a personalized discount (Werro et al. 2005). Discount rates can be associated with each fuzzy class: in the following example C1 (Commit Customer) has a discount rate of 10%, C2 (Improve Loyalty) one of 5%, C3 (Augment Turnover) 3%, and C4 (Don't Invest) 0%. The individual discount of an online customer could be calculated by the aggregation of the discounts of the classes he belongs to in proportion to his various degrees of membership.

The top rating customer Smith belongs 100% to class C1 because he has the highest possible turnover as well as the best loyalty behavior; the membership degree of Smith in class C1 would be written as Smith (C1:1.0, C2:0.0, C3:0.0, C4:0.0). Customer Brown belongs to all four classes and would be rated as (C1:0.28, C2:0.25, C3:0.25, C4:0.22). With fuzzy classification, the online customers of Fig. 4 receive the following discounts:

- Smith (C1: 1.0, C2: 0.0, C3: 0.0, C4: 0.0):  
 $1.0 \cdot 10\% + 0.0 \cdot 5\% + 0.0 \cdot 3\% + 0.0 \cdot 0\% = 10\%$
- Brown (C1:0.28, C2:0.25, C3:0.25, C4:0.22):  
 $0.28 \cdot 10\% + 0.25 \cdot 5\% + 0.25 \cdot 3\% + 0.22 \cdot 0\% = 4.8\%$
- Ford (C1:0.22, C2:0.25, C3:0.25, C4:0.28):  
 $0.22 \cdot 10\% + 0.25 \cdot 5\% + 0.25 \cdot 3\% + 0.28 \cdot 0\% = 4.2\%$
- Miller (C1: 0.0, C2: 0.0, C3: 0.0, C4: 1.0):  
 $0.0 \cdot 10\% + 0.0 \cdot 5\% + 0.0 \cdot 3\% + 1.0 \cdot 0\% = 0\%$

Using fuzzy classification for mass customization and personalization leads to a transparent and fair judgment: Smith gets the maximum discount and a better discount than Brown who belongs to the same class C1. Brown and Ford have nearly the same discount rate. They have comparable customer values although they belong to opposing classes. Miller, who is in the same class as Ford, does not benefit from a discount.

Applying the fuzzy classification model with personalized discounts has additional advantages: first, all online customers of a webshop are motivated to

improve their buying attitude and/or loyalty behavior. Second, only a small group of the premium class C1 gets the 10% discount; the same is true for classes C2 and C3. In other words, the total budget for personalized discounts will be smaller compared with conventional discount methods. The savings can then be used for acquisition or retention programs, i.e. online marketing campaigns.

### 3.3 Online Marketing Campaign

Launching a marketing campaign can be very expensive. It is therefore crucial to select a customer group with potential. Fuzzy classification offers considerable advantages when planning and selecting customer subgroups.

An example of a fuzzy-controlled marketing campaign is given in Fig. 5. Here, the strategy is to select loyal customers with low turnover. Using membership functions, a subset of customers in class C3 can be chosen. The application of membership functions allows marketers to dynamically modify the size of the target group in relation to the available campaign budget. Modifying the size of the target group is also a valuable mean to increase or decrease the homogeneity between the targeted customers (Nguyen et al. 2003).

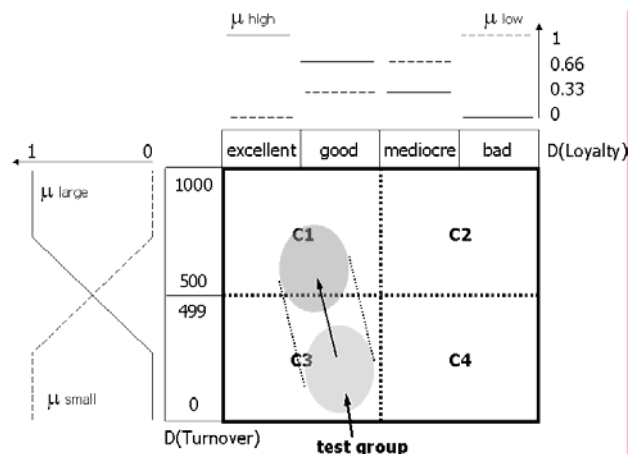


Figure 5: Development of the Target Group as a Result of a Marketing Campaign

Once the marketing campaign or testing process has been started, the fuzzy customer classes can be analysed again. It is important to find out if the money invested is moving the customers in the planned direction, i.e. improving their customer value.

With fuzzy classification, marketers can monitor the development of customers or customer groups (see Fig. 5). By comparing the value of a customer over time, it is possible to determine whether an online customer has increased, maintained, or decreased in customer value. The most useful application of monitoring customers could be the detection of churning customers: automated triggers can respond to the development of customer values; if a good customer begins to show churning behavior, an alert to the marketing department may help to retain this customer.



## 4 Hierarchical Fuzzy Classification

### 4.1 Analysis of Online Customers

The analysis of online customers compared to traditional ones has the advantage that a lot of information about the customers' behavior is automatically logged in the system.

In a webshop application, explicit and implicit information can be used for the analysis of the customer relationships. Explicit information is the data provided directly by the customer, like the orders or the product ratings and forum entries:

- Orders: the orders can determine the *turnover* and the *margin* of a customer as well as his *buying frequency*. Indirectly, the *payment delay* and the *return rate* can also be identified.
- Product ratings and forum entries: the ratings of products as well as forum entries can reveal the *involvement frequency*.

Implicit information is retrieved from the interaction and the behavior of the online user with the webshop:

- Clickstream: the clickstream information can establish the *visiting frequency* and the behavior of visitors.

The use of the order information is straightforward as most of the companies are using the turnover, margin, payment delay and return rate information in order to analyse and reward their customers. The clickstream data is a new source of information coming from the interaction of the users with the online shop. All the actions performed by the users are logged with a time stamp into the clickstream data allowing the system to determine the visiting frequency for each user (Lee et al. 2001). Many online shops provide their customers a way to communicate and to share knowledge by means of product ratings and forum entries. This social involvement can lead to a virtual community which increases the trust and the attachment towards the company (Rheingold 1993).

### 4.2 Hierarchy of Fuzzy Classification

In real applications, a fuzzy classification database schema can have a number of attributes, linguistic variables, and terms. This leads to a multi-dimensional classification space with a large number of classes. After combining all these attributes, it may not be possible to extract clear semantics for each resulting class. This problem, also present with sharp classifications, is partially resolved by the use of fuzzy classes. By having a continuous transition between the classes, fewer equivalence classes (linguistic terms) are required. The problem of complexity remains, however, if the number of dimensions (linguistic variables) increases exponentially with the number of classes.

In order to maintain classes with a proper semantic, a multi-dimensional fuzzy classification space can be decomposed into a hierarchy of fuzzy classification

levels (Werro et al., 2006). By grouping attributes of a given context in sub-classifications, it is possible to derive meaningful definitions for each of the classes. The decomposition of a multi-dimensional fuzzy classification space also reduces the complexity and allows optimization during the modeling phase.

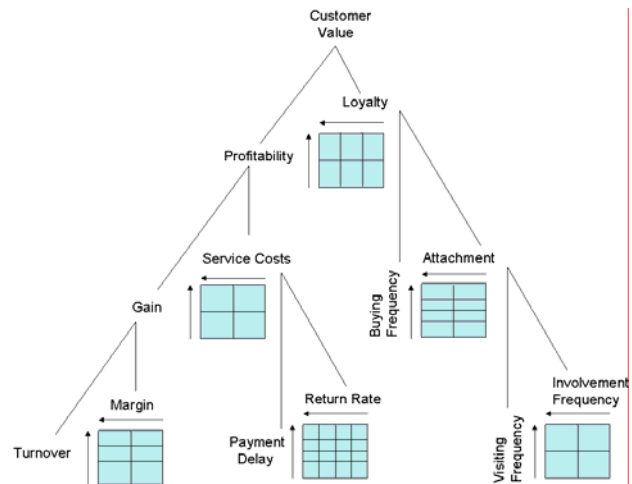


Figure 6: Customer Value as Hierarchical Fuzzy Classification

An example of a hierarchical fuzzy classification space could be the calculation and the controlling of customer equity. Customer equity not only has to deal with monetary assets (turnover, margins, service costs etc.) but also with hidden assets such as loyalty or attachment. Harrison (2000), for instance, proposes to express customer loyalty based on two dimensions, attachment and buying behavior. Taking these different aspects of customer equity into account, a hierarchy of fuzzy classes for the calculation of customer values could be derived (see Fig.6). In the proposed example, the customer value depends on the two linguistic variables profitability and loyalty, loyalty on buying frequency and attachment, attachment on visiting frequency and involvement frequency and so on. It is important to note that marketers can evaluate the fuzzy classification space in a more structured way and with clearly defined semantics at every level of the hierarchy. If, for instance, loyalty problems seem to occur for a given customer, then his buying, visiting and involvement behavior can be studied in more depth in order to evaluate retention programs.

### 4.3 Closed Loop for Controlling

In recent years, managers in a range of industries have been rethinking how to measure the performance of their businesses. They have recognized that a shift is needed towards treating financial statistics in the context of a broader set of measures. Edvinsson and Malone, for instance, propose the introduction of an intellectual capital report which brings together indicators from finance, customer base, process management, renewal, and development as well as from human resources

(Edvinsson & Malone 1997). The customer focus requires indicators such as market share, number of customers, customer equity, customers lost, average duration of customer relationship, ratio of sales contacts to sales responses, satisfied customer index and service expenses per customer.

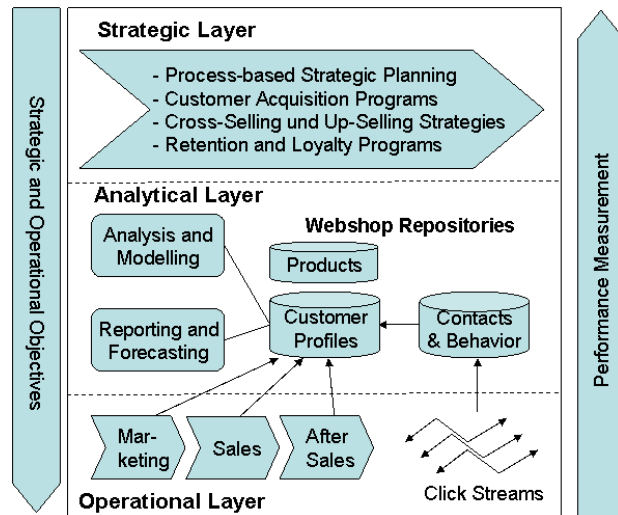


Figure 7: Closed Loop for Controlling Online Customers and their Behavior

Fig. 7 illustrates the closed loop for controlling online customer relationships. This loop has been implemented in the webshop eSarine with the help of a performance measurement system (Küng et al. 2001). At the strategic level, objectives for online customer acquisition, retention, and add-on selling must be defined, as must also the process and service quality goals of the webshop. The traditional tasks in marketing, sales and after-sales activities will be carried out. Applications for collaborative services such as customized push, personalized offers and care of e-communities will also be developed. In addition, all customer contacts information, i.e. click streams, has to be analyzed and stored in a contacts database.

The glue between the strategic and operational layer is the customer profile database with fuzzy classes, extended by the contacts database. The webshop administrator or a specialized team is responsible for a consistent customer database, and for analysis of the contacts and the behavior of online customers. The fuzzy classification model, with a corresponding query facility, allows the webshop owner to improve customer equity, launch loyalty programs, automate mass customization issues, and refine online marketing campaigns.

## 5 Conclusion

The fuzzy classification approach and the fCQL toolkit are more than just another concept and piece of software. Fuzzy classification can be seen as a management method and the fCQL toolkit is a powerful instrument for analysis and control of a business:

- *Strategic Management:* For the analysis of markets, fuzzy classification allows demographic, geographic, behavioral and psychographic market segmentations. It is more successful and realistic to fuzzily target markets and to fuzzily position brands or companies in their markets.
- *Customer Relationship Management:* For customer analysis and segmentation, fuzzy customer classes give the marketers a differentiated judgment of customers and customer groups. In addition, if customer value is calculated as an aggregated membership degree (see Section 4.2) then customer equity is based on both monetarily-based and hidden assets.
- *Supply Chain Management:* With a fuzzy approach, it is possible to classify, analyse and evaluate different suppliers and their delivery processes. A fuzzy supplier rating and/or fuzzy judgment of quality and time schedules of the delivery processes provides for more differentiated planning. For instance, improvements in the delivery system can be effected by observing moving targets in fuzzy classes.
- *Total Quality Management:* Quality measures are not only numeric; there are also qualitative measures. The equal treatment of quantitative and qualitative properties makes the fuzzy classification approach attractive for TQM. It is possible to fuzzily categorise, analyse and control materials, products, services and processes.
- *Risk Management:* In banking or insurance, individuals or companies have to be divided into risk classes. Very often, pricing components directly depend on risk levels. With a fuzzy classification, the calculation of risk degrees, creditworthiness or other indicators can be carried out with finer granularity.

Fuzzy classification helps to analyze and control qualitative and quantitative performance indicators in managerial application domains. If adequate data from marketing and finance is available in databases, fuzzy classification can be successfully used for performance measurement.

## Acknowledgement

At the IADIS International Conference on e-Society 2006, our research paper (Meier and Werro 2006) has been awarded as an outstanding paper. We thank Pedro Isaias who motivated us to extend the original conference paper and to submit it for a special issue of the Informatica Journal.

We also would like to thank our professional colleagues for developing the fCQL toolkit: Christian Mezger, Christian Nancoz, Christian Savary, Günter Schindler, and Yauheni Veryha.

In the design, development and extension of the webshop eSarine, we have been supported by Daniel Frauchiger, Henrik Stormer and a number of masters students of the University of Fribourg.

In addition, we thank Anthony Clark for helpful comments on an earlier version of this paper.

This research was supported, in part, by the Swiss Federal Office for Professional Education and Technology, under grant No. 8092.2 ESPP-ES.

## References

- [1] Blattberg R. C., Getz G., Thomas J. S., 2001. Customer Equity – Building and Managing Relationships as Valuable Assets. Harvard Business School Press, Boston.
- [2] Edvinsson L., Malone M. S., 1997. Intellectual Capital – Realizing your Company’s True Value by Finding its Hidden Brainpower. Harper Collins Publisher, New York.
- [3] Harrison T., 2000. Financial Services Marketing, Pearson Education, Essex.
- [4] Hruschka H., 1986. Market Definition and Segmentation Using Fuzzy Clustering Methods. International Journal of Research in Marketing, Vol. 3, No. 2, pp. 117-135.
- [5] Hsu T.-H., 2000. An Application of Fuzzy Clustering in Group-Positioning Analysis. Proceedings National Science Council of the Republic of China, Vol. 10, No. 2, pp. 157-167.
- [6] Küng P., Meier A., Wettstein T., 2001. Performance Measurement Systems must be Engineered. Communications of the Association for Information Systems, Vol. 7, Article 3.
- [7] Lee J., Podlaseck M., Schonberg E. and Hoch R., 2001. Visualization and Analysis of Clickstream Data of Online Stores for Understanding Web Merchandising, in Data Mining and Knowledge Discovery, 5, pp. 59-84.
- [8] Meier A., Savary C., Schindler G. and Veryha Y., 2001. Database Schema with Fuzzy Classification and Classification Query Language. Proceedings of the International Congress on Computational Intelligence - Methods and Applications (CIMA), Bangor, UK.
- [9] Meier A., Werro N., 2006. Extending a Webshop with a Fuzzy Classification Model for Online Customers. Proceedings of the IADIS International Conference on e-Society, Dublin, Ireland, Volume I, pp. 305-312.
- [10] Meier A., Werro N., Albrecht M., Sarakinos M., 2005. Using a Fuzzy Classification Query Language for Customer Relationship Management. Proceedings 31<sup>st</sup> International Conference on Very Large Data Bases (VLDB), Trondheim, Norway, pp. 1089-1096.
- [11] Nguyen P.T., Cliquet G., Borges A. and Leray F., 2003. Opposition between Size of the Market and Degree of Homogeneity of the Segments: a Fuzzy Clustering Approach (in French), in Décisions Marketing, 32, pp. 55-69.
- [12] Pine B. J., Davis S., 1999: Mass Customization – The New Frontier in Business Competition. Harvard Business School Press, Boston.
- [13] Rheingold H., 1993. The Virtual Community - Homesteading on the Electronic Frontier, Addison Wesley, New York.
- [14] Schindler G., 1998. Fuzzy Data Analysis through Context-Based Database Queries (in German), Deutscher Universitäts-Verlag, Wiesbaden.
- [15] Sheno S., 1995. Fuzzy Sets, Information Clouding and Database Security. In: Bosc, P., Kacprzyk, J.: Fuzziness in Database Management Systems. Physica Publisher, Heidelberg, pp. 207-228.
- [16] Varki S., Cooil B., Rust R. T., 2000. Modeling Fuzzy Data in Qualitative Marketing Research. Journal of Marketing Research, Vol. XXXVII, November, pp. 480-489.
- [17] Wedel M., Steenkamp H.-B. E. M., 1991. A Clusterwise Regression Method for Simultaneous Fuzzy Market Structuring and Benefit Segmentation. Journal of Marketing Research, Vol. XXVIII, November, pp. 385-396.
- [18] Werro N., Stormer H., Frauchiger D., Meier A., 2004. eSarine – A Struts-Based Webshop for Small and Medium-Sized Enterprises. Lecture Notes in Informatics EMISA2004 – Information Systems in E-Business and E-Government. Luxembourg, pp. 13-24.
- [19] Werro N., Stormer H. and Meier A., 2005. Personalized discount - A fuzzy logic approach. Proceedings of the 5th IFIP International Conference on eBusiness, eCommerce and eGovernment, Poznan, Poland, pp. 375-387.
- [20] Werro N., Stormer H. and Meier A., 2006. A Hierarchical Fuzzy Classification of Online Customers. Proceedings of the IEEE International Conference on e-Business Engineering (ICEBE), Shanghai, China, pp. 256-263.
- [21] Zimmerman H.-J. and Zysno P., 1980. Latent Connectives in Human Decision Making, in FSS, 4, pp. 37-51.
- [22] Zimmermann H.-J., 1992. Fuzzy Set Theory – and Its Applications. Kluwer Academic Publishers, London.

# Mobile Location-Based Gaming as Driver for Location-Based Services (LBS) – Exemplified by Mobile Hunters

Jörg Lonthoff and Erich Ortner  
 Technische Universität Darmstadt  
 Hochschulstraße 1, 64289 Darmstadt, Germany  
 E-mail: {Lonthoff, Ortner}@winf.tu-darmstadt.de

**Keywords:** mobility, location-based services (LBS), gaming, acceptance, application system

**Received:** March 3, 2007

*Location-Based Services (LBS) have entered the discussion in the early 1990ies, but they have not yet achieved a real breakthrough. Here, "Mobile Location-Based Gaming" (MLBG) could bring about the change. Location-based services may well achieve wider recognition because of the human play instinct. This work introduces MLBG and the adventure game "Mobile Hunters" – an implemented MLBG that uses the currently available cellular phone network to create a virtual playing field that represents the real world. This innovative way of playing seems strange at first, but it proves to be a helpful step towards context-based value-added services.*

*Povzetek: Opisana je igrica "Mobilni lovci" na mobilnih telefonih, kot primer novih mobilnih storitev LBS.*

## 1 Introduction

Growing Internet mobility due to various transmission methods such as broadband data transmission get mobile service providers interested in providing services that offer more than voice telephony. Modern cellular phones support GPRS, have a colour display and are usually Java-compliant. This meets the device's requirements for context-based services [18]. As GSM-based cell phones are widely used and – at least in Europe – the GSM-network is available almost everywhere, the context variable "location" seems useful for extending the relevant value-added services. For example, there are services for finding friends in the vicinity (Buddy Alert by Mobiloco – [www.mobiloco.de](http://www.mobiloco.de)) or, mobile navigation systems for cellular phones (NaviGate by T-Mobile – [www.t-mobile.de/navigate](http://www.t-mobile.de/navigate)).

But there has not yet been a real breakthrough for Location-Based Services (LBS). This is partly due to the costs involved. As yet, mobile network providers do not charge competitive prices for location requests [18]. On the other side, the general recognition of such services poses a central problem for LBS [1]. The possible solution presented in this work uses the human play instinct to achieve greater acceptance. Playing, the use of location-based services will become effortless and people will get interested in such systems. Mobile games are becoming a mass market. If this mass market can be served, prices for location requests will go down – thus providing the solution to a second central issue.

Supported by T-Systems International, in 2005 we developed the adventure game "Mobile Hunters" [9]. The game demonstrates what is possible with LBS and uses the currently available infrastructure mobile network

providers offer for creating a virtual playing field. Such a playing field can be adapted to the real world. The object of the game is a hunt. Players can either be a hunter who must find a fugitive or, a fugitive, who has to make sure he is not getting caught. Of course, this hunt will become eventful as there are a variety of obstacles.

This work will first summarize the current state of research in this field and then present "Mobile Location-Based Gaming" (MLBG). Then you will get to know the game "Mobile Hunters". After that, we will discuss the lessons learned from the game and possible further developments will be considered. At the end, a short conclusion will be drawn.

## 2 Related Work

### 2.1 Location-Based Services (LBS)

The added value of mobile services opens up opportunities for service providers to address a new dimension of the user: the user's spatiotemporal position. Such services are called Location-Based Services. LBS are based on a variety of localization methods for determining a user's position. In the field of so-called "Context-Aware Computing" [23] LBS provide location information as context references [3].

There are many possibilities of using the location reference in an application system [20, 25]. Unni and Harmon's classification of LBS is consumer-oriented. They divide LBS in information/directory services (e.g. yellow pages), tracking services (e.g. tracking assets), emergency services (e.g. police and fire response), navigation (e.g. route description) and location-based advertising and promotions (e.g. mobile coupons) [28].

All of these LBS are based on mobile positioning. Mobile positioning comprises all of the technologies for determining the location of mobile devices [28]. Basically, there is a difference between indoor and outdoor localization [2]. In the following sections we will focus on outdoor positioning technologies.

A position can be determined in different ways: using network-based technology (the network provides the position) using terminal-based technology (the device provides the position), and hybrid technology (a combination of the former technologies) [22].

#### Network-based Positioning

GSM networks offer basically six different methods of network-based localization [21]. Cell of Origin (COO) is the simplest mobile positioning technique. It identifies the cell in which a cellular phone is logged on. The positioning accuracy that can be achieved depends on the size of the cellular coverage area, which may range between 25 m and 35 km in diameter. In addition, there are more complex techniques such as Angle of Arrival (AOA), Time of Arrival (TOA), Time Difference of Arrival (TDOA), Signal Attenuation (SA) and the RadioCamera system.

#### Terminal-based Positioning

Cell of Origin (COO) can also be considered a terminal-based technique, as the desired cell ID can be read out directly from the device (terminal). To do so, however, a reference database is needed, which contains the geographic coordinates stored for each cell ID. The following further terminal-based techniques are available: Enhanced Observed Time Difference (E-OTD) as well as the satellite-based systems such as the global positioning system (GPS), or Assisted-GPS (A-GPS), which works without modifications to the cellular phone network infrastructure, except that the mobile device must possess a GPS receiver.

#### Hybrid Positioning

Hybrid localization techniques use network-based and terminal-based technology. The Location Trader developed by the Centre for Digital Technology and Management (CDTM) in Munich combines various localization methods. It works in an accumulative way and uses solely information that is open to the public, like cell ID, signal strength and fingerprints of signal patterns. An iterative localization database is built up by using a multitude of users (community). Explicit user interaction, heuristic computation and tabular search and replenish cause the data basis to adjust. Therefore, the goal is to develop a provider-independent localization method, which becomes more precise as more and more users are involved in this system [4].

## 2.2 Gaming

Mobile games for cellular phones are currently experiencing a growing demand. For the game developers cellular phones are just another platform for porting their console games [18]. There is a strong trend

towards more complicated 3D-games. This trend is supported by the current hardware developments, that is, the availability of high-performance cellular phones or smart phones, respectively.

#### Mobile Games

Games that allow direct communication with remote participants are of great interest (multi-player games). The multi-player games currently on offer can be played using a WAP portal or, locally by two people (infrared) or by several players (Bluetooth).

Games for PDAs (handhelds) are also very interesting. They are usually intended for one player. But games for several players become possible, if infrared, Bluetooth or WLAN are used.

#### Location-Based Games

But the above mentioned “mobile games” ignore the exciting new possibilities mobile devices like cellular phones provide via their inherent ability to maintain connectivity while on the move. So, one possibility could be to extend the virtual world of a game using location-based information. This extension allows users to play games that incorporate knowledge of their physical location and landscape, and additionally provides them with the ability to interact with both real and virtual objects within the space [18].

In location-based games the movements of a player (in the sense of a geographical change of location) influence the game. Nicklas et al. [15] suggest a classification of location-based games into Mobile Games, Location-Aware Games and Spatially-Aware Games.

As a location reference mobile games require merely one more player who is in the vicinity. The location information itself is not considered in the game. A typical example of this kind of game is “Snake”, a game of dexterity for two, which is delivered with the older Nokia cellular phone models. It can be played using infrared or Bluetooth.

Location-Aware Games include information about the location of a player in the game. A typical example might be a treasure quest whereby a player must reach a particular location.

Spatially-Aware Games adapt a real-world environment to the game. This creates a connection between the real world and the virtual world. The MLB “Mobile Hunters” presented in the following belongs to this category of games.

## 3 Mobile Location-Based Gaming (MLBG)

### 3.1 Definition and Delimitation

Mobile Location-Based Games (MLBG) are a special category of Location-Based Games.

**Definition:** A *MLBG* is a location-based game that can run on a mobile device. By using a communication

channel the game can exchange information with a game server or other players.

Applying this definition, the fields Location-Aware Games and Spatially-Aware Games become relevant. The following figure shows all terms relevant in MLBG.

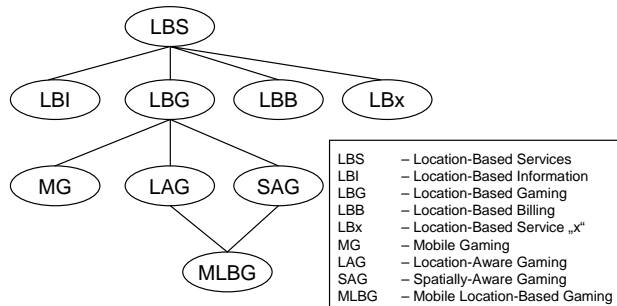


Figure 1. Taxonomy for Location-Based Services.

### 3.2 Characteristics

Important for MLBG are the type of device used, the communication and network infrastructure it is based on, the way positions are determined and the kind of game.

Devices such as cellular phones, smart phones and PDAs can be used, possibly laptops also. In addition to this rough classification, the device properties can serve for further distinction: the operating system, client programming (JavaVM, Web-Client/WAP-Client), the types of user interfaces available, as well as power consumption and processor power.

The relevant communication media are wide area networks such as WLAN, GSM and UMTS. These technologies vary in range and bandwidth.

In chapter 2 you will find a description of the techniques that can be used for determining positions for MLBG. The accuracy of a determined position depends on the technique used and on the network structure.

When looking at the type of game, two dimensions are of interest: the number of players and the type of game. There are single-player games and multi-player games. You can also play multi-player games alone, if players are simulated. Massive-Multiplayer-Games are a special type of game in which the end of the game is not defined. Players can actively participate in the game for some time and improve their ranking in the community associated with the game. Relevant genres of game would be role-playing games, scouting games, real-time strategy games and shoot-em-up games.

You will find a variety of game collections on the Internet that include MLBG. For example:

- [www.smartmobs.com/archive/2004/12/28/locationbased.html](http://www.smartmobs.com/archive/2004/12/28/locationbased.html),
- [www.we-make-money-not-art.com/archives/001653.php](http://www.we-make-money-not-art.com/archives/001653.php) and
- [www.in-duce.net/archives/locationbased\\_mobile\\_phone\\_games.php](http://www.in-duce.net/archives/locationbased_mobile_phone_games.php).

There, the distinction is made between cellular phone games, handheld/PDA games and others. The following are typical examples of MLBG: Botfighters by It's Alive ([www.botfighters.com](http://www.botfighters.com) – cellular phone, GSM,

SMS, J2ME, PC), Treasure Hunt by Treasure Hunt Mobile ([www.playtreasurehunt.com](http://www.playtreasurehunt.com) – cellular phone/PDA, J2ME, GPS), Supafly by It's Alive ([www.itsalive.com/supafly](http://www.itsalive.com/supafly) – cellular phone/PC, WAP/J2ME, SMS/MMS) and Seamful Game by Englands Equator Project (PDA, .net, WLAN, GPS).

In newer publications you will find studies [7] and overviews [11, 12, 13, 18] of pervasive games, including MLBG.

### 3.3 Challenges

One of the main challenges to MLBG is the problem of how to turn "classic" games into MLBG. Nicklas et al. [15] identified four central issues for the adaptation of board games to MLBG:

#### Adaptation of the playing field

The playing fields of board games are clearly delimited. The real world, in contrast, is continuous. This requires an adaptation of the playing field in the sense of a virtual playing field as an extension of the real world. Here, the unclear delimitation of virtual playing fields proves to be difficult.

#### Adaptation of the pawn in a game

The pawns that are being moved in a board game correspond to real-life players who are changing their positions in their environment.

#### The representation of cards or objects

The cards or objects drawn in a board game can be adapted to the virtual world as virtual cards or virtual objects that a player can take.

#### Adaptation of the moves in the game

Board games are based on rounds. One discrete move is made after another. In MLBG this restriction does not apply, as a player cannot simply remain motionless for one round. Playing rounds or sitting out for some time must be considered in the game concept in such a way that the MLBG can be played in real-time.

When a "classic" game is adapted to a MLBG, it is important to ensure that the point of the game is kept and that the duration of the game is adequate. The hype phases we observe are much shorter with games than with any other service offers. Once a game is considered boring or error-prone, acceptance drops.

Rashid et al. remark that in established game developer conferences, and numerous online discussions, many traditional game developers and executives express their doubts about location-based games [18]. Nevertheless, existing location-based games do contribute to operator revenue. For example, in Sweden, Botfighters made EUR 10 to 100 per person in 2002. Furthermore, the fact that the majority of location-based games launched thus far have generated a considerable amount of public interest, it is likely that they will eventually capture the interest of the less traditional gamers [18].



MLBG are fascinating: "This ability for you to actually use your real world movements to play the game means that you are no more playing a game... You are in the game!" [14].

Another challenge results from the characteristics of mobile networks. In MLBG it may happen that some of the players interrupt the connection for short periods of time. These interruptions must be covered (resuming), for example, using central session management.

## 4 "Mobile Hunters"

### 4.1 The Game

We started out with the idea to implement the well-known board game "Scotland Yard", also known as "Hunting Mr. X" [19]. This was difficult because the original game was based on rounds and we needed to change this for the real-time game. Another problem arose when we tried to distinguish the means of transportation. This forced us to modify the conception of the game.

Mobile Hunters is a multi-player game. The creation of a players' community reflects partly the notion of massive multi-player games. To be able to start a game session at least two players must participate. There are two possible roles in the game: One or more hunters and one fugitive. The hunters want to catch the fugitive before the specified playing time (default 30 minutes) is over. The fugitive must try to escape the hunters or, to prove that he is innocent by collecting a number of items (default 3) as proof of his innocence.

After logging on to the Mobile Hunters server with user name and password as authentication, a player can initiate a game. Once a game has been initiated, several more players can enter in the game. If a new player enters in the game, the game server checks whether the potential player's location is within an appropriate distance from the center of the playing field. This ensures



that the spatial distance between the players does not become too large. The person who initiated the game decides when the number of players is sufficient and then starts the game. The maximum number of players can be specified.


When the game is started, the game server randomly assigns the roles and informs the players about whether they are a hunter or a fugitive. When the game has been started, the game server randomly distributes all the items for hunters and fugitives on the playing field. After this initialization phase the game begins synchronously on all participating clients and the countdown for the playing time starts. In previously specified intervals (default 1.5 minutes) the current position of the fugitive appears on a map in the hunters' display. In the playing field there are a number of locked (virtual) boxes that players can open if their position is the same as the position of the box (geo-coordinate with specified radius = playing field). Some of the boxes are visible only for hunters. These boxes contain offensive weapons. The fugitive can only see the boxes that contain items for him. These are items for defense or proof of innocence. Table 1 gives an overview of the various items.

A player can attack if one player is located in the same position and the attacker has a weapon. If a hunter attacks another hunter, the person attacked will become incapable of action for some time (default 30 seconds), his client's menu will be hidden. In this attack the attacker loses his weapon. If a hunter attacks a fugitive, the fugitive can defend himself using a matching item for defense. If this happens, the attacker will become incapable of action for some time (default 30 seconds). If the fugitive cannot defend himself, the hunter wins.

The fugitive wins if he is able to find a certain number (default 3) of items that prove his innocence or, if no attack on him was successful during the duration (default 30 minutes) of the game.

Table 1. Overview of the items available for the players.

Recipient	Type	Instance	Purpose
Hunter	Attack		In the boxes the hunter will find offensive weapons he can use to attack the fugitive or, for attacking another hunter to make him incapable of action (e.g. handcuffs).
Fugitive	Proof		In the boxes the fugitive will find items that can prove his innocence (e.g. a theater ticket).

Fugitive	Defense		In the boxes the fugitive will find items he can use to defend himself against attacks by a hunter (e.g. paper clip for defense against handcuffs)
----------	---------	---	--

### 4.2 Fundamental Design Decisions

In the development of "Mobile Hunters", to facilitate acceptance of the game, we only used technologies that have already been accepted in the market and are, or will most certainly be, widely spread. We chose the GSM network, which is currently most widely used [15, 18], to ensure a wide range of application. Using GPRS, GSM networks support stable IP connections.

The mobile device we chose should not require further hardware (e.g. an external GPS receiver). The devices hardware requirements are at least 200 MHz processor power, 1 MB RAM and a minimal display resolution of 126 x 208 pixels. The software requirements are Java-compliance, a Symbian OS greater or equal version 7.0 in association with Mobile Information Device Profile (MIDP) 2.0 and Connected Limited Device Configuration (CLDC) 1.1. Therefore we opted for the Nokia 6680 cellular phone with an adequate display as playing device [10].

It was essential that the positioning technique would only use the available cellular phone network infrastructure and that it would not be necessary to make modifications to it. Furthermore, it was required that the cell ID could be read out from the cellular phone. This is why Cell of Origin (COO) seemed best suited for Mobile Hunters.

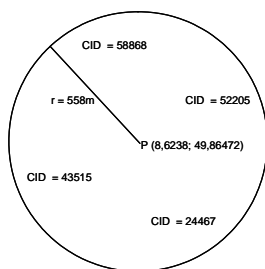


Figure 2. Cellular coverage area of a T-Mobile GSM base station located in Darmstadt, Germany.

COO evaluates the cell ID (CID) to which the cell phone is logged on. The cell ID is connected with the radiation range of a mobile base station. The cellular coverage area has a certain range (r) around the position (P) of the mobile base station. One mobile base station can have several radiation areas (CID), whereby these cells always refer to the same geographic position (P) of the mobile base station's location, mostly denoted in the World Geodetic System 1984 (WGS84) format (s. figure 2) [6].

Thereby the position accuracy depends on the size of the cellular coverage area (approx. 25-100 m in urban areas and up to 35 km in rural areas) [24].

### 4.3 "Mobile Hunters" Architecture

The implementation is based on a client/server architecture. Communication between the components is realized using the German T-Mobile GSM network. For data transfer we chose GPRS, communication takes place on application level via HTTPS. The Mobile Hunters server is a Java-based game server, which provides the user management and the game management. For determining each player's current position, we use T-Systems' research platform "Permission and Privacy Gateway" (PPGW). The PPGW also provides the Mobile Hunters server with maps of the areas in question. The following figure shows the overall architecture schematically.

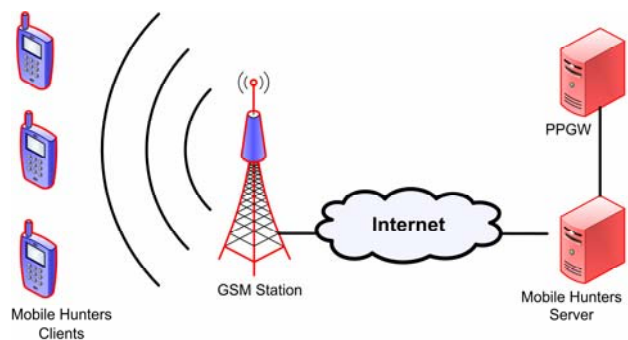


Figure 3. Overall architecture.

### 4.4 "Mobile Hunters" Client

The Mobile Hunters client (s. figure 4) is the game's user interface. It displays the current map segments and here, interaction with the other players takes place. Players also use the Mobile Hunters client to register and log on.



Figure 4. "Mobile Hunters" client on Nokia 6680.

Current prerequisites are a smart phone with Symbian-OS 7.0 and a J2ME framework. A class implemented in Symbian C++ is needed for reading out the cell ID the mobile device uses. The cell ID is transmitted to the Mobile Hunters server. The server requests the current status in specified intervals (default every 10 seconds). The necessary unique identification of each player is the user name he entered at the beginning of the game.

#### 4.5 "Mobile Hunters" Server

The Mobile Hunters server centrally controls the game. On the server, identification and authentication of the users takes place, as well as the game logic and the communication between PPGW and the clients. Every initialization, as well as the specification of the playing field, happens dynamically on the game server. The server was implemented in Java as an Apache Tomcat application on a Microsoft Windows 2003 Server platform and connected to a Microsoft SQL Server via JDBC/ODBC. This ensures a clear separation of application level and data level. The server establishes the connection to the PPGW for requesting the geographic data of each cell ID used in the game at any time. The PPGW has a reference database where the relevant geo-reference-data for each cell ID are stored. The geographic data that correspond to the cell IDs are returned by the PPGW via an XML-based interface. The server keeps a high-score list to challenge the gaming community. Using a configuration file, the parameters that can be controlled are passed to the server.

## 5 Lessons Learned

### 5.1 The Game's Concept

The adaptation of the game's concept to the virtual world turned out to be difficult because of the inaccuracy of positioning. Cell switches occur even if a player does not move. This may happen, for example, if several participants are logged on to a cell and, due to the limits of bandwidth or for reasons of optimization, a neighbouring cell is allocated to the mobile device.

Cellular coverage areas have different sizes: in an urban area approx. 25 m, in rural areas up to 35 km. The homogeneous structuring of a cellular phone network in hexagonal units that is described in literature is not found in practice. The cellular coverage areas of single cells overlap with those of other cells. Large cells cover smaller cells almost completely (s. figure 5). It is still in the discussion whether to exclude cells from the game if they are too large and, from which size on they should be banned.

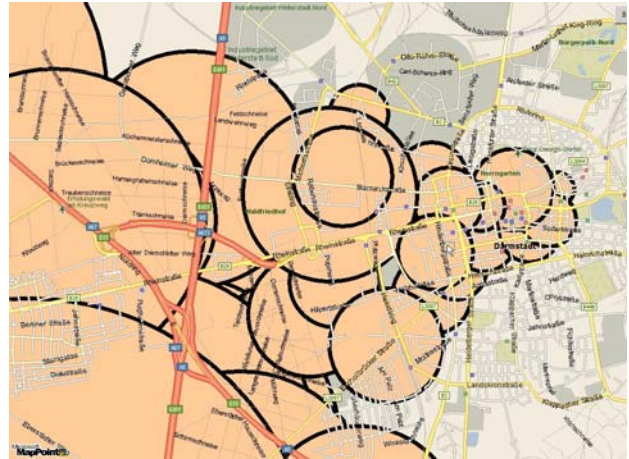


Figure 5. T-Mobile cells in Darmstadt, Germany.

In the real world it frequently happens that two players are standing side by side, but their mobile devices are logged on to different cells. "Mobile Hunters" is suited for playing in high-density urban areas, because the accuracy of positioning is adequate there. On the outskirts of a town or in rural areas, the game cannot be played due to the current means of positioning and the size of the cells.

For the game Mobile Hunters to be funny and exciting, a game duration of 30 minutes is recommended. The interval in which a fugitive's current position is displayed should be 1.5 minutes, the time a player is incapable of action should be 30 seconds and, the interval in which the cell ID is read ought to be less than one second.

### 5.2 Hardware and Implementation Experience

A powerful CPU such as the Nokia 6680 is necessary for playing this game. Session handling is essential for playing a MLBG. The user interface's design should be simple, so that no explanations will be necessary.

It is very important to develop the client's code in a way best suited for cellular phones. This means, the "renaissance" of well-tried implementation guidelines like lightweight data structures and prevention of unnecessary procedure calls [10].

Our mobile-specific modeling resulted in a very economic consumption of resources and of handling communication data. In a 30 minute game, only an average of 300 kB data is transferred. Half of that data volume is used for loading the graphics (maps and icons) at the beginning of the game.

In our attempt at optimizing the implementation, another important aspect needed to be considered. It is often the case that, at run-time, 90% of the executed program uses only 10% of the implemented code. It is therefore vital that the code optimization take care of precisely those 10% of the code.

JSME Wireless Toolkit offers a tool called "Profiler". It is ideally suited for finding out, which code is frequently used and can therefore be very helpful for code optimization. You will find a detailed description of

how to use this tool in [26]. The Profiler gives an overview of the entire code execution, as for example the numbers of iterations of a loop or a procedure call.

## 6 Enhancements

To achieve a higher positioning accuracy based on COO, we are trying to handle cell switches and overlaps by including all cell IDs that the mobile phone receives. This allows us to buffer positions and information can also be evaluated by neighbouring cells. This software-based solution is also known as Enhanced cell ID (E-CID) [28].

A more comprehensive solution would be to use more accurate positioning techniques like GPS, or in future, GALILEO. Assisted GPS (A-GPS) will open up interesting opportunities for mobile network providers.

To play Mobile Hunters globally, that is over great distances an additional virtualization level should be introduced. Using this additional level, each individual player's physical movements could be mapped as relative movements on the virtual playing field. This approach would eliminate the range restrictions of the playing field [27].

It would be a nice feature, if hunters were able to communicate during the game in order to act strategically. Communication between the hunters can become possible by using Instant Messaging or simply by sending text messages (SMS). In this case, another alternative would be Push-to-Talk or voice-telephony. Such communication is very interesting for network operators who want to increase their revenue.

The Mobile Hunters server could be enhanced to become an enabling service, which would provide potential creators of MLBG with the generic functions needed for the implementation of location-based games.

Furthermore, Mobile Location-Based Gaming could be extended by a variety of context parameters and thus reach the higher development stage of Mobile Context-Based Gaming. In addition, MLBG is an interesting medium for advertising and interactive marketing [5, 17].

## 7 Conclusion

The conclusion we draw is that we have partly succeeded in adapting the real world to a game's virtual world. It is possible to read out the cell ID from a variety of mobile devices using different techniques. However, no standard API is currently available to offer this functionality. This means that anyone who wants to create a game needs to develop a suitable API for every single device, if the game is to be widely used. Only Java-compliant devices with optional package JSR 179 (Location API for J2ME, released 12.07.05) are suitable for location requests. This problem may also be solved by deploying a middleware such as BREW (www.qualcomm.com/brew; [27]) or SNAP Mobile (snapmobile.nokia.com; [8]).

The experience gained in the field of gaming also applies to situations in private and business life. Possibly, brand-new application systems useful in every day life

can be developed and implemented. In this context "application system" is understood in a comprehensive way for all tasks in user and computer-based information processing [16].

The infrastructure for LBS is available. Now, key issues such as the ownership and management of location-specific data, transaction and data security, as well as consumer privacy are still to be resolved [28]. Gaming is suitable to get a feeling of LBS.

## Acknowledgement

Mobile Hunters was developed in a co-operation between the university department "Wirtschaftsinformatik I – Development of Application Systems", computer and information science, in Darmstadt and T-Systems International GmbH. Eight students designed and implemented the game as hands-on training in the field of information science in the winter term 2004/05. Based on this prototype, the "Mobile Hunters" team continued the development and stabilized the software. Here, my special thanks go to the "Mobile Hunters" team (Florian Seeanner, Tilman Laschinger and Michael Wolf) for their extraordinary dedication and to Thomas Leiber of T-Systems International for his invaluable visionary and practical support.

## References

- [1] Borriello, G et al., 2005, The disappearing computer: Delivering real-world ubiquitous location systems. In: *Communications of the ACM*, Vol. 48, No. 3, pp. 36-41.
- [2] Bulusu, N. et al., 2004, Self-Configuring Localization Systems: Design and Experimental Evaluation. In: *ACM Transactions on Embedding Computing Systems*, Vol. 3, No. 1, 2004, pp. 24-60.
- [3] Dey, A., 2001, Understanding and Using Context. In: *Personal and Ubiquitous Computing*, Vol. 5, No. 1, pp. 4-7.
- [4] Dornbusch, P.; Huber, M., 2003, Generierung von Ortsinformationen durch User-Communities. In: Uhr, W.; Esswein, W.; Schoop, E. (Eds): *Proceedings of Wirtschaftsinformatik 2003 – Medien, Märkte, Mobilität*, Vol. 1, Heidelberg [a. o.], 2003, pp. 175-186.
- [5] Han, S.; Cho, M.; Choi, M., 2005, Ubitem: A Framework for Interactive Marketing in Location-Based Gaming Environment. In: *Proceedings of the IEEE International Conference on Mobile Business (ICMB'05)*. Sydney, Australia, pp. 103-108.
- [6] Hansmann, U. et al., 2001. *Pervasive Computing Handbook*, Springer, Berlin, Germany.
- [7] Jegers, K.; Wiberg, M., 2006, Pervasive Gaming in the Everyday World. In: *Pervasive Computing*, Vol. 5, No. 1, pp. 78-85.
- [8] Levy, M., 2006, Developing Network Connected Mobile Games for J2ME Handsets Using SNAP Mobile. In: *Proceedings of Game Developers Conference (GDC) 2006*, San Jose, USA.

- [9] Lonthoff, J.; Leiber, T., 2005, Mobile Location Based Gaming als Wegbereiter für Location Based Services. In: *LNI-Proceedings of the 19. DFN-Arbeitstagung über Kommunikationsnetze*, Volume P-73. Düsseldorf, Germany, pp. 343-360.
- [10] Lonthoff, J.; Ortner, E.; Wolf, M., 2006, Implementierungsbericht Mobile Hunters. In: Roth, J.; Schiller, J.; Voisard, A. (Eds): *3<sup>rd</sup> GI/ITG KuVS Fachgespräch Ortsbezogene Anwendungen und Dienste*, Technical Report, Institut für Informatik, Freie Universität Berlin, 07.-08.09.06, Berlin, pp. 26-31.
- [11] Magerkurth, C. et al. (eds), 2004, *Proceedings of International Workshop on Gaming Applications in Pervasive Computing Environments (PerGames) 2005*. Vienna, Austria.
- [12] Magerkurth, C. et al. (eds), 2005, *Proceedings of 2<sup>nd</sup> International Workshop on Gaming Applications in Pervasive Computing Environments (PerGames) 2005*. München, Germany.
- [13] Magerkurth, C. et al. (eds), 2006, *Proceedings of 3<sup>rd</sup> International Workshop on Gaming Applications in Pervasive Computing Environments (PerGames) 2006*. Dublin, Ireland.
- [14] Mikoishi, 2004, *gunslingers*. In <http://www.gunslingers.mikoishi.com>.
- [15] Nicklas, D. et al., 2001, Towards Location-based Games. In: *Proceedings of the International Conference on Applications and Development of Computer Games in the 21st Century: ADCOG 21*. Hongkong Special Administrative Region, China, pp. 61-67.
- [16] Ortner, E., 2005. *Sprachbasierte Informatik – Wie man mit Wörtern die Cyber-Welt bewegt*. EAGLE-Verlag, Leipzig, Germany.
- [17] Rashid, O.; Coulton, P.; Edwards, R., 2005, Implementing Location Based Information/Advertising for Existing Mobile Phone Users in Indoor/Urban Environments. In: *IEEE-Proceedings of the 4<sup>th</sup> International Conference on Mobile Business (ICMB'05)*, 2005, Sydney, Australia, pp. 377-383.
- [18] Rashid, O. et al., 2006, Extending Cyberspace: Location Based Games Using Cellular Phones. In: *ACM Computers in Entertainment*, Vol. 4, No. 1, 2006, pp. 1-16.
- [19] Ravensburger, 1983. *Scotland Yard*. Ravensburg, Germany.
- [20] Richter, U. et al., 2004, Location-based Services: Konkurrenz durch lizenzfreie Alternativen. In: *VDE Kongress 2004*, Berlin, Germany, pp. 65-70.
- [21] Röttger-Gerigk, S., 2002. Lokalisierungsmethoden. In: *Handbuch Mobile-Commerce*. Springer, Berlin, Germany, pp.419-426.
- [22] Samsioe, J.; Samsioe, A., 2002, Competitor Analysis in the Location Based Service Industry. In: *IEEE-Proceedings of First International Conference on Mobile Business (ICMB'02)*, 2002, Athens, Greece.
- [23] Schilit, B. et al., 1994, Context-Aware Computing Applications. In: *Workshop on Mobile Computing Systems and Applications*. Santa Cruz, USA, pp. 85-90.
- [24] Schiller, J., 2003, *Mobilkommunikation*, 2<sup>nd</sup> ed., Pearson Studium, Munich, Germany, 2003.
- [25] Schiller, J.; Voisard, A., 2004. *Location-based services*. Morgan Kaufmann Publishers, San Francisco, USA.
- [26] Shivas, M., 2003, *J2ME Game Optimization Secrets*. In: [http://www.developer.com/ws/j2me/article.php/10945\\_2234631\\_1](http://www.developer.com/ws/j2me/article.php/10945_2234631_1).
- [27] Tarumi, H.; Matsubara, K.; Yano, M., 2004, Implementations and Evaluations of Location-Based Virtual City System for Mobile Phones. In: *Proceedings of the IEEE Global Telecommunications Conference Workshops*. Dallas, USA, pp. 544-547.
- [28] Unni, R.; Harmon, R., 2003, Location-Based Services: Models for Strategy Development in M-Commerce. In: *IEEE-Proceedings of Portland International Conference on Management of Engineering and Technology 2003 (PICMET'03)*, Portland, USA, pp. 416-424.



# A Group Learning Management Method for Intelligent Tutoring Systems

Eliane Pozzebon<sup>1,2</sup>, Janette Cardoso<sup>2</sup>, Guilherme Bittencourt<sup>1</sup> and Chihab Hanachi<sup>2</sup>

<sup>1</sup> Santa Catarina Federal University, DAS  
88040-900 Florianópolis, Brazil,  
E-mail: (eliane.gb@das.ufsc.br)

<sup>2</sup> Université Toulouse 1, IRIT  
F-31042 Toulouse Cedex, France  
E-mail: (jcardoso, hanachi@univ-tlse1.fr)

**Keywords:** learning in group, collaborative learning, intelligent tutoring systems, multi-agents systems

**Received:** February 17, 2007

*In this paper we propose a group management specification and execution method that seeks a compromise between simple course design and complex adaptive group interaction. This is achieved through an authoring method that proposes predefined scenarios to the author. These scenarios already include complex learning interaction protocols in which student and group models use and update are automatically included. The method adopts ontologies to represent domain and student models, and object Petri nets to specify the group interaction protocols. During execution, the method is supported by a multi-agent architecture.*

*Povzetek: Grupno učenje je podprto s scenariji, modeli, ontologijami, agenti, Petri mrežami.*

## 1 Introduction

Although the research on *Artificial Intelligence in Education (AI-ED)* can be traced back to the 80's, when the first ideas on *Intelligent Tutoring Systems (ITS)* were introduced, presently it is going through an accelerated evolution process, mainly due to innovative computer technologies, such as hypermedia, Internet and virtual reality [1] [2]. Nevertheless, the conceptual gaps between authoring systems and authors and between instructional planning and tutoring strategy for dynamic adaptation are challenges that have not yet been overcome [23]. These challenges are especially complex in Intelligent Tutoring Systems in which one considers, beside an individual interaction, a group interaction. In this case, the ITS should not only support the domain presentation for a single student, but also manage the group interactions.

ITSs that allow group work present different degrees of group interaction control. At one extreme, we have systems that only make available the communication tools that allow the group interaction (chat, mail, forum, cooperative editors, etc), leaving all the problem solving and coordination activities under human responsibility. At the other extreme, we have systems that control all the details of the group interaction, following well defined and rigid protocols. In the former, the author instructional planning task is at least as hard as in traditional group work planning. In the latter, the lack of flexibility makes it difficult to

achieve dynamic adaptation and to share and reuse ITS components across domains.

In this paper we propose a group management specification and execution method that seeks a compromise between these two extremes. To provide a reliable and flexible interaction mechanism, the method includes a formal specification language that allows the definition of arbitrarily complex learning interaction protocols, here called *scenarios*. These scenarios are specified by the authoring tool developers. The group activity author only provides the contents and customizes the chosen scenario using an authoring interface. To provide an adaptive behavior the method explores the structure of the domain and student models of the underlying ITS. This is possible because the method is intended to be applied to ITSs created using the FAST multi-agent ITS building tool [3] [15], in which these models are specially designed to facilitate adaptiveness.

To tackle the compromise between simple group activity design and complex adaptive group interaction, the following project decisions were adopted in the development: (i) an explicit representation, using ontologies, of the knowledge that describes the domain, student and group activity models, including their relationship, (ii) the use of a multi-level control process to increase the flexibility of the behavior without sacrificing the specification simplicity, (iii) the use of an expressive formalism, *Object Petri Nets (OPN)* [25], to specify the group



interaction protocols. OPN are a formalism combining coherently *Petri nets (PN)* theory and the *Object-Oriented (OO)* approach. While PN are very suitable to express the dynamic and possibly concurrent and open behavior of a protocol, the OO approach permits the modeling and the structuring of its active (actor) and passive (information) entities. In our case, actors correspond to teacher and students, while information corresponds to domain and scenarios.

The paper is organized as follows: Section 2 presents an overview on the related work. Section 3 introduces the FAST ITS building tool. Section 4 explains how a group interaction scenario is specified and Section 5 presents a simple example of a scenario, discussing in particular what a group interaction is. Finally, in Section 6, we present some conclusions and future works

## 2 Related Work

Organization modelling is recognized as an essential mechanism for structuring the design of *Multi-Agent Systems (MASs)* and coordinating their executions. Indeed, this approach provides high level concepts, such as groups, roles, protocols or commitments, useful to structure and rule, at a macro level, the coordination of the different agents involved in a MAS. All these reasons have led to an increased development of agent methodologies (GAIA, MOISE, AALADIN, etc.) structured around organizational concepts (see [20] for a survey). In most of these methodologies, protocols and groups are considered as basic building blocks of an organizational oriented approach of MASs. This is the approach we have followed in this paper by structuring our MAS around groups and protocols: while groups constitute an interaction space for agents, protocols define the rules to enter or leave a group and play a role within a group. The concept of organization (also groups, institutions, communities, etc.) within MAS has been discussed in several papers [10], [13], [9], [11], [12], [24], [19], [26], [16], [14], [28].

Regarding agent-based protocols, [7] provides an interesting survey of the different specification formalisms, and concludes that Petri Nets provide good software engineering properties to specify, validate and execute concurrent protocols. Our work is also related to [16] in which the adequacy of the Petri Net with Objects formalism, to describe real world protocols, is shown. Systems focusing on the concept of group have also been used in the context of ITS [22], [14], [27], [21] and [18]. In this paper, we do not address the automatic group formation problem. This issue is treated, for example, in NetClass [21] using the learner model, the author information and a sociometric test (that measures the degree of cohesion among students). In WhiteRabbit [27], the groups are created from the analysis of the user model based on the keywords (about his projects, experience, etc.) and also on the conversations.

Among ITSs that share our goal of simplifying course development, an interesting example is the Cognitive Tutor Authoring Tools (CTAT) project. It assists in the creation and delivery of ITS based on model tracing [17]. The main goal of this project is to provide tools to reduce the amount of artificial intelligence (AI) programming expertise required to implement ITSs. The project authoring tools support the development of two types of tutors: Cognitive Tutors and Example-Tracing tutors. Cognitive tutors contain a cognitive model that simulates the student thinking in order to monitor student activities and to provide pedagogical assistance during problem solving. In contrast, Example-Tracing Tutors do not contain a cognitive model: to develop a tutor of this kind, the author needs to specify a recording of possible student actions and corresponding feedback messages. Although Example-Tracing Tutors do not require IA programming, they are specific to the given set of problems and cannot deal with student actions which are not pre-specified by the author [17], i.e. they lack adaptiveness.

An example of an ITS that uses multi-agent technology is the DOCTA [5] system. It uses intelligent agents for collaborative learning to support collaboration in a learning scenario on gene technology. Agent system consists of two components: a Student Assistant agent (SA-agent) and an Instructional Assistant agent (IA-agent). Both agents observe and detect problems in the collaboration and knowledge-building process among students, but their presentations are different.

Another example is the COLER system [6] that addresses both social and task-oriented aspects of group learning. It helps students collaborate while solving Entity Relationship modeling problems. Unlike previous work, generally emphasizing dialogue analysis or expert models, this work proposes a new approach to support collaboration that identifies learning opportunities based on the differences between problem solutions and tracking levels of participation. This work demonstrates how intelligent agents can produce reasonable collaboration advice in domains for which structured problem solutions exist by using a few basic knowledge sources, and illustrates several methods for knowledge evaluation and reasoning of complex knowledge-based systems.

## 3 FAST ITS Building Tool

FAST [3] is a domain independent authoring tool to implement multi-agent Intelligence Tutoring Systems. Courses developed using FAST are based on the conceptual model MATHEMA [8]. This model proposes an ITS architecture that consists of three modules (see Figure 1): the *Tutoring Agent Society (TAS)*, the *Student Interface* and the *Instructor Interface*. The student interface provides access to the system and the instructor interface allows the monitoring of the course. The TAS consists of a multi-agent system where each *Tutor Agent (TA)* contains a

complete ITS focused on a sub-domain of the course target domain. Each of the intelligent tutoring agents in the TAS is responsible for one sub-domain. MATHEMA provides a modeling scheme for these sub-domains that is divided into two views: *external view* and *internal view*.

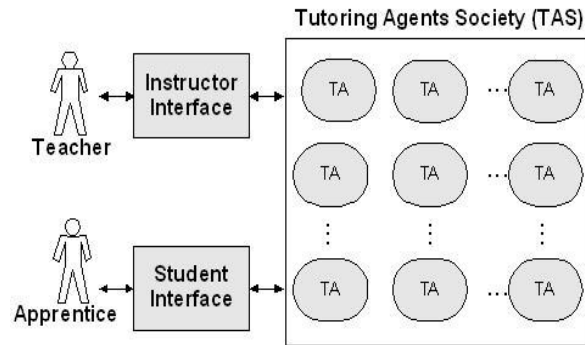


Figure 1: MATHEMA System Architecture.

The external view is a domain knowledge partitioning scheme, based on epistemological assumptions, that guides the author during course development. This partitioning is performed according to two main dimensions: *context* and *depth*. Along the context dimension the domain knowledge is partitioned according to a set of different points of views about its contents. For each particular context, the depth dimension partitionates the domain knowledge according to the methodologies used to deal with its contents. Each pair context/depth is associated with a sub-domain, to be dealt with by one of the TAS agents.

The internal view proposes to organize the knowledge associated with each sub-domain into a set of curricula. Each curriculum is progressively refined according to three levels of detail: *pedagogical units*, *problems* and *interaction support units*. At the *pedagogical unit* level, each curriculum, that describes a possible sequence of sub-domain contents to be presented to the student, is refined into a set of partially ordered pedagogical units, possibly with prerequisites relationships. At the *problems* level, each pedagogical unit is refined into a set of problems, also partially ordered and possibly with prerequisites relationships. Finally, at the *interaction support units* level each problem is associated with a set of interaction units with the student, that support the problem solving activities, such as explanations, examples and exercises.

The domain knowledge of any ITS developed using the FAST tool presents the structure defined by this internal view. This fact allows the construction of group interactions that, although not domain dependent, can explore the domain structure, going beyond the simple communication support between group members and the instructor. This is possible because these group interactions can use the same problem solving activities already defined in the context of the underlying ITS.

A further advantage is that the student model, used in group interaction management, can be defined as an extension of the student model in the underlying ITS. In such a way that the group interaction manager can explore the preferences and previous results obtained by each student in the context of individual learning during her/his interaction with the underlying ITS. Both, domain and student models, are represented using ontologies. These ontologies are briefly described in the next subsections.

### 3.1 Domain Model

The domain model contains definitions of all the concepts in the internal view of the MATHEMA model. A course is represented as an instance of the domain model and contains all the information provided by the author. This information is of two types: properties and contents. Examples of properties are prerequisite relationships, degree of detail, level of difficulty, etc. Contents is what is presented to the student, typically an interactive page encoded into predefined HTML pages templates.

The ontology described in [15] includes concepts to define prerequisite order graphs, that can be used to define the relationship among pedagogical units or problems and concepts to represent specific types of interaction support units, whose contents are also specified by the author (see Figure 2). These concepts correspond to the elements of the internal view of the MATHEMA model.

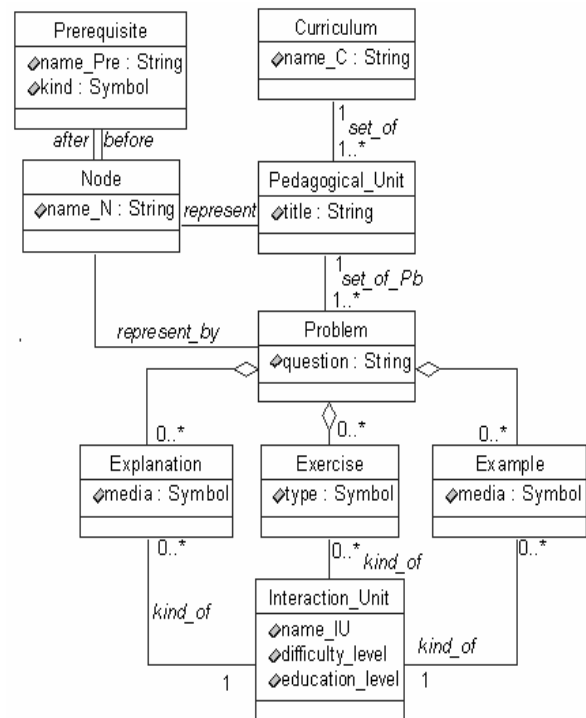


Figure 2: Domain Model.

In particular, the Problem concept (and its sub-concepts) is reused in the definition of the Content Unit concept of the proposed group management method (see Section 4.2).

### 3.2 Student Model

The student model, proposed in [15], contains definitions of all the concepts necessary to characterize a student and her/his history of interactions with the system. Its contents include static information, such as education level, based on a preliminary test, and preferences; and also dynamic information that consists of descriptions of the student activities during all her/his sessions of interaction with the system. This student model was extended to include the information necessary to group interaction. This also includes static information, such as preferences, and dynamic information, such as the record of the student performance during group activities.

## 4 Group Management Method

The goal of the proposed group management method is to allow the specification and execution of complex group activities, without burdening the author with the task of specifying how the student and group activity models should be taken into account and updated during the interaction. To support the proposed method, the conceptual model proposed by the FAST tool was extended to include the definition of *group activity*. A group activity involves the *developer* who specifies the scenario library where the group activities are stored; the *author* who chooses and instantiates a suitable scenario to build an actual group activity; and the *instructor* who supervises the group activity, determining the beginning and end of the activity, and verifying student feedback. Each group activity is necessarily based on an underlying ITS built up using the FAST tool. It presents two

levels: the *specification level* and the *execution level*, as shown in Figure 3.

The specification level main concepts are Group and Scenario. A group consists of a set of students. A scenario consists of an operational definition of the group activity. Scenarios are defined by the developer and stored into a scenario library. They are built using predefined activity units that can be reused in different scenarios.

The execution level consists of a multi-agent system that performs a group activity based on an instance of the Scenario concept, as defined in the specification level. To define such an instance, the author chooses the more adequate scenario from the scenario library, provides the contents, and customizes the scenario parameters (e.g., student level requirements, minimal and maximal number of group members, etc). This information is compiled into an OPN able to manage the group activity, in which the tokens are instances of the Group concept. Once the scenario and group instances are defined, the group activity can be made available to the students to be executed under the supervision of the instructor.

The concepts involved in these two levels of the group activity are described in more detail in the next subsections.

### 4.1 Specification Level

The concepts involved in the specification level (see left side of Figure 3) include: Group, Role, Scenario, Prerequisites, Activity units (Management and Content units), Prerequisites and Interaction Protocol.

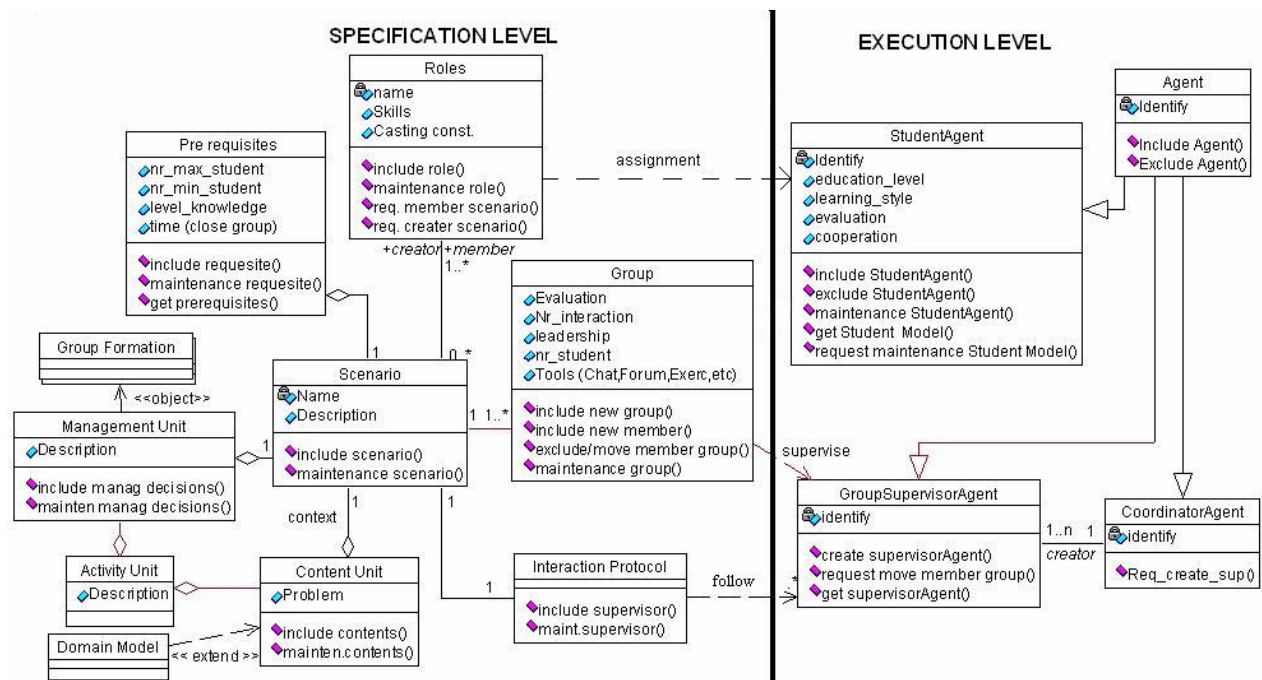


Figure 3: Group Activity Model.

The Group concept joins a set of students already inscribed in the underlying ITS and, optionally an instructor. The members of a group can be assigned to different roles.

The Role concept structures the members of a group into classes according to their participation in a scenario. Each Role is defined by: its name; the required skills (that an agent must meet to be authorized to play that Role); and the casting constraints (such as the maximum number of agents that may play that role, the condition required to play it, etc). Some examples of roles are: Team Leader, and Plain Member.

The Scenario concept consists of an operational definition of the group activity. It includes prerequisites, activity units and an interaction protocol.

The Prerequisite concept defines the initial conditions for a given scenario, e.g., the minimal and maximal number of group members, the situation of these members with respect to the course of the underlying ITS, etc.

The Activity Unit concept defines the different activities that occur in a given scenario. There are two types of activity units:

- Management units: used to define the typical activities of group interactions, e.g., group formation, problem distribution, wait for the first solution, waiting for all solutions, group member instruction, all group members instruction, etc; and
- Contents units: used to define the problem solving tasks associated with a given scenario. These tasks are defined using the Problem specification (that includes Interaction Units) of the domain model of the underlying ITS (see Section 3.1). The content definitions are provided by the author, using the authoring interface of the FAST tool [4] [15] and can be used either in the context of group learning or individual learning.

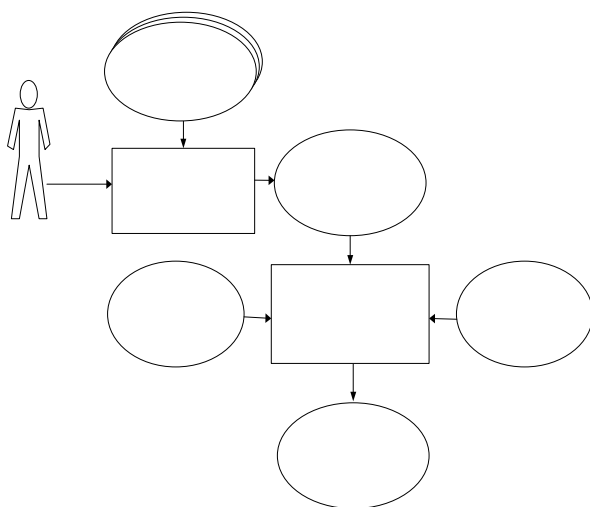


Figure 4: Specification of an Interaction Protocol.

The Interaction Protocol concept contains the operational specification of the group activity, i.e., the

order in which the activity units are executed in a given scenario. It is represented by a two level hierarchical OPN. The specification of an instance of an interaction protocol is a two step process (see Figure 4). The first step consists in selecting a scenario from the scenario library and instantiating all its attributes.

The second step compiles this information to produce the OPN that defines the interaction protocol. The compilation process automatically integrates the domain model of the underlying ITS and the use and update of the student models into the conditions of the Petri Net transitions. This integration provides the adaptive character of the interaction protocol.

### 4.2 Execution Level

The execution level is defined by a multi-agent architecture inspired by Ferber [07] and represented in Figure 5. The concepts involved in the execution level (see right side of

Figure 3) are: Agent, Student Agent, Group Supervisor Agent and Coordinator Agent.

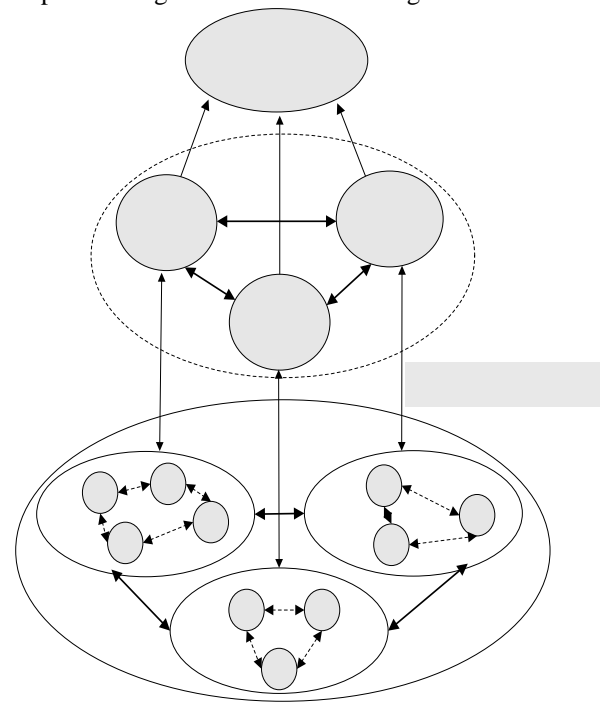


Figure 5: Multi-Agent Architecture.

A Student Agent (SA) represents a student and has a role assigned to it. Each Student Agent stores internally the information of the student model relevant to the group management, e.g., the group activities in which the student has participated, statistics about the role of the student in these groups (leader or not), the number of group communications in which she was involved, etc. It can also consult the student model stored in the TAS agents of the underlying ITS (see Section 3.2).

A Group Supervisor Agent (GSA) supervises a group activity according to the OPN associated with

the interaction protocol defined in the scenario instance. In this OPN, the tokens are instances of the Group concept. This allows the management units, in the interaction protocol, to consult and update group attributes. In a first step, students can be included as group members. Once the student set is available, it can be used to identify the associated student agents, e.g., to allow a broadcast message to be sent or to consult the student models stored in the TAS of the TAS, e.g., to check the performance of the students in a given content unit.

The Coordinator Agent is responsible for the creation and destruction of Group Supervisor Agents at run time, for the permanent storage of all the relevant information about the group activities, and also for monitoring individual learning in order to detect opportunities for group learning activities. The Coordinator Agent also provides an interface for the instructor, through which she/he can monitor the group activity.

## 5 An Example of a Group Activity

To clarify the notions introduced in the previous section, we present an example of a simple group activity and show how it can be instantiated into a concrete group management process.

The group activity is intended to develop the “divide-and-conquer” strategy in problem solving. It supposes a problem that can be partitioned into a certain number of sub-problems. Each sub-problem may be solved independently and their solutions have to be combined to solve the original problem.

### 5.1 General Description

According to the specification level of a group activity, defined in Section 4.1, we must define the following concepts: group, roles, scenario, prerequisites, management units, content units and interaction protocol.

**Group:** the activity needs at least one student per sub-problem.

**Roles:** the activity includes two roles: sub-problem solver and solution integrator. The solution integrator role should be assigned to one or more students that will be responsible for the integration of the sub-problems solutions. The choice of these students can be done dynamically, e.g., the first to complete a sub-problem solution or the best graded in the underlying ITS. Finally, the sub-problem solver role is assigned to all the students that participate in the activity.

**Scenario:** it is defined by the following concepts.

**Prerequisites:** the members of the groups involved in the activity should have the necessary background to solve the problem being considered.

**Management units:** The following management units are necessary to control the scenario:

- Group formation: assignment of the problem solver roles associated with each sub-problem.
- Sub-problem distribution.
- Monitoring of the sub-problem solutions.
- Coordination of the interaction among group members who incorrectly implemented the interface between their solutions.
- Assignment of the integration group.

**Content units:** The contents of the scenario, to be provided by the author through the FAST authoring interface (see Section 3), consist of the following problem descriptions:

- A general explanation of the problem and its sub-problems.
- For each sub-problem:
  - a detailed explanation.
  - one or more examples of similar problem solutions.
  - two types of exercises: one that tests the correctness of the sub-problem solution and one that verifies whether the solution correctly implements the expected interface with the other sub-problem solutions.
- An exercise that tests the correctness of the combined sub-problem solutions.

It should be noted that these contents are instances of the problem (and interaction unit) concepts of the domain model ontology described in Section 3.1 and may also be used in the context of an individual interaction with the underlying ITS.

**Interaction protocol:** The top level of the interaction protocol corresponding to this scenario is represented by the Petri Net shown in Figure 6.

In our context tokens contain an instance of the group concept. For legibility reasons, we have omitted the object dimension of a OPN that is the values of tokens, the Preconditions, Actions and Emission Rules of transitions. The resulting *abstract OPN* just keeps the aspects related the behavioral structure of the protocol. However, from this conventional abstract structure, standard Petri net properties can be proved, such as the presence of *loops* or *cycle* (sequences of transitions that can be infinitely repeated, *deadlocks* (blocking state from which no transition may occur), the (un-)accessibility of a *goal*, *final* or a *home state*, the *boundness* (infinite growing of the number of tokens) or the *lost of tokens* in a hole place.

### 5.2 Scenario Instance

To instantiate the group activity for the “divide-and-conquer” problem solving strategy, we implement a group activity based on an already existing individual learning ITS for the domain of *Structure of Information*, an undergraduate discipline of the Control and Automation Engineering course at the Santa Catarina Federal University, Brazil [3]. This ITS

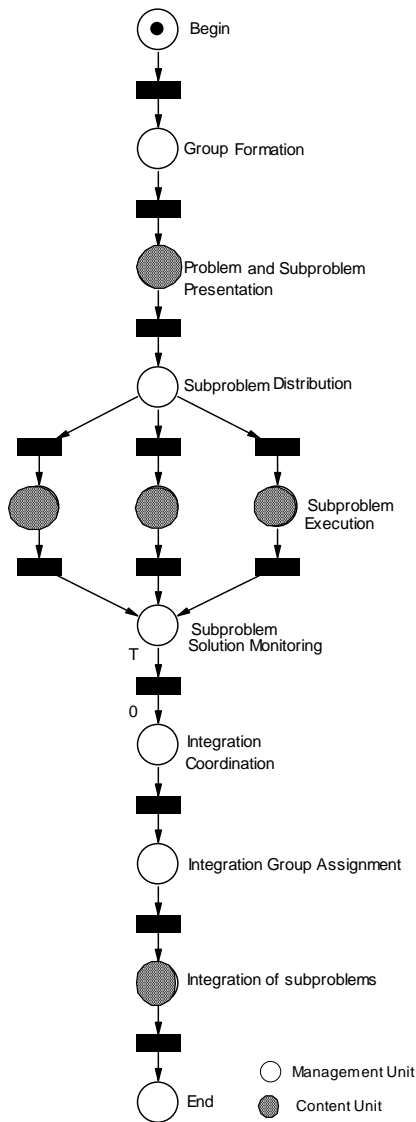


Figure 6: Petri net modelling the interaction protocol of a scenario.

was built using the FAST tool and meets the definitions given in Section 3.

The problem to be solved during the group activity is defined as follows:

- Problem description: given a programming language that supports integer arithmetic operations, how can it be extended to support operations for other types of numbers (rational, float and complex).
- Sub-problems: arithmetic operation packages for each of the three new types of numbers, including appropriate conversion functions.
- Integration: a dispatch function package that integrates all four types of numbers.

The implemented group activity is intended to be developed during a presential course in which the course teacher is the instructor. The instances of the relevant concepts involved in the group activity definition of Section 5.1 are defined as follows.

**Roles:** The sub-problem solver role is assigned to all students in the class room and the solution integrator role is assigned to the students that are members of the first group to solve the assigned sub-problem successfully.

**Prerequisites:** the students that participate in the activity must have completed the necessary pedagogical units (basic programming, abstract data types).

**Management units:** the implemented activity uses a synchronous group formation in which an invitation is sent to all the students in the classroom. The students should answer with the identification of their preferred sub-problem. The system controls the maximum size of each group automatically. The necessary prerequisites are also verified for each student. The Group Formation place in the top level OPN (see Figure 6) is exploded into the bottom level OPN shown in Figure 7. Problem distribution is generated automatically. Monitoring of sub-problem solutions is

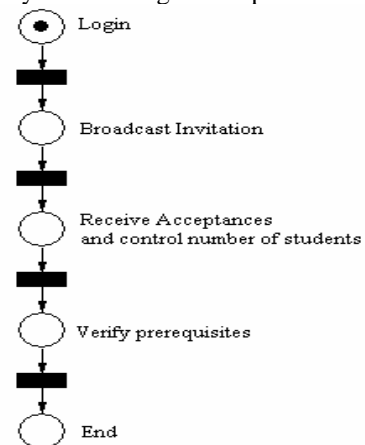


Figure 7: Petri Net Modelling the Group Formation Protocol.

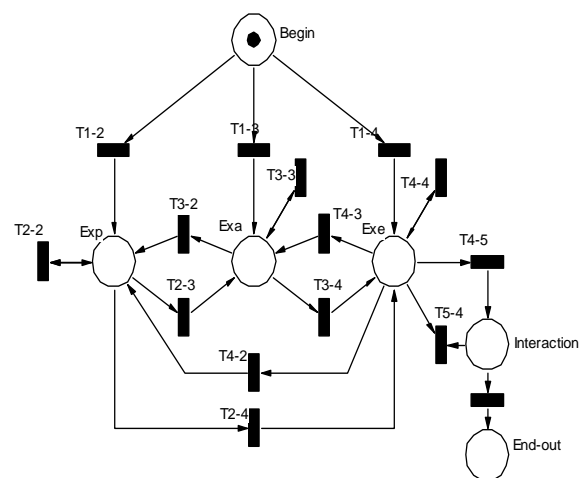


Figure 8: Petri Net Modelling the Problem Solving Protocol.



based on correctness exercises included in the contents units. Interface problem coordination is performed under the instructor responsibility, through a chat tool.

**Content units:** the bottom level OPN that implement the sub-problem and integration problem units are implemented using the FAST tool. Their general form is shown in Figure 8, where *Exp*, *Exa* and *Exe* are interaction units that present to the students, respectively, explanations, examples and exercises.

## 6 Conclusion

This paper has presented a method, based on ontologies and Petri nets, to allow the development of group learning in the context of *Intelligent Tutoring Systems (ITSs)*. While ontologies represent domain and student models in a shareable format, Petri nets formally specify group interaction protocols. The method is intended to be used with ITSs that are built using the FAST authoring tool. This allows the method to explore the student and domain models of these ITSs to increase the adaptiveness of the interaction. The method includes a library of group activities scenarios, previously defined by the system developers. To build a group activity instance, the teacher chooses one scenario, customizes its parameters and provides the contents of the activity. This information is compiled into an object Petri net that guides the group activity. The use and update of the group activity and student models is automatically included in the transitions of this Petri net.

Presently, the proposed method only allows the management of intra group activities, coordinating the tasks performed by the students that are member of a group. We intend to extend the method to allow the management of inter group activities, increasing the complexity of possible scenarios. Ongoing work also includes the enhancement of the student model attributes that are relevant to group activities and the development of further group activity scenarios.

As future work we intend to develop a friendlier authoring tool to develop these group activity scenarios.

## Acknowledgements

This work was partially supported by CAPES/Cofecub (processes 400/02 and 0212/05-9) and CNPq (Brazilian National Research Council) (process 140005/2004-8).

## References

- [1] Alpert, S.R., Singley, M.K., Fairweather, P.G., 1999. Deploying intelligent tutors on the web: An architecture and an example. *JAIE*, 10:183–197.
- [2] Brusilovsky, P. 2000. Adaptive hypermedia: From intelligent tutoring systems to web-based education. *LNCS, Intelligent Tutoring Systems 2000*.
- [3] Cardoso, J., Bittencourt, G., Frigo, L.B., Pozzebon, E., Postal, A., 2004. MathTutor: A multi-agent intelligent tutoring systems. In 1st IFIP Conf. on AI Applications and Innovations, WCC'04, pages 22-27.
- [4] Cardoso J., Bittencourt, G., Frigo, L.B., Pozzebon, E., 2004. Petri nets for authoring mechanisms. In XV Simpósio Brasileiro de Informática na Educação (SBIE'2004), ISBN 85-7401-161-4, pages 378–387, Manaus, AM, Brazil.
- [5] Chen, W. & Wasson, B. 2004. Intelligent Agents Supporting Distributed Collaborative Learning. In: Lin, O., red. *Designing Distributed Learning Environments with Intelligent Software Agents*. IDEA Publishing Group;
- [6] Constantino-González M., Suthers, D., Santos, J.E.G., Coaching Web-based Collaborative Learning based on Problem Solution Differences and Participation, in *International Journal of Artificial Intelligence in Education*, 2003, vol 13, 263 - 299
- [7] Cost R.S., Chen Y., Finin T., Labrou Y. and Peng Y., Modeling Agent Conversations with Colored Petri Nets, In *Proc of the Workshop on Specifying and Implementing Conversation Policies*, Seattle, May 1999, pp. 59-66.
- [8] Costa, E.B., Lopes, M.A., Fereda, E, 1995. Mathema: A learning environment based on a multiagent architecture. In *LNAI - Advances in Artificial Intelligence*, volume 991, pages 141–150.
- [9] Coutinho, L., Sichman, J.S., Boissier, O, 2005. Modeling Organization in MAS: A Comparison of Models. in: *First Workshop on Software Engineering for Agent-oriented Systems*, Uberlândia.
- [10] Cuesta, P., Gomez, A., Gonzalez, J.C., Rodriguez, F., A Framework for Evaluation of Agent Oriented Methodologies. The MESMA Approach for AOSE. *Proceedings of Fourth Iberoamerican Workshop on Multi-Agent Systems (Iberagents'2002)*, at IBERAMIA'2002, the VIII Iberoamerican Conf. on Artificial Intelligence, Malaga, Spain.
- [11] Dignum, M.V., Vázquez-Salceda, J., Dignum, F.P.M., 2004. OMNI: Introducing social structure, norms and ontologies into agent organizations. In P. Bordini & et al. (Eds.), *PROMAS 2004* (pp.183-200). Heidelberg: Springer.
- [12] Esteva, M., Padget, J., Sierra, C., 2001. Formalizing a language for institutions and norms. *Proceedings of the Eighth International Workshop on Agent Theories, Architectures, and Languages (ATAL-2001)*, pp 106–119.
- [13] Ferber, J., Gutknecht, O., Michel, F., 2003. From Agents to Organizations: an Organizational View of Multi-Agent Systems. 4th Int. Workshop on agent-oriented software engineering, IV AOSE 2003, Melbourne, Australia.

- [14] Frasson, C., Martin, L., Gouarderes, G., Aimeur, E., 1998. Lanca: A distance learning architecture based on networked cognitive agents. In *Lectures Notes in Computer Science. Intelligent Tutoring Systems. Proceedings of 4th International Conference*, 1:594–603, August 1998.
- [15] Frigo, L., Cardoso, J., Bittencourt, G., 2005. Adaptive Interaction in Intelligent Tutoring Systems. In : HT-2005, Intern. Workshop on Combining Intelligent and Adaptive Hypermedia Methods/ Techniques in Web-based Education Systems, Salzburg, Autriche.
- [16] Hanachi, C., Sibertin-Blanc, C., 2004. Active Middle-Agents in Multi-Agent Systems. Dans : *Autonomous Agents and Multi-Agent Systems*, Kluwe Academic Publishers Netherlands, V. 8 N. 3, p. 131 - 164, avril 2004
- [17] Harrer, A., McLaren, B., Walker, E., Bollen, L., Sewall, J. (2005). Collaboration and Cognitive Tutoring: Integration, Empirical Results, and Future Directions. In C.-K. Looi et al. (Eds.), *Proceedings of the 12th International Conference on Artificial Intelligence in Education* (pp.266-273). Amsterdam: IOS Press.
- [18] Hernandez-Dominguez, A., 1998. An Architecture of Cooperative Learning in a Distance Education context *International Conference on Engineering Education*, 1998, August 17-20, Rio de Janeiro, Brazil.
- [19] Horling, B., Lesser, V., 2004. Quantitative Organizational Models for Large-Scale Agent Systems. *Proceedings of the International Workshop on Massively Multi-Agent Systems*, Melbourne, LNCS 2935, pp 214-230, December 2004. Kyoto, Japan
- [20] Iglesias C., Garijo M., Gonzales J. C., A survey of Agent-Oriented Methodologie", in J.P. Müller, M.P. Singh and A. S. Rao, eds, *Proceedings of the Fifth International Workshop on Agent Theorie, Architecture and Languages (ATAL 98)*, LNAI, vol 1555, Springer-Verlag, Heidelberg, 1999.
- [21] Labidi, S., Souza, C.M., Nascimento, E., 2004. NETClass: Cooperative Learner Modeling in Web-Based Environment *6th International Conference on Computer Based Learning in Science CBLIS*, Nicosia.
- [22] Miao, Y., Pinkwart, N., Hoppe, U., 2006. Conducting situated learning in a collaborative virtual environment. *Proceedings of the 5th International Conference on Web Based Education*, pp. 7-12. Anaheim, CA: ACTA Press. 2006
- [23] Mizoguchi, R., Bourdeau, J., 2000. Using ontological engineering to overcome common AI-ED problems. *J. of AI in Education*, 11(2):107–121, (2000).
- [24] Odell, J., Nodine, M., Levy, R., 2005. A Metamodel for Agents, Roles, and Groups. *Agent-Oriented Software Engineering (AOSE)* V, James Odell, P. Giorgini, Jörg Müller, eds., LNCS, Springer, Berlin.
- [25] Sibertin-Blanc, C., 1985. High-level Petri nets with data structures. In *European Workshop on Application and Theory of Petri Nets*, pp 141–170.
- [26] Silva, V., Choren, R., Lucena, C., 2004. A UML based approach for modeling and implementing multi-agent systems. In [AAMAS 2004], pages 914.921.
- [27] Thibodeau, M., Belander, S., Frasson, C., 2000. White rabbit matchmaking of user profiles based on discussion analysis using intelligent agents. *Proceedings of 5th International Conference, ITS 2000*, 1:113–122, June 2000. Montreal/Canada.
- [28] Zambonelli, F., Jennings, N., Wooldridge, M., 2001. Organisational Abstractions for the Analysis and Design of Multi-Agent Systems. In: Ciancarini P., Wooldridge, M. (eds.): *Agent Oriented Software Engineering*, LNCS 1957, Springer-Verlag.



# Reliability, Availability and Security of Wireless Networks in the Community

Carsten Maple, Geraint Williams and Yong Yue  
 Institute for Research in Applicable Computing  
 University of Bedfordshire  
 Luton Campus, Park Square, Luton, Beds, LU1 3JU, UK  
 E-mail: carsten.maple@beds.ac.uk, geraint.williams@beds.ac.uk, yong.yue@beds.ac.uk

**Keywords:** security, WEP, WLAN, WPA, risk assessment, risk management, threat analysis.

**Received:** March 12, 2007

*Wireless networking increases the flexibility in the home, work place and community to connect to the Internet without being tied to a single location. Wireless networking has rapidly increased in popularity over recent years. There has also been a change in the use of the internet by users. Home users have embraced wireless technology and businesses see it as having a great impact on their operational efficiency. Both home users and industry are sending increasingly sensitive information through these wireless networks as online delivery of banking, commercial and governmental services becomes more widespread. However undeniable the benefits of wireless networking are, there are additional risks that do not exist in wired networks. It is imperative that adequate assessment and management of risk is undertaken by businesses and home users. This paper reviews wireless network protocols, investigates issues of reliability, availability and security when using wireless networks. The paper, by use of a case study, illustrates the issues and importance of implementing secured wireless networks, and shows the significance of the issue. The paper presents a discussion of the case study and a set of recommendations to mitigate the threat.*

*Povzetek: Zanesljivost, varnost in dosegljivost brezžičnih omrežij.*

## 1 Introduction

The use of the Internet is increasing by home users has been continually increasing for some time, and this increase can also be seen in the context of businesses, banking and governmental services. According to the Office for National Statistics, more than half of the households in the UK had some form of Internet connection in 2005. The number of households that have an Internet connection is a statistic that has steadily risen from just over 30% of households in 2000, to 54% of households in 2005 [Office for National Statistics 2007]. In an article in *New Scientist* in 2005, Graham-Rowe predicted the growth of broadband Internet connection and wireless networks in the home up until the year 2009. It was predicted that in 2006 the percentage of households in Western Europe with both a broadband connection and wireless network would be approximately 12% [Graham-Rowe 2005]. As the number of households that utilise an Internet connection, so too does the range of services that are accessed.

In recent years, industry has recognised the benefits in teleworking. Companies are no longer concerned that those working away from the office are simply not working. Indeed, if anything, those working away from the office now tend to work too much. In 2004 in all US government departments, there 751, 844 employees deemed to be eligible for teleworking out of 1, 749, 998,

a total of 43% of the workforce. Of those eligible for teleworking, 102, 291 undertook the activity at least some of the time, 14% of those eligible or 6% of the total workforce.. In the UK similar statistics were witnessed in 2001. In the spring of 2001, 2.2 million people in the UK, representing 7.4 per cent of the total labour force, worked from home at least one day a week using both a telephone and a computer to undertake their duties. This had represented a dramatic growth over previous years. The total number of teleworkers in the UK in 2001 had increased by nearly 70 per cent over the period 1997 to 2001 [Office for National Statistics 2002]. According to the *Sunday Times* the percentage of staff working one morning or afternoon a week at home has more than doubled in the past two years. The newspaper reports that according to research for *The Sunday Times* 100 Best Companies to Work For 2007 List, the figure has risen from 14.1% to 31.7% [Thomas 2007].

The nature of how teleworkers operate has also changed. In 2001, 71% of teleworkers operated in different places using home as a base. This percentage had increased to 75% of all teleworkers operating in different places using home as a base. The increase in the number of teleworkers operating in a different place using home as a base had increased from 1.05 million in 2001 to 1.77 million in 2005, an increase of 60% [Office

for National Statistics 2005]. This increase in the mobility of teleworkers presents a significant increase in risk of security compromises.

Year	Homeworkers			Teleworkers		
	Works mainly at home	Work in different places	Total	Works mainly at home	Work in different places	Total
2001	673	1,916	2,590	433	1,051	1,484
2002	698	2,062	2,760	484	1,381	1,865
2003	707	2,207	2,915	516	1,553	2,069
2004	761	2,243	3,004	562	1,599	2,161
2005	768	2,324	3,092	603	1,774	2,377

An increase in the number of employees' teleworking represents a risk that can undermine security in business. However, there are other trends that point to an increase in risk to the public in general as well as businesses. In the UK a high percentage of adult users use e-government services to retrieve information; however, the actual level of interaction with the service is low, especially compared to Europe [Eurostat Online European Statistics 2006]. With traditional governmental services, a citizen usually visits a government office and is authenticated by presenting papers. Identification to an official is replaced with a digital service where credentials are electronic and identity theft can become an increasing problem. Security of the citizen's connection is an important part of ensuring a secure e-government service. Where the services are provided to the citizens of a country the term c2g (Citizen to Government) is used in much the same way as b2c and b2b are used for business to consumer and business to business, respectively. The UK government has defined c2g as the online relationship between its citizens and various government departments they communicate with.

In Europe the eEurope 2005 Action Plan sets 7 targets for e-services. These are:

- **Interactive public services:** Member States should have ensured that basic public services are interactive, where relevant, and accessible for all. The Commission and Member States must agree on a list of public services for which interactivity and interoperability are desirable. Relevant issues include exploiting the potential of broadband networks and multi-platform access, and addressing access for people with special needs;

- **Public procurement:** Member States should carry out a significant part of public procurement electronically, cutting costs and raising efficiency in government procurement. The European Parliament and Council should adopt as quickly as possible the legislative package on procurement;

- **Public Internet Access Points (PIAP's):** All citizens should have easy access to PIAP's, preferably with broadband connections, in their communes or municipalities. In establishing PIAP's, Member States should use structural funds and work in collaboration with the private and/or voluntary sector, where necessary;

- **Broadband connections:** Member States should aim to have had broadband connections for all public administrations by 2005. Authorities should not discriminate between technologies when purchasing connections;

- **Interoperability:** The Commission presented a staff working paper on the importance of interoperability for e-Government services at the 2003 e-Government Ministerial Conference and intends to propose a European interoperability framework for pan-European e-Government services that provides a series of recommendations and defines generic standards with regard to organizational, semantic and technical aspects of interoperability, offering a comprehensive set of principles for European co-operation in e-government;

- **Culture and Tourism:** The Commission, in co-operation with Member States, the private sector and regional authorities, will define and launch e-services to promote Europe and to offer user-friendly public information. Building on the Communication "Working together for the future of European tourism", the Commission is now developing a European Tourism Portal. This work is being undertaken through an ETD project that involves a collaboration between EC3, LiXto, ITC-irst, Siemens Austria and Tiscover and is currently in the second stage of development;

- **Secure communications between public services:** The Commission and Member States have proposed to examine the possibilities to establish a secure communications environment for the exchange of classified government information.

The EU has agreed a list of 20 basic services that should be available as part of e-government. These consist of 12 relating to citizens and a further 8 services are aimed at businesses:

Citizen services	Business Services
Job search	Social contribution for employees
Income taxes	Corporate tax
Social security benefits	VAT
Personal documents	Registration of a new company
Car registration	Submission of data to the statistical office
Application for building permission	Custom declaration
Declaration to the police	Environment-related permits
Public libraries	Public procurement
Birth and marriage certificates	
Enrolment in higher education	
Announcement of moving	
Health-related services	

It is also these 20 services that have been used by the EU and researchers for benchmarking the performance of e-government accessibility.

For e-government to be a success with the citizens, they must be able to connect to the government infrastructure in a dependable manner where there is

reliability, availability and security of connection. A government has great control of the infrastructure connecting its offices and departments together but not the connection to business and in particular its citizens rely on external third parties. Additionally, some of the responsibility for the dependability will rely on the citizen’s ability to set-up a connection which implies a level of knowledge that is likely that the citizen does not have [Furnell 2005].

An important aspect of the e-government is how its citizens access the systems and the infrastructure extends beyond the government offices and officials to wherever the citizens access the services within the community.

In a household where the Internet connection is either direct to the PC or a wired network, anybody trying to access the network will have to come through the modem/termination unit and if it has been configured correctly there should be a firewall and/or NAT set-up. In a wireless networked environment, the security of the network is not as controlled.

## 2 WLAN security

Wireless networking has experienced a huge increase in popularity over the last couple of years. The necessary hardware is widely available to consumers, it is affordable, and relatively easy to install and configure. Gateway devices, such as "routers" or "firewalls", that allow users to share a broadband connection with and protect multiple computers on a home network have been utilised for some time and have increased in popularity as more users in the home see the need for the use of an Internet connection and access to the same peripherals, such as printers. The addition of wireless capabilities to these gateway devices gives the user the convenience of taking a computer anywhere in the house without running wires through floors and attics.

### 2.1 Operation of 802.11 networks

Wireless communication under 802.11 [802.11 Working Group 2006] comes in two flavours: the Independent Basic Service Set (IBSS) or ad-hoc mode, and the Basic Service Set (BSS) Infrastructure mode. The IBSS mode allows wireless clients to talk directly to each other without a central controlling mechanism. To join such a network all that is needed is the Service Set Identifier (SSID) and the channel it operates on. This mode is generally considered to be insecure. Although it supports Wired Equivalent Privacy (WEP), it is not supported by WiFi Protected Access (WPA); however under 802.11i and the introduction of WPA2 ad-hoc networks will support better encryption. It is these ad-hoc wireless networks form the basis of mesh wireless networks being implemented by cities across the world. The Cloud, a wireless broadband service provider, has announced the creation of meshes in 9 cities across the UK. The BSS mode uses a controlling mechanism, normally by an access point to control the communication. The BSS functions in a manner similar to a switched wired network while the IBSS operates in a manner comparable

to a hub based wired network. In BSS mode all traffic is routed through the access points (AP) even between peer wireless devices. There is an extension to the BSS where multiple BSS can be connected via a wired network connecting the access points known as the Extended Service Set (ESS).

An SSID is the name of a wireless local area network (WLAN). All wireless devices on a WLAN must employ the same SSID in order to communicate with each other. The SSID on wireless clients can be set either manually, by entering the SSID into the client network settings, or automatically, by leaving the SSID unspecified or blank. SSIDs are case sensitive text strings containing a sequence of alphanumeric characters with a maximum length of 32 characters. All access points come with a default SSID set, some of the more common ones are:

Manufacturer	Default SSID
Cisco	tsunami
3Com	101
Lucent/Cabletron	RoamAbout Default Network Name
Various	Default SSID
Compaq	Compaq
Addtron,	WLAN
Intel	intel
Linksys	linksys
Various	Wireless

Wireless networks work in two modes: the standard default 802.11 mode of broadcasting their SSID and a mode in which the SSID is not broadcast. The former is known as an open system and the latter is a modified closed system which is a propriety addition to the standard.

An open system will broadcast management beacon frames at a fixed interval that contain capability information of the access point. This is intended to enable wireless clients to detect the closest access point and if there is a stronger signal with the same SSID, it re-associates with the signal allowing roaming between access points.

A closed system does not broadcast the SSID as part of the beacon management frame and a client must have prior knowledge of the SSID to enable it to join a closed system using a probe request. However, some access points will respond to probe requests that contain a blank SSID or contain an SSID set to “any”. This behaviour is modifiable on some access points, typically enterprise class equipment.

The Microsoft Windows XP SP 1-based Wireless Zero Configuration service suffers from what Microsoft calls "behaviour by design." If the wireless network is set to so that it does not broadcast its SSID, Microsoft's wireless manager periodically drops its non-broadcasting WiFi connection in response to the presence of a broadcasting SSID-based network.

### 2.2 Detecting wireless networks

There are two basic techniques for locating wireless access points: passive or active searching based on 802.11. Both approaches are easy to implement and only



need the most basic equipment and set-up. Each of these techniques employs a process known as “sniffing”, where the wireless card listens for management packets. There is a great deal of commercial and open source software that can do this and will also record the details of the packets; the wireless client on a device does this as part of its normal function. Sniffing software takes this functionality and applies additional techniques to allow listening to take place on all possible channels; by switching the channel the card is working on at a regular interval.

Additionally, certain wireless card chipsets are more flexible and are more suitable for use with sniffing software. The best cards are ones that can be placed into what is called the Monitor mode, also known as RFMON mode, which is similar to promiscuous mode on wired network interface cards. It allows the wireless card to sniff all the traffic that the card receives instead of sniffing all the traffic from the associated network. Usually, the card is unable to transmit or otherwise be used when in this mode. It is also used for passive stumbling, a technique used in wardriving where the wireless card listens for base stations instead of actively probing them to determine their presence.

Passive searching is to purely listen for the transmitted management beacon frames. It only needs to be within receiver range to detect a network; no traffic is generated by passive sniffing. Passive sniffers are also capable of recording data packets for additional dissection. However, they require a card and driver capable of radio frequency (RF) Monitor support, which enables raw packet detection. They cannot detect a non-beaconing network with no data traffic, though it is possible to record packets for analysis for encryption breaking and MAC address searching.

Active searching uses the probe response/request feature where a probe request sent from the client to the access point for information results in the probe response frame if the access point is so configured. Inexpensive wireless access points intended for home use do not normally allow the user to disable the beaconing mechanism; this level of configuration is normally found on enterprise class equipment. This method does not need traffic to be transmitted across the network for a network to be found; however, it does generate packet traffic that can be detected by intrusion detection systems. This form of active scanning using probe requests is not completely effective at finding all wireless access points.

An alternative active method of scanning to detect networks, where both SSID broadcast and probe response are only actuated by a probe request with a valid SSID, is to passively listen for a communication session between the access point and client and then issue a disassociation request. This causes the client to break the communication link and after a short period of time, it issues a probe request and re-associates with the access point allowing the management frames to be captured and the SSID to be identified.

Typically, the minimum wardriving kit consists of a laptop, wireless network card and sniffing software

though many use a Global Positioning System (GPS) unit to provide geographical location information. Additional mapping utilities are needed to generate maps of locations of detected wireless networks if the positional data is recorded. GPS is an American system of 24 orbiting satellites that can provide a positional fix to a resolution of a metre though most commercial equipment is less accurate than this.

### 2.3 Wireless security measures

The main insecurity with wireless networks compared to wired networks is the ease of accessing the transmission medium used, i.e. with a wired network to sniff packets, there has to be a physical access to the network whilst with wireless networks, the transmission is easily available outside the physical building. Insecurities on wireless networks other than those caused by the ease of accessing the transmission media are the same as for a wired network, i.e. packets can be sniffed if sent in clear text across wires if someone has packet sniffing software on the same segment of the network as the packet is being transmitted across.

In order to establish some form of protection for wireless networks, the WEP algorithm, has been developed as part of the 802.11 standard. It restricts access to the network to those who has the same key and is supposed to give the equivalent privacy as those on a wired network. However, a number of flaws have been discovered in the WEP algorithm [Fluhrer, et al 2001, Mead and McGraw 2003], which seriously undermine the security of the system and leaves the system open to a number of attacks:

- Passive attacks to decrypt traffic based on statistical analysis.
- Active attack to inject new traffic from unauthorised mobile stations, based on known plaintext.
- Active attacks to decrypt traffic, based on tricking the access point.
- Dictionary-building attack that, after analysis of about a day's worth of traffic, allows real-time automated decryption of all traffic.

It is practical to mount these attacks using only inexpensive off-the-shelf equipment. It is recommended that anyone using an 802.11 wireless network does not rely on WEP for security, but rather employ other security measures to protect the wireless network. The effectiveness of the attacks applies to both the 40-bit and the so-called 128-bit versions of WEP equally well.

The 802.11 standard uses Ethernet packets across wireless networks, Ethernet uses a Media Access Control (MAC) address which is a hardware address that uniquely identifies each node of a network and is configured as part of network interface card (NIC) and regulated by the IEEE. This allows a further method of securing a network by the use of MAC filters, which restricts access to an AP to authorised MAC addresses only. Most APs provide this capability for checking the MAC address of the station before allowing it to connect to the network, thus providing an additional control

layer. However this approach requires that the list of MAC addresses be configured and maintained. This can be circumvented since MAC addresses are transmitted as part of the Ethernet frames and can be read from captured packets. It is, however, possible to spoof the MAC address of node using various software such as SMAC, a MAC address modifying utility for Windows operating systems, regardless of whether the manufacturers allow this option or not. This allows an intruder to alter the MAC address of the node to match a known node on the network.

Further security is available using WiFi Protected Access (WPA and WPA2) which was created in response to the serious weaknesses in WEP. WPA implements the majority of the IEEE 802.11i standard, and is intended as an intermediate measure to take the place of WEP while 802.11i is prepared. It is designed to work with all wireless network interface cards, but not necessarily with first generation wireless access points. WPA2 implements the full standard, but does not work with some older network cards. There are two modes of operation for WPA:

- Personal mode or Pre-Shared Key (PSK) mode is designed for home and small office networks that cannot afford the cost and complexity of an 802.1X authentication server. Each user must enter a passphrase to access the network. The passphrase may be from eight to 63 ASCII characters or 64 hexadecimal digits (256 bits). The passphrase may be stored on the user's computer at their discretion under most operating systems to avoid re-entry. The passphrase must remain stored in the WiFi access point.
- Enterprise mode is designed for larger offices and enterprises various Extensible Authentication Protocol (EAP) types are now supported in enhanced WPA/WPA2 compared to just the Temporal Key Integrity Protocol (TKIP) in the original WPA and EAP-TLS in WPA2

Additional security can be obtained by deploying standard TCP/IP security protocols over the connections, such as IPSec and the use of Virtual Private Network (VPN).

## 2.4 Legal issues

The legality or rather the illegality of accessing a wireless networks is not in question; under most legal systems accessing a computer system without permission is illegal. However, the activity of wardriving is more of a grey area as often the legality is dependent on local laws. Law enforcement officials, increasingly concerned about wireless networks, say the possibilities for mischief run the gamut. A wireless hacker's intentions could be as malevolent as identity theft or as benign as using a neighbour's Internet connection to check e-mail or scan the newspaper online. Sometimes they drain other people's bandwidth to illegally download movies and other copyrighted material or access pornography.

Others are pranksters, who maliciously lock people out of their wireless networks just for fun, others still are those that are spammers, using unauthorised Internet access to send masses of unsolicited e-mail.

In the UK, it should be said that listening to broadcasts in the Industrial, Science and Medical (ISM) band is not illegal as this is a license exempt band, indeed the 802.11 standard is written so that wireless networks broadcast their presence and hence only the built-in mechanism of the transport system is used. This effectively allows passive scanning to take place, and since active scanning uses the mechanisms built into the standard to identify a network. Though a slight increase in electrical power usage may be witnessed or a very slight delay in a transmission of a packet may be caused, this is not considered to be illegal as it is part of the client functionality to identify and attach to a network with a hidden SSID. Whilst this is the case, there may be moral and ethical considerations of what is done with the information gathered from active scanning. Any activity resulting in an association with an access point, even accidentally, could be considered illegal - accessing information or sending data across the network without the permission of the network administrators would be considered illegal in many places around the world.

If a disassociation packet is issued, causing a client to lose contact with the network and reconnect to the network to generate extra traffic this would result in a degradation in performance of the network and this may be deemed illegal in some countries.

An interesting consideration is whether it should be illegal if, by using passive means, enough information is collected to decode the WEP key and no further use of the key made. No connection is made to the network, its performance is unaffected and the information is transmitted publicly and any information broadcast on the ISM can be listened to. However, information vital for the security of the network has been gained leading to a potential compromise.

## 2.5 Availability and reliability

Interference to wireless networks can come from a number of sources, the 802.11 standard uses the 2.4GHz and the 5GHz bands which are generally unlicensed or license exempt around the world. The 2.4GHz band which is part of the ISM band is crowded with a large number of devices ranging from wireless baby monitors, microwave ovens to cordless phones. Wireless networks working in this band have to contend with all these devices, plus other wireless access points in the same area.

Although the 2.4GHz band is divided into 14 channels (not all are available around the world), the bandwidth of each channel is sufficient that adjacent or near adjacent channels interfere with each other. Of the 11 channels commonly used in the USA or the 13 channels used in the majority of Europe, a maximum of three channels are spread enough apart to avoid interference problems.

One of the problems in hiding or disabling the SSID is that it becomes harder for those configuring wireless networks to identify nearby networks operating on the same or adjacent channels.

Additionally most emerging radio technologies for Wireless Personal Area Networks, such as the Bluetooth protocol, are designed also to operate in the 2.4GHz ISM band. Since both Bluetooth and IEEE 802.11 devices use the same frequency band and are likely to come together in a laptop or may be close together at a desktop, interferences may lead to significant performance degradation.

802.11 is a Collision Sense Multiple Access Protocol where each wireless point checks for another wireless point transmitting: if it detects a transmission it will wait for a random length time delay and try again; if two transmit at the same time, both detect a collision and wait for a random time delay and try again. The more points on the same channel within the range, the more likelihood of a collision exists and increased collision results in reduced data transmission rates. However, the 14 usable channels allocated worldwide allow wireless networks in close proximity to each other to use different channels. The drawback is that the channels are separated by 5MHz; each channel has a bandwidth of 22MHz causing it to interfere with adjacent channels which actually means that in the 11 or 13 channel implementation, it is only possible to utilise three channels concurrently without any overlap in frequency, typically taken as channels 1, 6 and 11.

## 2.6 Security risks

With a wired network, the only possible access for an intruder would be through the broadband connection and normally through a firewall in the case of a broadband modem. With Network Address Translation (NAT) and security software on the individual computers, there is a fairly comprehensive layered defence system whose effectiveness would depend on the ability of the installer or the default settings.

A wireless network allows access from outside the property onto the network behind the broadband modem with its firewall and NAT. This means instead of intruders needing to get through the broadband connection, the network has to deal with intruders connected directly to the network. Since the wireless router is designed to provide a wireless Internet connection and its range can reach as far as 150 feet it can often reach many public roads and nearby homes. An attacker could be outside a home or business with a laptop with a WiFi card and the right software, gain access to private information on the network.

Instead of gaining access the public side of the gateway device, the intruder connects directly to the network on the private side of the gateway device, completely bypassing any hardware firewall between the private network and the broadband modem. Many people assume that since they are behind a firewall their private network is safe, letting down their guard, sharing drives, and generally being less careful about security. The

intruder can take advantage of this by perusing devices and gaining access to confidential data such as personal information (financial data, tax records and wills) and work-related information such as confidential specifications and trade secrets that the victim brings home from the office. By employing a sniffer an intruder can also sniff email or FTP user names and passwords since they are usually transmitted in clear text. This allows unauthorised access to email accounts or web servers without the victim's knowledge.

Another risk posed by lack of security is identity theft. Whilst there has been no direct conviction for this there is certainly a great deal of evidence that it occurs. By using information such as tax returns and resumes obtained by compromising a home network, it is possible to use the name, address, date of birth and National Insurance number to create a bogus identity. The increasing use of on-line services for banking and government makes it more tempting to use alternative identities or to gather information on an individual to impersonate them. A poll by Winmark Research, on behalf of RSA Security, found that two-thirds of consumers used the same password to access different types of websites - from email to bank accounts [Leyden 2004]. One third even admitted to sharing passwords with friends and family, massively increasing the risk of fraud. In the survey, the most common password categories were family names such as partners or children (15%), followed by football teams (11%) and pets (8%), the most common password being "admin". Many workers who regularly had to change their passwords kept them on a piece of paper in their desk drawers, or stored them with a Word document. With the ability to search people's computers via a wireless attack, it makes it easier for identity theft to occur. Secured networks could be jeopardised as a result of human vulnerabilities such as lack of awareness and adherence to usage policies [Bhagyavati, et al 2004].

## 3 Case study of Luton

During Dec 2005 and Jan 2006, an extensive wardrive around Luton was undertaken as part of a research project into the extent of security in wireless networks in the community. The wardrive was conducted by one of the authors using equipment only capable of detecting 802.11b and g networks and works only in the 2.4GHz band range. It uses active scanning and only detects networks that respond to a general probe request. The analysis looked at number of secure and unsecured networks, distribution of channels and percentage of the population of Luton and combined it with data from the last national census conducted in 2001 and statistics from the National Statistics Office on number of households with Internet and Broadband access. The wardriving survey of Luton shows similarities with those conducted in London and Bristol in the UK and in Frankfurt and Paris in Europe [Jaques 2005].

Luton is a typical town in the UK within the southeast of the country about 30 miles north of London on the M1 with good communication links. It has a population of

187000 living in 70775 households (Data obtained from the 2001 Census results), covering an area of 436 hectares and used to be a major automotive manufacturing town in the 60's and 70's. Looking at the national average of 55% households having Internet access, about 38900 households in the town would have Internet access; however in the southeast, the percentage of households with Internet access raises to 64% and if Luton was typical of towns in this region, the estimate would be 45300 households.

Tentative results from the wardriving survey shows that for the whole of Luton, it is expected that there are around 4000 wireless networks with 50% of these being unsecured. These results are only for the 2.4Ghz band and 802.11b and g wireless networks. This means that there may be networks that have not been detected and the number of networks would then be higher.

There are 24,479,439 households in the UK, of which it is estimated that around 13,463,700 have Internet access. It is also thought that 10% of these have a wireless network, meaning an estimated 1,346,370 wireless networks. If the figure of 50% of all wireless networks being insecure holds throughout the country, it would mean that there were approximately 673,185 unsecured wireless networks in the UK. Confirmation of the order of magnitude of wireless networks can be seen in a report by Contractor UK [Contractor UK 2005] quoting research by IDC who estimated there were 958,000 wireless networks in the UK, with this figure expected to double to almost two million by the end of the year.

A further characteristics examined in the case study was the channel usage of the detected networks. As expected, channels 1, 6 and 11 were the most commonly used channels with channel 11 being used by 52% of the detected networks. The actual distribution of the channels has not been fully analysed but it can be expected that in some areas the wireless networks are not running at the best possible bandwidth expected from ideal conditions due to the devices suffering from collisions generated by nearby networks. The average user, however, will be unaware that such collisions may be taking place and may well apportion blame elsewhere.

A further characteristics examined in the case study was the ratio of ad-hoc to infrastructure and the usage of SSID's. In the survey, 2% of the detected networks were in IBSS or ad-hoc mode, and the majority detected were infrastructure mode. In a sample of 2363 networks, there were 1077 distinct SSID's; however, the majority of the SSID's found were the original default settings of the equipment.

## 4 Recommendations

There are a great deal of sources available that provide information regarding the construction of secure wireless networks. These sources can be found easily and can be implemented with little difficulty for those with some technical knowledge and confidence. However, most of those installing a wireless network do not have technical

knowledge nor confidence and may fail to understand the problems associated with wireless networks.

We recommend that to improve the availability and reliability of a wireless network, a person configuring the network should conduct a small site survey using widely available free of charge software. The results of this survey can then inform the choice of location and channel. Due to the large increase in use of wireless networks, one should also check regularly to examine if the conditions around the location have changed and adjust settings accordingly. This may well be beyond the ability of most of the general public and additionally they may not have the tools to conduct the survey. In such cases this could be a service that is sourced and should be of low cost. It is also possible using most wireless client software, to display a list of available networks but this will not normally show those with hidden SSIDs. However, as a minimum it is possible to use the client software to detect wireless networks and the channel they are working over and then setting the channel of the access point to one that will have the least interference.

There are a great deal of information sources that recommend users stop or to disable the SSID broadcast; we strongly recommend against this action. The disabling of an SSID broadcast offers very little increase in security to anyone attempting to access a network, it only stops beacon broadcast on the access point. Essentially, disabling SSID broadcast just stops the inclusion of SSID's in the broadcast beacon frame which is the one of the five SSID broadcast mechanisms. Using Microsoft Windows XP SP 1-based Wireless Zero Configuration service to manage the wireless network card suffers from what Microsoft calls "behaviour by design." If the wireless network is not set to broadcast the SSID, Microsoft's wireless manager periodically drops the non-broadcasting WiFi connection in response to the presence of a broadcasting SSID-based network. Thus, in practice, it does not improve security to stop the SSID broadcast and doing so can cause problems with wireless networks. The inclusion of the SSID in the standard was to aid management of wireless networks and it really should be used for doing this. Allowing users of wireless networks to identify channels being used and select other channels can reduce interference. Whilst those with advanced knowledge can examine networks in the vicinity, the average user will not be able to examine those that withhold the SSID and so collisions may be rife. This can be seen in the case study in which 52% of networks were operating on the same channel.

A recommended approach to wireless security would be to use a layered approach with MAC filtering and WPA or WPA2 at the access point. The use of IPSec and VPNs on the network and ensuring the machines on the network are protected. Ultimately, it may be the best to follow the practice to put publicly accessible servers into a Demilitarised Zone (DMZ) and put the wireless access point into a firewalled section of the network with rules governing communications to the rest of the network; however until equipment for the home can support this, it will remain a security weakness.

In general it is recommended to apply extra caution to wireless connections in a public area as they may not provide as much security as wired Internet connections. In fact, many "hotspots" - wireless networks in public areas like airports, hotels and restaurants - reduce their security. Unless a security token is used, it may be decided that accessing an online bank account through a wireless connection is not worth the security risk of a snooper capturing your packets and decoding them.

One of the most important recommendations would be to the manufacturers of wireless networking equipment to provide information in an easy to understand format on setting up wireless network and possible for governments to put pressure on the manufactures to do so. A better informed public will result in better set-up wireless networks and hence better availability, reliability and security of them.

## 5 Conclusions

In this paper we have discussed some of the key concerns surrounding the security of wireless networks. We have highlighted a number of weaknesses in existing protocols and configurations of wireless networks including how these weaknesses can be exploited. The paper has also considered some of the legality aspects of accessing information regarding the configuration of a wireless network as well as the accessing of transmitted or stored information on the network.

A case study has been presented that demonstrates the extent of the problem and this study is to be used as a basis for further work. Additional equipment will be used later in the study to detect 802.11, b and g across both the 2.4GHz and 5GHz bands. The research investigates wireless networks within the community and looks at aspects of reliability, security and whether education or training would help reduce potentially insecure networks and improve the reliability and availability of them. The wardrive identifies the number of wireless networks and their distribution around the town.

We have presented a number of recommendations that can ensure the greater security of wireless networks. These recommendations require action from both manufacturers and those configuring a wireless network, most often the end user of the equipment.

## References

- [1] 802.11 Working Group, 2006. <http://grouper.ieee.org/groups/802/11/>
- [2] Bhagyavati, Summers, W.C. and DeJoie, A. 2004. Wireless security techniques: an overview, *Proceedings of 2004 Conference for Information Security Curriculum Development*, Kennesaw, Georgia,.
- [3] Contractor UK, 19 January 2005. Wireless hackers creep nearer to UK homes. <http://www.contractoruk.com/news/001908.html>
- [4] Eurostat Online European Statistics, [http://epp.eurostat.cec.eu.int/portal/page?\\_pageid=1073,46870091&\\_dad=portal&\\_schema=PORTAL&p\\_product\\_code=IR111](http://epp.eurostat.cec.eu.int/portal/page?_pageid=1073,46870091&_dad=portal&_schema=PORTAL&p_product_code=IR111)
- [5] Fluhrer, S., Mantin, I. and Shamir, A. 2001. Weaknesses in the Key Scheduling Algorithm of RC4, *Selected Areas in Cryptography 2001, Lecture Notes in Computer Science*, Vol. 2259, pp. 1-24. Springer.
- [6] Furnell, S. 2005. Why users cannot use security, *Computers & Security*. Vol. 24, pp. 274-279.
- [7] Graham-Rowe, D., 22 January 2005. Wireless boom is hackers' heaven, *New Scientist*, <http://www.newscientist.com/article.ns?id=dn6894>
- [8] Jaques, R. 10 Mar 2005. UK firms haemorrhaging data to drive-by hackers: Unsecured Wi-Fi in one third of all wireless networks. <http://www.vnunet.com/vnunet/news/2126948/uk-firms-haemorrhaging-drive-hackers>
- [9] Leyden, J. 20 April 2004. Brits are crap at password security. [http://www.theregister.co.uk/2004/04/20/password\\_surveys/](http://www.theregister.co.uk/2004/04/20/password_surveys/)
- [10] Mead, N.R. and McGraw, G. 2003. Wireless Security's Future, *IEEE Security and Privacy*, 1 (4), pp. 68-72.
- [11] Office for National Statistics (ONS), 2002. Teleworking in the UK [http://www.statistics.gov.uk/articles/labour\\_market\\_trends/Teleworking\\_jun2002.pdf](http://www.statistics.gov.uk/articles/labour_market_trends/Teleworking_jun2002.pdf)
- [12] Office for National Statistics (ONS), 2006. Home-based working using communication technologies [http://www.statistics.gov.uk/articles/labour\\_market\\_trends/teleworking\\_Oct05.pdf](http://www.statistics.gov.uk/articles/labour_market_trends/teleworking_Oct05.pdf)
- [13] Office for National Statistics (ONS), 2007. Monthly, On-line edition <http://www.statistics.gov.uk/statbase/Product.asp?vlnk=8251>
- [14] Thomas, Z. 11Feb 2007. Best companies see surge in working from home <http://www.timesonline.co.uk/tol/news/uk/article1496840.ece>

# A Model and Framework for Online Security Benchmarking

Graeme Pye and Matthew J. Warren  
 School of Information Systems, Deakin University  
 Geelong, Victoria, Australia, 3217  
 E-mail: graeme@deakin.edu.au, mwarren@deakin.edu.au

**Keywords:** online, security, benchmarking.

**Received:** March 9, 2007

*The variety of threats and vulnerabilities within the online business environment are dynamic and thus constantly changing in how they impinge upon online functionality, compromise organizational or customer information, contravene security implementations and thereby undermine online customer confidence. To nullify such threats, online security management must become proactive, by reviewing and continuously improving online security to strengthen the enterprise's online security measures and policies, as modelled. The benchmarking process utilises a proposed benchmarking framework to guide both the development and application of security benchmarks created in the first instance, from recognized information technology (IT) and information security standards (ISS) and then their application to the online security measures and policies utilized within online business. Furthermore, the benchmarking framework incorporates a continuous improvement review process to address the relevance of benchmark development over time and the changes in threat focus.*

*Povzetek: Razvito je novo testno okolje za preizkušanje varnosti internetnega poslovanja.*

## 1 Introduction

Online security measures and policies are essential for protecting both the information of any online business and that of its customers. It is also imperative to maintaining an online business's competitive edge, building trust, customer confidence and enhancing the business reputation while maintaining and operating a secure online business environment (Standards Australia 2001).

A key finding of the 2004, 2005 and 2006 AusCERT surveys (AusCERT 2004, 2005, 2006) is that organizations are showing a preparedness to protect their IT systems across three areas: the use of information security policies, their practices and procedures; the use of information security standards or guides; and the number of organizations with qualified, experienced and trained personal. This indicates that Australian organizations are placing greater importance on managing the security of their information systems against latent online security threats and vulnerabilities. Similarly, the online components of Australian businesses can seek to deal with such security issues by applying the minimal best practice security recommendations outlined within the current Australian and New Zealand Information Security Management Standard, AS/NZS ISO/IEC 17799:2001 (Standards Australia 2001). Alternatively, they can apply the recommendations of various reports, guidelines, frameworks and security best practice publications that deliver advice on securing an online business implementation (NOIE 2002). Also in 2007, the Australian Federal Government allocated \$13.6 million

over four years in the national budget to improve e-security at a national level, to raise awareness of e-security issue for home users and small businesses including teaching e-security within schools (Coonan, H, 2007) and this will help to improve e-security awareness and in the long term security management.

Therefore, the authors propose that the development and application of online security benchmarks can provide both guidance and an assessment methodology for online security measures and policies. Furthermore, by incorporating a regime of continuous improvement, any online business can proactively strengthen security measures and policies through periodic revision. This paper establishes a dynamic benchmarking model applicable to the online business environment and proposes a managerial framework for establishing, reviewing and continuously improving online security benchmarks that also takes into consideration the passage of time, while still remaining applicative to Australian online business.

## 2 Benchmarking and Continuous Improvement

Benchmarking in traditional business models plays a major role in the ongoing assessment of business performance and return on investment. Similarly, benchmarking is also applicable to an online business or component thereof as a method to gauge performance and promote continuous improvement of online business processes (McGaughey 2002).



## 2.1 Benchmarking

Traditional business has utilized the systematic evaluation provided by benchmarking as a standard against which to compare and measure performance (Koch & Robertson 2002) and as an analysis tool focused on competitive performance factors such as costs, strategies and products within their competitive business domains. From such analyses an understanding can be gained of how the business compares with its peers and to what extent it deviates from the ‘norm’ or established benchmarks, over a given number of parameters (Codling 1996).

Therefore, benchmarking enables identification and targeting of business areas that are not meeting the established benchmark measures. Furthermore, by coupling the element of continuous improvement to the benchmarking process, sub-standard business areas become the focus of regular audits, assessment and ongoing monitoring through a periodic review process. Additionally, continuous improvement applied through revision of the benchmark itself incorporates the betterment of the set benchmark standard in an ongoing continuous manner too.

## 2.2 Continuous Improvement

Continuous improvement is the process of revising and improving upon previous assessment criteria to raise the level of functionality, improve efficiency and strengthen the assessment criteria with each application of the continuous improvement process (McGaughey 2002). Ideally, the continuous improvement process itself is an endless circular process that aims to establish higher goals with every iteration and reappraisal of the assessment criteria (Zajacek 2002). Benchmarking with continuous improvement will establish benchmarks that are continually reviewed, improved upon and strengthened in an ongoing manner. This proactive concept begins to address the dynamic environment of online business security, through the application of continuous improvement benchmarks for online security measures and policies.

## 3 A Benchmarking and Online Business Security Model

The concept of proactive online security benchmarking utilizes the continuous improvement principles of Total Quality Management alluded to by Saylor (1996). However, management of online security measures and policies has tended to be reactive and addressed only the perceived threats and vulnerabilities to the online business at the point in time of their implementation. Hence, online security generally remains static and is not reviewed, assessed or upgraded until after a detected security incident has run its course (Kolokotronis et al. 2002).

One way to address this reactive perception of security is through the utilization of continuous improvement benchmarking techniques that proactively

assess security measures and policies in an ongoing manner. This ensures that both the business online security measures and policies and the security benchmarks are continually improved, strengthened and remain up-to-date as time passes. By applying benchmarking that incorporates a systematic self-assessment regime, but vigilance is still required so that the application of ‘best practice’ online security does not lapse and become static once again and therefore incapable of addressing and meeting the changing nature of online threats and vulnerabilities.

Figure 1 depicts the authors’ perception of a traditional reactive online business security model and conceptually illustrates how the application of a proactive benchmarking model would exist and function to enhance online business security measures and policies.

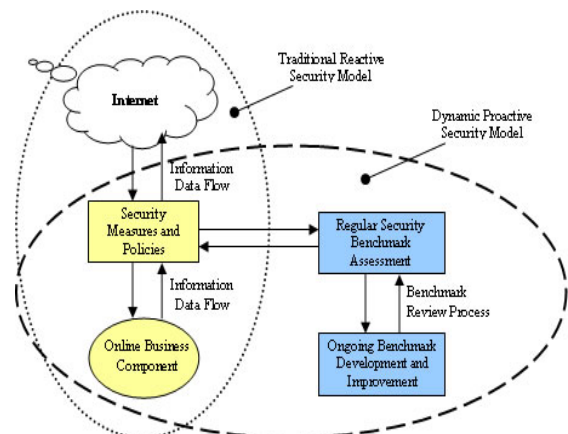


Figure 1: Online Security Benchmarking Model.

The essence of this proactive benchmarking process is the continual security benchmark review and improvement leading to the strengthening of online security measures and policies. Therefore, the results of this process will better reflect the business and customer security expectations for secure operation and continue to meet the economic expectations of the online business enterprise (McGaughey 2002). While this model (Figure 1) illustrates in general terms where online security benchmarking would be applied within the online business environment, there are a number of differing areas within an online business environment that need further consideration in terms of vulnerability types and the likely security criterion that should be benchmarked within these areas.

## 4 Online Business Security Criteria for Benchmarking

In general, online threats and vulnerabilities can fall within five broadly applicable criteria to online business security, namely Organizational Security, Infrastructure Security, Application Security, Network/System Security and User Management Security. Each of these security criteria addresses a specific area of online business security and incorporates security standards’ based recommendations that relate to the establishment and

development of security benchmarks applicable to online business security (Pye & Warren 2003).

#### 4.1 Organizational Security

The diffusion and complexity of technology within the online business environment demands that a reasonable level of security be implemented and maintained. This requires online organizations to develop and establish considered plans that organize and efficiently implement online security measures and policies via a functional and systematic security management plan (BSI 2003) and some of the key areas that should be benchmarked are identified as follows (Standards Australia 2001).

Organizational Security Management Policy Benchmarks ensure that an unambiguous security management policy is applied and readily available across the organization that clearly states the direction and goals for security management by outlining the organization's approach to online security.

Information Security Management Benchmarks relate to the management practices within an online business and ensures that a consistent organizational wide approach to secure management and storage of information within the possession of the business applies consistently at all levels of personnel.

Personnel Security Management Benchmarks ensure that adequate background checks are in place prior to the employment of new staff to address and minimize the risks of human error, theft, fraud and misuse of the organization's assets.

Security Incident Reporting Benchmarks measure the management and prompt activation of contingency plans that are paramount to minimizing damage from security incidents and malfunctions. It is imperative to report all security incidents as soon as practicable to the designated point of managerial contact irrespective of its magnitude to evoke a response action and ongoing incident monitoring for further subsequent analysis.

All staff and external contractors must be aware of the reporting procedures for incidents to minimize the potential impact on organizational assets of security breaches, threats, weaknesses or malfunctions.

#### 4.2 Infrastructure Security

Infrastructure security is a vital consideration to the overall security of buildings, offices and the equipment within the physical boundaries of the online business. Physical security controls should utilize physical barriers to protect assets from unauthorized access, damage, interference, and removal (Australian Standards 2001) and benchmarked as follows.

Physical Security Management Benchmarks ensure measures are in place to prevent unauthorized access, damage and interference to the premises by physically protecting critical equipment and sensitive business information within the building. Therefore, the emplacement of the online business processing hardware should be within a clearly defined secure area behind appropriate barriers and entry controls.

Equipment Security Management Benchmarks appraise equipment asset security to prevent loss, theft and damage and guard against the compromise of sensitive information and the protection of such equipment from physical threats and potential environmental hazards.

General and Media Management Benchmarks evaluates security controls to protect against disclosure, modification and theft by unauthorized persons by ensuring the controlled handling of computer media and its disposal is physically protected with procedures to protect documentation, computer storage media, data and system documentation from damage, theft and unauthorized access.

#### 4.3 Application Security Benchmarks

Software applications are integral to a functional online business and protective policies and measures should be in place to combat malicious software and establish appropriate electronic communications standard using encryption techniques to protect data in storage or during transmission across insecure public networks, benchmarks can address the following aspects (Standards Australian 2001).

Malicious Software Security Management Benchmarks measure the safeguards against the introduction of malicious software into the system, user education, security controls, and policies used to protect business information assets.

Electronic Mail Security Management Benchmarks ensure that the controls in place protect against loss, modification or misuse of information exchanged, email access should be controlled, monitored and compliant with the relevant legislation and user behaviour education should be adopted to reduce risk.

Encryption Management Benchmarks ensure that the cryptographic controls adopted to safeguard confidentiality, integrity or authenticity of information transmitted across public networks.

#### 4.4 Network/System Security

Computer networks and systems convey the communication and exchange of data for online business and it is imperative that security controls and policies are in place to regulate how such systems are utilized and accessed. These safeguards protect the information within internal and public networks as well as supporting the protection of the network infrastructure and suggested benchmarks are as follows (Standards Australia 2001).

Network/System Communication Control Benchmarks relating to the control of internal and external network communication and network services is necessary to ensure users who have access do not compromise the security measures and policies in place (Australian Standards 2001). Firewalls define the online boundary by providing communication control between internal networks and the publicly accessible external networks (BSI 2003).

System Security Management Benchmarks ensure planning is in place to minimize the risk of system failures to ensure system availability, adequate capacity, and enough resources exist for future expansion and growth of the online business (Standards Australia 2001).

Network/System Security Management Benchmarks focus on protecting the information and supporting infrastructure of the local network within the physical boundaries of the online business and includes all the activities and controls for securing the network (Standards Australia 2001).

System Use Monitoring Benchmarks measure the effectiveness of system monitoring in detecting unauthorized activities and records deviations from access control policy by logging system events to provide an audit trail and evidence in the event of security breaches or incidents caused by internal or external users (Standards Australia 2001)

#### 4.5 User Management Security

Validation and authentication of internal business staff and online customers can deliver a protective barrier to unauthorized access and this should cover the entire life-cycle of the user's access to services from new registrations to final deregistration by benchmarking the following aspects (Standards Australia 2001).

Password Management Benchmarks measure the strength and enforces regular changing of passwords for validating the identity of the user prior to the granting of access to a particular online service or system (Standards Australia 2001).

Authentication Management Benchmarks ensures that authentication mechanisms for online business systems and applications identify the users uniquely and authenticate prior to permitting further interaction between the online business system and the user.

The pursuit of business activities and transactions across public networks involving the exchange of sensitive data exposes the online business to potentially damaging threats and vulnerabilities that may result in fraudulent activity, contract disputes and the disclosure or modification of sensitive information.

Therefore, the application of regular benchmarking of online business security measures and policies will provide the online business with the ability to review their security status and continue to strengthen and update online security measures and policies. Pursuant to this view however is the need for a framework in which to develop these security benchmarks.

### 5 Online Security Benchmark Framework

Before developing meaningful online security benchmarks, it is necessary to ascertain a starting point to devise the essential benchmarking elements necessary to assess online business security. An internationally recognized published standard is an obvious starting point for developing online security benchmarks, such as the Australian and New Zealand Standard AS/NZS

ISO/IEC 17799:2001 (2001). In utilizing the Australian and New Zealand Standard (2001) as the minimal benchmarks for online security, it is also important to have benchmarks that indicate improved online security goals. One such security standard that sets a higher security baseline than the Australian and New Zealand Standard is the German IT Baseline Manual (BSI 2003).

The premise for the German IT Baseline Manual (BSI 2003) being at a higher baseline standard than the Australian and New Zealand Standard is supported by a comparative evaluation of information security, baseline standards undertaken by Brooks and Warren (2001). This research concluded that the information within the Australian and New Zealand Standard (2001) focuses on enhancing information security awareness and authorization security at the minimal 'best practice' level. While the German IT Baseline Protection Manual (BSI 2003), documented baseline security features that were of a higher minimal standard, although its focus is technically orientated towards the implementation of security controls needed to secure IT systems. Through applying the Australian and New Zealand Standard (2001) as the minimum benchmark threshold and the German IT Baseline Protection Manual (BSI 2003) as a reasonable benchmark to aim for, this premise can then be applied to benchmarking online security to measure the current status of an online business's security criteria. Furthermore, a framework can standardize benchmark development and deliver consistent and methodical application of such security benchmarks.

#### 5.1 The Online Security Benchmark Framework

Table 1 illustrates the online security benchmark framework that incorporates the Australian and New Zealand Standard (2001) information as the initial minimal security requirement benchmark and similarly applies the recommended German IT Baseline Protection Manual (BSI 2003) information as an improved security benchmark that an online business should endeavor to achieve.

1	Initial Security Benchmark	Minimum Benchmark: Australian and New Zealand Standard (2001). Maximum Benchmark: German IT Baseline Protection Manual (2003).
2	Online Security Assessment	Online Security Assessment Against Applicable Benchmark: Pass/Fail.
3	Current Benchmark Analysis	Analysis of Current Online Security Benchmark: Pass/Fail
4	Continuous Improvement Analysis	Future Security Benchmark Development and Implementation.

Table 1: Online Security Benchmark Framework.

The framework shown in Table 1, consists of a four stage process that can be applied as a circular development process to initially create a security

benchmark and then continue to improve upon such benchmarks developed, this with the following assessment methodology can then be applied to benchmark or determine the security status of online business policies and measures, while further guiding the proactive development and implementation of continuously improved benchmarks for online security.

## 5.2 Methodology for Applying the Security Benchmark Framework

The methodical application of the online security benchmark framework within an online business is essential to correctly apply the benchmarking concept to online security policies and measures. The following outlines the methodology to apply when benchmarking online business security.

Initially, it should be determined within the security criteria of the online business, which security measure or policy is applicable to which particular benchmark, this ensures consistency of assessment and sets a minimum and maximum security benchmark.

Next, an assessment of the particular online business security measure or policy in comparison to the minimum security benchmark listed determines if it meets the minimum benchmark standard.

Then an assessment analysis on the current benchmark establishes its effectiveness and assists the development of an improved security benchmark as the new minimum benchmark thus strengthening the security benchmarking criteria with a higher-level benchmark.

The final step in the methodology promotes proactive benchmark development and ongoing continuous improvement research. This is to determine new and improved security benchmarks that exceed the current minimum security benchmark and to potentially define a new maximum benchmark as recorded in the framework, with the intention of becoming the future minimum benchmark.

However, there is a need to be mindful of the initial level of security advice provided by the relevant information security standard, as this is the underpinning foundation of the security benchmarking process and the initial security benchmark development starting point, irrespective of the current online security environment.

## 6 Information Security Standards (ISS)

Therefore, as the initial online security benchmarks are set, based on the relative Australian Standard New Zealand Standard (2001) and German IT Baseline Protection Manual (2003), there remains the point that the intention of such technology ISS is what do they actually deliver. The intention of the ISS is to deliver comprehensible and precise meanings to describe various technological actions as is related to IT security in a consistent, unambiguous and comprehensive manner. The Standard itself has to be flexible enough to allow for innovation and reconcile user issues while still delivering guidance for resolving both technical and political

problems. Therefore, Standards are particularly useful as a starting point for solving particular issues and are very appropriate where many systems function in a similar manner. However, while Standards remain regarded as technical document devices for the achievement of specific ends and essential to progress, they can also be a hindrance to innovative creativity and can encourage mediocrity due to their very convenience as a required minimum measurement (Libicki, 1995).

Therefore, it is with this in mind that the authors have developed the Online Security Benchmark Framework as a means to initially create a specific online security benchmark and to encourage continuous improvement by employing a circular process whereby the security benchmark is continually improved. Continuous improvement is imperative to proactive security benchmarking development to avoid, falling into the trap of mediocrity whereby the security benchmark is set once using the relevant security Standard and never improved upon or reassessed until necessary.

Furthermore, the aim of the authors is to encourage continuous improvement as a means of negating the passage of time and ensuring that the security benchmarks established by an organization will continue to measure the effectiveness and strengthen security measures and policies applicable to protecting the online security of an organization.

## 7 The ISS Verses Security Benchmarking Over Time

An often ill-considered aspect of developing and implementing online security measures and organizational security policies is the consequence of time. When considering that every person, society and environment occupies a point in time that is dynamically changing as time passes, all too often security measures and policies are created and implemented initially but thereafter remain unchanged as time passes. The authors submit that a lack of appreciation of the perception of time passing is a security weakness, but not from the perspective of security where a 'set and forget' static approach is generally applied.

From a philosophical aspect the 'perception of time' expression exists as one of three states: past, present and future and these states are perceived by changes or events in time and what is perceived as the present and what is going on right now. Furthermore, perceptions of past, present and future are important for social enquiry and action as they draw on past events that influence the present, but may not determine the future specifically although may enable a perceived range of likely futures (Le Poidevin 2004). Therefore, by incorporating an appreciation of the perception of time in the benchmark development process, the authors consider that online security can indeed utilize past and present perceptions of time related to the changes in the online security environment to promote continuous improvement. Thus as applied in the online security benchmarking framework (see Table 1), the application of a continuous

improvement process can deliver continually improved upon online security benchmarks into the future.

The advantage of adopting a continuous improvement benchmarking process is that over time the security benchmarks will be strengthened through regular review and updating. Whereas the ISSs' only deliver a static security reference that is applicable at a specific historical point in time, which is valuable in providing a starting point for the establishment of the initial online security benchmark. However, over time the ISS only represents a historical reflection of security and is very slow to update in comparison to the proposed benchmarking framework for online security. Therefore, the online security benchmarking framework offers an improved method of proactive online security maintenance of policies and measures that continues to take into consideration, not only the historical security aspect, but also adopt changes relevant to the present situation that can or may be incorporated and applied in the future. By utilizing the perception of time in the security benchmarking development process, this is addressing the dynamic nature of the online environment and potentially insulates the online components and protects the information of the organization in a way that is proactive to the online environment and continues to strengthen security measures and policies.

## 8 Conclusion

The benchmarking model and framework illustrated here for online security measures and policies, is designed to deliver guidance, manageability and consistency to the development, ongoing protection and improvement to the security features of an online business. Thus enabling a business to develop applicable security benchmarks, determine their current online security status and implement a continuous improvement plan to improve and strengthen their online security measures and policies. The technology ISS offers a starting point for development of online security benchmarks only and fails to address the ongoing changes and developments that occur in the online environment that can adversely impinge upon the secure function of an organization's online business component.

The practical application of this research would be beneficial to raising the awareness of security issues and policies within an online business and would perhaps encourage a culture of security awareness within any organization. This may be a consequence of the benchmarking regime as it lends itself to the management, monitoring and continuous improvement of online security measures and policies protecting the information within a business's possession, but this remains speculative. An indicative example of an application of the benchmarking methodology would be the benchmarking the security of an electronic supply chain, where a number of individual co-operating supply chain members would be able to apply the online security benchmarking framework and methodology proposed here, to ensure that all members of the electronic supply

chain meet and continue to improve their security status, as proposed by Pye et al (2005).

Additionally, further research is still required to assess the effectiveness, value and cost that the application of the security benchmarking techniques alluded to in this paper would impose on the business itself and whether these benchmarking techniques deliver practical assessments of the online security status of any business. Additionally, further research is required to determine if the continuous benchmarking techniques, as outlined in the framework (see Table 1) would prove to be advantageous to the assessment, application and monitoring of IT governance requirements of any business that utilizes information systems as an underpinning support to their online business, business processes and ambitions.

## 9 References

- [1] AusCERT (2004) Australian Computer Crime and Security Survey, AusCERT. URL: <<http://www.auscert.org.au/render.html?it=2001>> Accessed: May 2004.
- [2] AusCERT (2005) Australian Computer Crime and Security Survey, AusCERT. URL: <<http://www.auscert.org.au/images/ACCSS2005.pdf>> Accessed: May 2005.
- [3] AusCERT (2006) Australian Computer Crime and Security Survey, AusCERT. URL: <<http://www.auscert.org.au/images/ACCSS2006.pdf>> Accessed: May 2006.
- [4] Brooks W., Warren M.J. (2001) A Security Evaluation Criteria for Baseline Security Standards. Technical Report TR C 01/18, Deakin University, Geelong.
- [5] BSI (2003) IT Baseline Protection Manual. Federal Agency for Security in Information Technology. Bundesamt für Sicherheit in der Informationstechnik (Multimedia CD-ROM).
- [6] Codling S. (1996) *Best Practice Benchmarking*. Gulf Publishing Company, Houston, Texas.
- [7] Coonan H. (2007) Improving e-security for home users and small business – Media Release, May, 2007, Accessed 11<sup>th</sup> May 2007, URL: [http://www.minister.dcita.gov.au/media/media\\_releases/improving\\_e-security\\_for\\_home\\_users\\_and\\_small\\_business](http://www.minister.dcita.gov.au/media/media_releases/improving_e-security_for_home_users_and_small_business).
- [8] Koch H., Robinson P.E. (2002) Evaluating Electronic Commerce Initiatives with Benchmarks: Insights from Three Case Studies, *Eighth Americas Conference on Information Systems*. pp.1251-1258.
- [9] Kolokotronis N., Margaritis C., Papadopoulou P., (2001) An integrated approach for securing electronic transactions over the web, *Benchmarking: An International Journal* Vol: 9 (2): pp.166–181.
- [10] Le Poidevin R. (2004) The Experience and Perception of Time. The Stanford Encyclopedia of Philosophy, URL: <<http://www.plato.stanford.edu/archives/win2004/entries/time-experience>> Accessed: May 2006.

- [11] Libicki M.C. (1995) *Information Technology Standards. The Quest for the Common Byte*, Digital Press, Newtown, MA.
- [12] McGaughey R.E., (2002) Benchmarking business-to-business electronic commerce, *Benchmarking: An International Journal* **Vol. 9** (5): pp.471-484.
- [13] NOIE (2002) trusting the internet. small business guide to e-security, NOIE, URL: <[http://www.noie.gov.au/publications/NOIE/trust/trusting\\_the\\_internet.pdf](http://www.noie.gov.au/publications/NOIE/trust/trusting_the_internet.pdf)> Accessed: December 2002.
- [14] Pye G., Pierce J.D., Warren M.J., Mackay D.R. (2005) Supply Chain Security: The Need for Continuous Assessment, *Supply Chain Practice* **Vol. 7** (1): pp.4–16.
- [15] Pye G., Warren M.J. (2003) Development of I.T. Evaluation Criteria for Common E-business Security Issues. Technical Report TR C 03/12, Deakin University, Geelong.
- [16] Saylor J.H. (1996) *TQM Simplified. A Practical Guide*, 2<sup>nd</sup> Ed. McGraw-Hill New York.
- [17] Schneider G.P., Perry J.T. (2001) *Electronic Commerce*, 2<sup>nd</sup> Ed., Course Technology.
- [18] Standards Australia (2001) *Information Technology - Code of practice for information security management. AS/NZS ISO/IEC 17799:2001*, Standards Australia, Barton.
- [19] Zajacek M. (2002) Continuous Development Process. URL: <<http://www.unimelb.edu.au/development/wag/index.html>> Accessed: July 2003.





# An Approach to Extracting Interschema Properties from XML Schemas at Various “Severity” Levels

Pasquale De Meo, Giovanni Quattrone and Domenico Ursino  
 Università Mediterranea di Reggio Calabria  
 Via Graziella, Località Feo di Vito  
 89122 Reggio Calabria, Italy  
 E-mail: demeo@unirc.it, quattrone@unirc.it, ursino@unirc.it

Giorgio Terracina  
 Dipartimento di Matematica  
 Università della Calabria  
 Via Pietro Bucci  
 87036 Rende (CS), Italy  
 E-mail: terracina@mat.unical.it

**Keywords:** XML schemas, synonymies, homonymies, hyponymies, overlappings, interschema property extraction

**Received:** June 7, 2006

*This paper presents an approach for the semi-automatic, uniform extraction of synonymies, hyponymies, overlappings and homonymies holding among concepts of different XML Schemas. The proposed approach is specialized for XML, is almost automatic and “light”. As a further, original, peculiarity, it is parametric w.r.t. a “severity level” against which the extraction task is performed. First the paper presents an overview of the interschema property extraction approaches already presented in the past, as well as a set of criteria for classifying this kind of approaches. After this, it describes the proposed approach in all details, illustrates various theoretical results, presents the experiments we have performed for testing it and compares it with the interschema property extraction approaches previously proposed in the literature.*

*Povzetek: Opisan je polavtomatski postopek za ekstrakcijo sinonimov iz XML shem.*

## 1 Introduction

The Web is becoming the reference infrastructure for most of the applications conceived to handle the interoperability among partners. As a matter of fact, it is presently playing a key role for both the publication and the exchange of information among organizations. In order to make Web activities easier, the World Wide Web Consortium proposed XML (eXtensible Markup Language) for unifying representation capabilities, typical of HTML, and data management features, typical of classical DBMSs.

The exploitation of XML is crucial for improving the interoperability of Web partners; as a matter of fact, this language provides a uniform format for exchanging data among them. However, XML usage alone is not enough for guaranteeing such a cooperation. In fact, the heterogeneity of data exchanged over the Web regards not only their formats but also their semantics. The use of XML allows format heterogeneity to be faced; the exploitation of XML Schemas allows the definition of a reference context for exchanged data and is a first step for handling semantic diversities; however, in order to completely and satisfactorily manage these last, the knowledge of interschema properties (see Section 2.1), possibly holding among concepts

belonging to different sources, is necessary.

The most common interschema properties previously considered in the literature are synonymies and homonymies. A *synonymy* between two concepts indicates that they have the same meaning. An *homonymy* between two concepts denotes that they refer to different meanings, yet having the same name. In the past some approaches have been also proposed for deriving other interschema properties, e.g., hyponymies and overlappings. A concept  $C_1$  is a *hyponym* of a concept  $C_2$  (that is, in its turn, a *hyponym* of  $C_1$ ) if  $C_1$  has a more specific meaning than  $C_2$ . As an example, “PhD Student” is a hyponym of “Student”. An *Overlapping* holds between two concepts if they are not synonymous but share a significant set of properties.

For a more detailed survey about the semantic relationships possibly occurring between two concepts the reader is referred to [24]. In this paper semantic relationships are defined and classified according to different perspectives and disciplines, such as linguistics, logics and cognitive psychology. From a comparison between the definitions of [24] and those introduced in this paper we can observe that: (i) our definition of synonymy exactly matches the definition of synonymy provided in [24]; (ii) our concept of homonymy can be regarded as a special case of the con-

cept of antinomy specified in [24]; specifically, in that paper, an antinomy exists between two terms if they denote opposite (or, at least, different) concepts; our definition of homonymy, instead, requires that two terms indicate different concepts and, in addition, that they share the same name; (iii) our hyponymy property corresponds to the inclusion relationship specified in [24]; (iv) our overlapping property is similar to some kinds of meronymic relationship introduced in [24] (these last indicate that a part of a concept  $A$  is somehow related to a part of a concept  $B$ ).

Owing to the enormous increase of the number of available information sources, all the approaches for interschema property extraction currently proposed in the literature are *semi-automatic*; specifically, they require the human intervention only during a pre-processing phase and for the validation of obtained results. The rapid development of the Web leads each interschema property extraction approach to operate on a great number of sources; this requires a further effort for conceiving approaches with less manual intervention.

Since the possible interschema properties to consider are numerous and various, the capability of *uniformly* deriving distinct properties appears to be a crucial feature for a new interschema property derivation approach. As a matter of fact, different strategies for extracting distinct interschema properties could lead to different interpretations of the same reality; this is a situation that must be avoided.

Finally, the large number of currently available information sources makes it evident the necessity that an interschema property derivation approach should be “light”, i.e., it should minimize the exploitation of thresholds and/or weights whose tuning requires a lot of efforts.

This paper provides a contribution in this setting and proposes an approach for uniformly extracting synonymies, hyponymies, overlappings and homonymies from a set of XML Schemas.

Our approach satisfies all the desiderata mentioned above. In fact, (i) *it is almost automatic*; specifically, it requires the user intervention only in few specific cases. (ii) *it is “light”*; specifically, it does not exploit thresholds or weights; as a consequence, it does not need a tuning activity. However, in spite of this “lightness”, obtained results are precise and satisfactory, as shown in Section 5. (iii) *it allows the derivation of the various interschema properties within a uniform framework*; such a framework consists of a set of maximum weight matchings computed on suitable bipartite graphs. (iv) *it is specific for XML*; in fact, the framework underlying our approach has been defined for directly covering the XML specificities (see, below, Section 3). (v) *it allows the choice of the “severity level” against which the property extraction task is performed*; such a feature derives from the consideration that applications and scenarios possibly benefiting of derived interschema properties are numerous and extremely various. In some situations the extraction process must be very severe in that it can state the existence of an interschema property between two concepts only if this fact is confirmed

by various clues. In other situations, the extraction task can be looser and can assume the existence of an interschema property between two concepts if it has been derived by some computation, without requiring various confirmations. At the beginning of the extraction activity our approach asks the user to specify the desired severity level; this is the only information required to him until the end of the extraction task, when he has to validate obtained results.

It is worth pointing out that, in the past, we have proposed some algorithms for deriving synonymies and homonymies specifically conceived to operate on XML Schemas [5]. They do not exploit thresholds and weights and consider a “severity” level; as a consequence, they follow the same philosophy as the approach we are presenting here; however, they are not able to derive hyponymies and overlappings. In this context the approach presented here can be considered an advancement of this research line and provides a further component allowing the construction of a framework for uniformly deriving a large variety of interschema properties among a great number of XML Schemas.

## 2 Background

### 2.1 An overview of the interschema property extraction approaches

In [9] the system CGLUE is proposed. It exploits machine learning techniques for deriving *semantic matchings* between two given ontologies  $O_1$  and  $O_2$ . In [11] the authors propose a formal method, based on fuzzy relations, capable of performing the semantic reconciliation of heterogeneous data sources.

In [17] the authors propose Cupid, a system that detects semantic matchings holding between two schemas. First, Cupid represents input schemas by means of trees. Then it computes a coefficient, named *linguistic similarity* for each pair of schema elements. After this Cupid computes the *structural similarity* coefficient by means of a suitable tree-based algorithm. Finally it combines linguistic and structural similarity coefficients to derive semantic matchings.

In [1] the authors describe MOMIS, a system devoted to handle both integration and querying activities on heterogeneous data sources. MOMIS follows a “semantic approach” to interschema property extraction, based on an intensional study of information sources.

In [14] a statistical framework for performing schema matching tasks on Web query interfaces (i.e., on data sources containing the results of the execution of queries posed through Web interfaces) is proposed. In their approach, the authors hypothesize the presence, for each application context, of a “hidden schema model” which acts as a unified generative model describing how schemas are derived from a finite vocabulary of attributes.

In [4] the authors propose a matching algorithm for measuring the structural similarity between an XML document

$D$  and a DTD  $T$ . This algorithm assigns a score (called *similarity measure*) to  $D$ , indicating how much  $D$  is similar to  $T$ . The approach represents both  $D$  and  $T$  as labelled trees. In [20, 21] the system DIKE is proposed. This system is devoted to extract interschema properties from E/R Schemas. DIKE has been conceived to operate with quite a small number of information sources; as a consequence, it privileges accuracy to computation time. This system exploits a support dictionary containing an initial set of (generally lexical) similarities constructed with the support of human experts during a training phase. The extraction task is graph-based and takes into account the “context” of the concepts into examination; it exploits a large variety of thresholds and weights in order to better adapt itself to the sources which it currently operates on; these thresholds and these weights must be tuned during the training phase.

## 2.2 Classification criteria

In the literature various classification criteria have been proposed for comparing schema matching approaches (see, for example, [23]). They allow the approaches to be examined from various points of view. Specifically, the criteria appearing particularly interesting in our context are the following:

**Schema Types.** Some matching algorithms can operate only on a specific kind of data sources (e.g., XML, relational, and so on); these approaches are called *specific* in the following. On the contrary, other approaches are able to manage every kind of data sources; we call these approaches *generic* in the following. A generic approach is usually more versatile than a specific one because it can be applied on data sources characterized by heterogeneous representation formats. On the contrary, a specific approach can take advantage of the peculiarities of the corresponding data model.

**Instance-Based versus Schema-Based.** In order to detect interschema properties, schema matching approaches can consider data instances (i.e., the so-called *extensional information*) or schema-level information (i.e., the so-called *intensional information*). The former class of approaches is called instance-based; the latter one is known as schema-based. An intermediate category is represented by *mixed approaches*, i.e., those ones exploiting both intensional and extensional information. Instance-based approaches are generally very precise because they look at the actual content of the involved sources; however, they are quite expensive since they must examine the extensional component of the involved sources. On the contrary, schema-based approaches look at the intensional information only and, consequently, they are less expensive; however, they could be less precise. Finally, the results of an instance-based approach are valid only for the sources it has been applied to, whereas the results of a schema-based approach are valid for all those sources conforming to the considered schemas. As a consequence, instance-based and mixed approaches are more suited for those application contexts characterized

by few sources and requiring very accurate results, whereas schema-based approaches are more suited for those application contexts involving a great number of sources.

**Individual versus Combinatorial.** An *individual* matcher exploits just one matching criterion; on the contrary, *combinatorial* approaches integrate different individual matchers to perform schema matching activities. Combinatorial matchers can be further classified as: (i) *hybrid matchers*, if they directly combine several schema matching approaches into a unique matcher; (ii) *composite matchers*, if they combine the results of several independently executed matchers; they are sometimes called multi-strategy approaches. The individual matchers are simpler and, consequently, less time consuming than the combinatorial ones; however, the results they obtain are generally less accurate than those returned by combinatorial matchers.

**Matching Cardinality.** Some approaches have been conceived to derive only semantic similarities between two single components of different schemas (*1:1 matchings*). Other approaches are capable of deriving also semantic similarities between one single component of a schema and a group of components of the other schemas (*1:n matchings*) or between two groups of components of different schemas (*m:n matchings*).

**Exploitation of Auxiliary Information.** Some approaches could exploit auxiliary information (e.g., dictionaries, thesauruses, and so on) for their activity; on the contrary, this information is not needed in other approaches. Auxiliary information represents an effective way to enrich the knowledge that an approach can exploit. However, in order to maintain its effectiveness, the time required to compile and/or retrieve it must be negligible w.r.t. the time required by the whole approach to perform its matches. For this reason, pre-built or automatically computed auxiliary information would be preferred to the manually provided one.

## 3 Preliminaries

### 3.1 Neighborhood definition

We start to illustrate the definition of the neighborhood of an element or an attribute in an XML Schema by introducing the concept of  $x$ -component, that allows both elements and attributes of an XML Schema to be uniformly handled.

**Definition 3.1** Let  $S$  be an XML Schema; an  $x$ -component of  $S$  is either an element or an attribute of  $S$ .  $\square$

An  $x$ -component is characterized by its name, its typology (indicating if it is either a complex element or a simple element or an attribute) and its data type.

**Definition 3.2** Let  $S$  be an XML Schema; the sets of its  $x$ -components, its keys and its keyrefs are denoted as  $XCompSet(S)$ ,  $KeySet(S)$  and  $KeyrefSet(S)$ , respectively. The union of  $XCompSet(S)$ ,  $KeySet(S)$  and  $KeyrefSet(S)$  is denoted as  $ConstructSet(S)$ ; it forms the set of constructs of  $S$ .  $\square$

We now introduce some functions that allow the strength of the relationship existing between two x-components  $x_S$  and  $x_T$  of an XML Schema  $S$  to be determined. These functions are:

- *veryclose*( $x_S, x_T$ ), that returns *true* if and only if: (i)  $x_T = x_S$ , or (ii)  $x_T$  is an attribute of  $x_S$ , or (iii)  $x_T$  is a simple sub-element of  $x_S$ ; in all the other cases it returns *false*;
- *close*( $x_S, x_T$ ), that returns *true* if and only if: (i)  $x_T$  is a complex sub-element of  $x_S$ , or (ii)  $x_S$  and  $x_T$  are two complex elements of  $S$  and there exists a `keyref` element stating that an attribute of  $x_S$  refers to a `key` attribute of  $x_T$ ; in all the other cases it returns *false*;
- *near*( $x_S, x_T$ ), that returns *true* if and only if either *veryclose*( $x_S, x_T$ ) = *true* or *close*( $x_S, x_T$ ) = *true*; in all the other cases it returns *false*;
- *reachable*( $x_S, x_T$ ), that returns *true* if and only if there exists a sequence of x-components  $x_1, x_2, \dots, x_n$  such that  $x_S = x_1, near(x_1, x_2) = near(x_2, x_3) = \dots = near(x_{n-1}, x_n) = true, x_n = x_T$ ; in all the other cases it returns *false*.

We are now able to introduce the concept of Connection Cost from an x-component  $x_S$  to an x-component  $x_T$ . It is a measure of the correlation degree existing between  $x_S$  and  $x_T$  and indicates how much the concept expressed by  $x_T$  is “close” to the concept represented by  $x_S$ .

**Definition 3.3** Let  $S$  be an XML Schema and let  $x_S$  and  $x_T$  be two x-components of  $S$ . The *Connection Cost* from  $x_S$  to  $x_T$ , denoted by  $CC(x_S, x_T)$ , is defined as: (i) 0 if *veryclose*( $x_S, x_T$ ) = *true*; (ii) 1 if *close*( $x_S, x_T$ ) = *true*; (iii)  $\mathcal{C}_{ST}$  if *reachable*( $x_S, x_T$ ) = *true* and *near*( $x_S, x_T$ ) = *false*; (iv)  $\infty$  if *reachable*( $x_S, x_T$ ) = *false*.

Here  $\mathcal{C}_{ST} = \min_{x_A} (CC(x_S, x_A) + CC(x_A, x_T))$  for each  $x_A$  such that *reachable*( $x_S, x_A$ ) = *reachable*( $x_A, x_T$ ) = *true*.  $\square$

We are now provided with all tools necessary to define the concept of neighborhood of an x-component.

**Definition 3.4** Let  $S$  be an XML Schema, let  $x_S$  be an x-component of  $S$  and let  $j$  be a non-negative integer. The  $j^{th}$  neighborhood of  $x_S$  is defined as:  $nbh(x_S, j) = \{x_T \mid x_T \in XCompSet(S), CC(x_S, x_T) \leq j\}$   $\square$

The next proposition provides an estimation of the maximum number of distinct neighborhoods for an x-component; the interested reader can find its proof in the Appendix available at the address <http://www.mat.unical.it/terracina/informatica07/Appendix.pdf>.

**Proposition 3.1** Let  $S$  be an XML Schema; let  $x_S$  be an x-component of  $S$ ; let  $m$  be the number of complex elements of  $S$ ; then  $nbh(x_S, j) = nbh(x_S, m - 1)$  for each  $j$  such that  $j \geq m$ .  $\square$

The next proposition determines the worst case time complexity for computing all neighborhoods of all x-components of an XML Schema  $S$ . The interested reader can find its proof in the Appendix.

**Proposition 3.2** Let  $S$  be an XML Schema and let  $n$  be the number of its x-components. The worst case time complexity for computing all neighborhoods of all x-components of  $S$  is  $O(n^3)$ .  $\square$

### 3.2 Neighborhood comparison

Given two x-components  $x_{1_j}$  and  $x_{2_k}$  and two corresponding neighborhoods  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$ , there could exist different relationships between them.

Specifically, three possible relationships, namely *similarity*, *comparability* and *generalization*, could be taken into account. All of them are derived by computing suitable objective functions on the maximum weight matching associated with a bipartite graph obtained from the x-components of  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$ .

In the following we indicate by  $BG(x_{1_j}, x_{2_k}, v) = \langle NSet(x_{1_j}, x_{2_k}, v), ESet(x_{1_j}, x_{2_k}, v) \rangle$  the bipartite graph associated with  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$ ; when it is not confusing, we shall use the notation  $BG(v)$  instead of  $BG(x_{1_j}, x_{2_k}, v)$ . In  $BG(v)$ ,  $NSet(v) = PSet(v) \cup QSet(v)$  represents the set of nodes; there is a node in  $PSet(v)$  (resp.,  $QSet(v)$ ) for each x-component of  $nbh(x_{1_j}, v)$  (resp.,  $nbh(x_{2_k}, v)$ ).  $ESet(v)$  is the set of edges; there is an edge between  $p \in PSet(v)$  and  $q \in QSet(v)$  if: (i) a synonymy between the names of the x-components  $x_p$  and  $x_q$ , associated with  $p$  and  $q$ , is stored in the reference thesaurus; (ii) the cardinalities of  $x_p$  and  $x_q$  are *compatible*; (iii) their data types are *compatible* (this last condition must be verified only if  $x_p$  and  $x_q$  are attributes or simple elements).

Here, the cardinalities of two x-components are considered compatible if the intersection of the intervals they represent is not empty. The motivation underlying this assumption is that cardinalities represent constraints associated with the involved concepts and, therefore, contribute to define their semantics; as a consequence, completely disjoint intervals are a symptom that the two concepts have different semantics. Compatibility rules associated with data types are analogous to the corresponding ones valid for high level programming languages.

The maximum weight matching for  $BG(v)$  is the set  $ESet'(v) \subseteq ESet(v)$  of edges such that, for each node  $x \in PSet(v) \cup QSet(v)$ , there is *at most one* edge of  $ESet'(v)$  incident onto  $x$  and  $|ESet'(v)|$  is maximum (for algorithms solving the maximum weight matching problem, see [12]).

As previously pointed out, in our approach, all neighborhood comparisons are performed by computing the maximum weight matching on a suitable bipartite graph. The reasoning underlying this choice is the following: all types of neighborhood comparison (i.e., similarity, comparability and generalization) aim to determine how much two neighborhoods are somehow close. A neighborhood is a set of x-components. Generally speaking, two sets are close if they share a sufficiently large number of their elements. In our application context, two x-components belonging to two different neighborhoods can be considered as a shared x-component if a synonymy exists between them. Then, the maximum weight matching on the bipartite graph constructed from two neighborhoods allows the maximum number of pairs of synonymous x-components belonging to the neighborhoods to be determined; as a consequence, it allows the derivation of the maximum number of x-components that can be considered shared between the two neighborhoods and, therefore, the computation of the closeness degree of the two neighborhoods at hand.

### 3.2.1 Neighborhood similarity

Intuitively, two neighborhoods (and, more in general, two sets of objects) are considered similar if most of their components are similar.

In order to determine if  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$  are similar, we construct  $BG(x_{1_j}, x_{2_k}, v)$  and, then, compute the objective function  $\phi_{BG}(v) = \frac{2|ESet'(v)|}{|PSet(v)| + |QSet(v)|}$ .

Here  $|ESet'(v)|$  represents the number of matches associated with  $BG(v)$ , as well as the number of pairs of x-components  $\langle x_{1_j}^p, x_{2_k}^q \rangle$  such that  $x_{1_j}^p \in nbh(x_{1_j}, v)$ ,  $x_{2_k}^q \in nbh(x_{2_k}, v)$  and a synonymy between the names of  $x_{1_j}^p$  and  $x_{2_k}^q$  is stored in the reference thesaurus.  $|PSet(v)| + |QSet(v)|$  denotes the total number of nodes in  $BG(v)$ , as well as the total number of x-components associated with  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$ . The coefficient 2 at the numerator of  $\phi_{BG}$  is necessary to make the numerator and the denominator comparable; in fact,  $|PSet(v)| + |QSet(v)|$  refers to x-components whereas  $|ESet'(v)|$  regards pairs of x-components. Finally,  $\phi_{BG}(v)$  represents the share of matching nodes in  $BG(v)$ , as well as the share of similar x-components present in  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$ . The formal definition of the neighborhood similarity is given below.

**Definition 3.5** Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $x_{1_j}$  (resp.,  $x_{2_k}$ ) be an x-component of  $S_1$  (resp.,  $S_2$ ). Two neighborhoods  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$  are similar if, given the bipartite graph  $BG(x_{1_j}, x_{2_k}, v)$ ,  $\phi_{BG}(v) > \frac{1}{2}$ . □

This definition assumes that  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$  are similar if  $\phi_{BG}(v) > \frac{1}{2}$ ; such an assumption derives from the consideration that two sets of objects can be considered similar if the number of similar components is greater than the number of the dissimilar ones or, in other words, if the number of similar

components is greater than half of the total number of components.

The following theorem states the worst case time complexity for determining if two neighborhoods are similar. Its proof is provided in the Appendix.

**Theorem 3.1** Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $x_{1_j}$  (resp.,  $x_{2_k}$ ) be an x-component of  $S_1$  (resp.,  $S_2$ ). Let  $p$  be the maximum between  $|nbh(x_{1_j}, v)|$  and  $|nbh(x_{2_k}, v)|$ . The worst case time complexity for determining if  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$  are similar is  $O(p^3)$ . □

### 3.2.2 Neighborhood comparability

Intuitively, two neighborhoods  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$  are comparable if there exist at least two (quite large) subsets  $XSet_j$  of  $nbh(x_{1_j}, v)$  and  $XSet_k$  of  $nbh(x_{2_k}, v)$  that are similar. Similarity between  $XSet_j$  and  $XSet_k$  is computed by constructing a bipartite graph  $BG(XSet_j, XSet_k)$  starting from the x-components of  $XSet_j$  and  $XSet_k$ , and by computing  $\phi_{BG}$  in a way analogous to that we have seen in Section 3.2.1. Comparability is a weaker property than similarity. As a matter of fact, if two neighborhoods are similar, they are also comparable. However, it may be that two neighborhoods are not similar but are comparable because they have quite large similar subsets. The formal definition of neighborhood comparability is provided below.

**Definition 3.6** Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $x_{1_j}$  (resp.,  $x_{2_k}$ ) be an x-component of  $S_1$  (resp.,  $S_2$ ). Two neighborhoods  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$  are comparable if there exist two subsets,  $XSet_j$  of  $nbh(x_{1_j}, v)$  and  $XSet_k$  of  $nbh(x_{2_k}, v)$ , such that: (i)  $|XSet_j| > \frac{1}{2}|nbh(x_{1_j}, v)|$ ; (ii)  $|XSet_k| > \frac{1}{2}|nbh(x_{2_k}, v)|$ ; (iii)  $\phi_{BG}(XSet_j, XSet_k) > \frac{1}{2}$ . □

In this definition, conditions (i) and (ii) guarantee that  $XSet_j$  and  $XSet_k$  are representative (i.e., quite large); we assume that this happens if they involve more than half of the components of the corresponding neighborhoods. Finally, condition (iii) guarantees that  $XSet_j$  and  $XSet_k$  are similar.

The following theorem states the worst case time complexity for verifying if two neighborhoods are comparable. Its proof can be found in the Appendix.

**Theorem 3.2** Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $x_{1_j}$  (resp.,  $x_{2_k}$ ) be an x-component of  $S_1$  (resp.,  $S_2$ ). Let  $p$  be the maximum between  $|nbh(x_{1_j}, v)|$  and  $|nbh(x_{2_k}, v)|$ . The worst case time complexity for determining if  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$  are comparable is  $O(p^3)$ . □

**Corollary 3.1** Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $x_{1_j}$  (resp.,  $x_{2_k}$ ) be an x-component of  $S_1$  (resp.,  $S_2$ ). If  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$  are similar, then they are also comparable. □



### 3.2.3 Neighborhood generalization

Consider two neighborhoods  $\alpha$  and  $\beta$  and assume that: (1) they are not similar; (2) most of the x-components of  $\beta$  match with x-components of  $\alpha$ ; (3) most of the x-components of  $\alpha$  do not match with x-components of  $\beta$ . If all these conditions hold, then it is possible to conclude that the reality represented by  $\alpha$  is richer than that represented by  $\beta$  and, consequently, that  $\alpha$  is more specific than  $\beta$  or, conversely, that  $\beta$  is more general than  $\alpha$ . As an example,  $\alpha$  could be the set of attributes and sub-elements describing the concept *PhD Student* whereas  $\beta$  might be the set of attributes and sub-elements describing the concept *Student*. The following definition formalizes this reasoning.

**Definition 3.7** Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $x_{1_j}$  (resp.,  $x_{2_k}$ ) be an x-component of  $S_1$  (resp.,  $S_2$ ). We say that  $nbh(x_{1_j}, v)$  is more specific than  $nbh(x_{2_k}, v)$  (and, consequently, that  $nbh(x_{2_k}, v)$  is more general than  $nbh(x_{1_j}, v)$ ) if: (i) they are not similar and (ii) the objective function  $\varphi_{BG}(x_{1_j}, x_{2_k}, v) = \frac{|ESet'(v)|}{|QSet(v)|}$ , associated with the bipartite graph  $BG(x_{1_j}, x_{2_k}, v)$ , is greater than  $\frac{1}{2}$ ; here  $BG(x_{1_j}, x_{2_k}, v)$  has been described in Section 3.2,  $ESet'(v)$  represents the set of matching edges associated with  $BG$  whereas  $QSet(v)$  is the set of nodes of  $BG$  corresponding to the x-components of  $nbh(x_{2_k}, v)$ .  $\square$

The reasoning underlying Definition 3.7 derives from the observation that  $\varphi_{BG}(x_{1_j}, x_{2_k}, v)$  represents the share of x-components belonging to  $nbh(x_{2_k}, v)$  matching with the x-components of  $nbh(x_{1_j}, v)$ . If this share is sufficiently high then most of the x-components of  $nbh(x_{2_k}, v)$  match with the x-components of  $nbh(x_{1_j}, v)$  (condition (2)) but, since  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$  are not similar (condition (1)), most of the x-components of  $nbh(x_{1_j}, v)$  do not match with the x-components of  $nbh(x_{2_k}, v)$  (condition (3)). As a consequence, it is possible to conclude that  $nbh(x_{1_j}, v)$  is more specific than  $nbh(x_{2_k}, v)$  or, conversely, that  $nbh(x_{2_k}, v)$  is more general than  $nbh(x_{1_j}, v)$ .

The following theorem states the worst case time complexity for verifying if a neighborhood is more specific than another one. Its proof is provided in the Appendix.

**Theorem 3.3** Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $x_{1_j}$  (resp.,  $x_{2_k}$ ) be an x-component of  $S_1$  (resp.,  $S_2$ ). Let  $p$  be the maximum between  $|nbh(x_{1_j}, v)|$  and  $|nbh(x_{2_k}, v)|$ . The worst case time complexity for determining if  $nbh(x_{1_j}, v)$  is more specific than  $nbh(x_{2_k}, v)$  is  $O(p^3)$ .  $\square$

## 4 Extraction of interschema properties

In this section we illustrate our approach for the extraction of interschema properties. As pointed out in the Introduction, we shall concentrate our attention on the following properties: (i) *Synonymies*: a synonymy indicates

that two x-components have the same meaning. (ii) *Hyponymies/Hypernymies*: given two x-components  $x_S$  and  $x_T$ ,  $x_S$  is a hyponym of  $x_T$  (that is, in its turn, the hypernym of  $x_S$ ) if  $x_S$  has a more specific meaning than  $x_T$ . (iii) *Overlappings*: roughly speaking, given two x-components  $x_S$  and  $x_T$ , an overlapping holds between them if they are neither synonymous nor one a hyponym of the other but there exist non-empty sets of attributes and sub-elements  $\{x_{S_1}, x_{S_2}, \dots, x_{S_n}\}$  of  $x_S$  and  $\{x_{T_1}, x_{T_2}, \dots, x_{T_n}\}$  of  $x_T$  such that, for  $1 \leq i \leq n$ ,  $x_{S_i}$  is synonymous with  $x_{T_i}$ . (iv) *Homonymies*: an homonymy states that two x-components have the same name and the same typology, but different meanings.

Our approach exploits a thesaurus storing lexical synonymies holding among the terms of a language; specifically, it uses the English language and WordNet [19]. If necessary, different (possibly existing) domain-specific thesauruses could be used in the prototype implementing our approach; they can be provided by means of a suitable, friendly interface.

### 4.1 Derivation of candidate pairs

In order to verify if an interschema property holds between two x-components  $x_{1_j}$ , belonging to  $S_1$ , and  $x_{2_k}$ , belonging to  $S_2$ , it is necessary to examine their neighborhoods. Specifically, our approach operates as follows. First it considers  $nbh(x_{1_j}, 0)$  and  $nbh(x_{2_k}, 0)$  and verifies if they are comparable. In the affirmative case, it is possible to conclude that  $x_{1_j}$  and  $x_{2_k}$  refer to analogous “contexts” and, presumably, define comparable concepts. As a consequence, the pair  $\langle x_{1_j}, x_{2_k} \rangle$  is marked as *candidate* for an interschema property. However, observe that  $nbh(x_{1_j}, 0)$  (resp.,  $nbh(x_{2_k}, 0)$ ) takes only attributes and simple sub-elements of  $x_{1_j}$  (resp.,  $x_{2_k}$ ) into account; as a consequence, it considers quite a limited context. If a higher severity level is required, it is necessary to verify that other neighborhoods of  $x_{1_j}$  and  $x_{2_k}$  are comparable before marking the pair  $\langle x_{1_j}, x_{2_k} \rangle$  as candidate. Such a reasoning is formalized by the following definition.

**Definition 4.1** Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $x_{1_j}$  (resp.,  $x_{2_k}$ ) be an x-component of  $S_1$  (resp.,  $S_2$ ). Let  $u$  be a severity level. We say that the pair  $\langle x_{1_j}, x_{2_k} \rangle$  is *candidate for an interschema property* at the severity level  $u$  if  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$  are comparable for each  $v$  such that  $0 \leq v \leq u$ .  $\square$

It is possible to define a boolean function *candidate* that receives two x-components  $x_{1_j}$  and  $x_{2_k}$  and an integer  $u$  and returns *true* if  $\langle x_{1_j}, x_{2_k} \rangle$  is a candidate pair at the severity level  $u$ , *false* otherwise.

The following theorem states the computational complexity for the detection of candidate pairs. Its proof is immediate from Theorem 3.2 and Definition 4.1.

**Theorem 4.1** Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $x_{1_j}$  (resp.,  $x_{2_k}$ ) be an x-component of  $S_1$  (resp.,  $S_2$ ). Let  $u$

be a severity level. Finally, let  $p$  be the maximum between  $|nbh(x_{1_j}, u)|$  and  $|nbh(x_{2_k}, u)|$ . The worst case time complexity for verifying if  $\langle x_{1_j}, x_{2_k} \rangle$  is a candidate pair at the severity level  $u$  is  $O(u \times p^3)$ .  $\square$

## 4.2 Derivation of synonymies, hyponymies, overlappings and homonymies

Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $x_{1_j}$  (resp.,  $x_{2_k}$ ) be an x-component of  $S_1$  (resp.,  $S_2$ ). In order to verify if a synonymy, a hyponymy, an overlapping or an homonymy holds between  $x_{1_j}$  and  $x_{2_k}$  it is necessary to examine their neighborhoods and to determine the relationships holding among them. The following definition formalizes this reasoning:

**Definition 4.2** Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $x_{1_j}$  (resp.,  $x_{2_k}$ ) be an x-component of  $S_1$  (resp.,  $S_2$ ) and let  $u$  be a severity level.

- A synonymy holds between  $x_{1_j}$  and  $x_{2_k}$  at the severity level  $u$  if: (i)  $candidate(x_{1_j}, x_{2_k}, u) = true$ ; (ii)  $nbh(x_{1_j}, v)$  and  $nbh(x_{2_k}, v)$  are similar for each  $v$  such that  $0 \leq v \leq u$  (see Section 3.2.1).
- $x_{1_j}$  is said a hyponym of  $x_{2_k}$  (that, in its turn, is said a hypernym of  $x_{1_j}$ ) at the severity level  $u$  if: (i)  $candidate(x_{1_j}, x_{2_k}, u) = true$ ; (ii)  $nbh(x_{1_j}, 0)$  is more specific than  $nbh(x_{2_k}, 0)$  (see Section 3.2.3).
- An overlapping holds between  $x_{1_j}$  and  $x_{2_k}$  at the severity level  $u$  if: (i)  $candidate(x_{1_j}, x_{2_k}, u) = true$ ; (ii)  $x_{1_j}$  and  $x_{2_k}$  are not synonymous; (iii)  $x_{1_j}$  is neither a hyponym nor a hypernym of  $x_{2_k}$ .
- An homonymy holds between  $x_{1_j}$  and  $x_{2_k}$  at the severity level  $u$  if: (i)  $candidate(x_{1_j}, x_{2_k}, u) = false$ ; (ii)  $x_{1_j}$  and  $x_{2_k}$  have the same name; (iii)  $x_{1_j}$  and  $x_{2_k}$  are both elements or both attributes.  $\square$

It is possible to define a boolean function  $synonymy(x_{1_j}, x_{2_k}, u)$  (resp.,  $hyponymy(x_{1_j}, x_{2_k}, u)$ ,  $overlapping(x_{1_j}, x_{2_k}, u)$ ,  $homonymy(x_{1_j}, x_{2_k}, u)$ ), that receives two x-components  $x_{1_j}$  and  $x_{2_k}$  and an integer  $u$  and returns *true* if a synonymy (resp., a hyponymy, an overlapping, an homonymy) holds between  $x_{1_j}$  and  $x_{2_k}$  at the severity level  $u$ , *false* otherwise.

As for the computational complexity of the interschema property derivation, it is possible to state the following theorem whose proof can be found in the Appendix.

**Theorem 4.2** Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $x_{1_j}$  (resp.,  $x_{2_k}$ ) be an x-component of  $S_1$  (resp.,  $S_2$ ). Let  $u$  be a severity level. Finally, let  $p$  be the maximum between  $|nbh(x_{1_j}, u)|$  and  $|nbh(x_{2_k}, u)|$ . The worst case time complexity for computing  $synonymy(x_{1_j}, x_{2_k}, u)$ ,  $hyponymy(x_{1_j}, x_{2_k}, u)$ ,  $overlapping(x_{1_j}, x_{2_k}, u)$ ,  $homonymy(x_{1_j}, x_{2_k}, u)$  is  $O(u \times p^3)$ .  $\square$

**Corollary 4.1** Let  $S_1$  and  $S_2$  be two XML Schemas. Let  $u$  be a severity level. Let  $m$  be the maximum between the number of complex elements of  $S_1$  and  $S_2$ . Finally, let  $q$  be the maximum cardinality of a neighborhood of  $S_1$  or  $S_2$ . The worst case time complexity for deriving all interschema properties holding between  $S_1$  and  $S_2$  at the severity level  $u$  is  $O(u \times q^3 \times m^2)$ .  $\square$

## 5 Experimental results

### 5.1 Introduction

In this section we provide a detailed description of the experiments we have carried out for testing the performance of our approach. We have performed a large variety of experiments, devoted to test the various aspects of our approach; they will be presented in the next subsections.

It is worth pointing out that some of our tests have been inspired to ideas and methodologies illustrated in [6]; in this paper, the authors propose a catalogue of criteria for comparing some of the most popular interschema property extraction systems, namely, *Autoplex* [2], *Automatch* [3], *COMA* [7], *Cupid* [17], *LSD* [8], *GLUE* [10], *SemInt* [16] and *SF (Similarity Flooding)* [18]. In our opinion, this is a very interesting effort and we have decided to exploit the same criteria (and, whenever possible, the same sources) for testing the performance of our approach. This choice allowed us to obtain an objective evaluation of our approach, as well as to make a precise comparison between it and the other systems evaluated by [6].

### 5.2 Characteristics of the exploited sources

In our tests we have exploited a large variety of XML Schemas relating to disparate application contexts; specifically, we have considered XML Schemas relating to Biomedical Data, Project Management, Property Register, Industrial Companies, Universities, Airlines, Scientific Publications and Biological Data. In our tests, we have compared all pairs of XML schemas within a particular domain.

Biomedical Schemas have been derived from various sites; among them we cite: <http://www.biomediator.org>. XML Schemas relating to Project Management, Property Register and Industrial Companies have been derived from Italian Central Government Office sources and are shown at the address: <http://www.mat.unical.it/terracina/tests.html>. XML Schemas relating to Universities have been downloaded from the address: <http://anhai.cs.uiuc.edu/archive/domains/courses.html>. XML Schemas relating to Airlines have been found in [22]. XML Schemas relating to Scientific Publications have been supplied by the authors of [15]. Finally, Biological Schemas have been downloaded from the addresses: <http://smi-web.stanford.edu/projects/>

helix/pubs/ismb02/schemas/,  
[http://www.cs.toronto.edu/db/clio/data/GeneX\\_RDB-s.xsd](http://www.cs.toronto.edu/db/clio/data/GeneX_RDB-s.xsd) and <http://www.genome.ad.jp/kegg/soap/v3.0/KEGG.wsdl>.

As far as exploited thesauruses are concerned, we have used WordNet for XML Schemas relating to Project Management, Property Register, Industrial Companies, Universities, Airlines and Scientific Publications. On the contrary, for Biomedical and Biological Schemas we have exploited the Biocomplexity Thesaurus, a biological domain specific thesaurus available at the address: <http://thesaurus.nbii.gov>.

Examined sources were characterized by the following properties, expressed according to the terminology and the measures of [6]:

**Number of schemas:** we have considered 35 XML Schemas whose characteristics are reported in Table 1; this number of schemas is quite similar to those considered by the authors of the other approaches for performing their evaluation; they are reported in Table 2. From this table it is possible to see that the number of schemas exploited by the other approaches for carrying out their evaluation activity ranges from 2 to 24.

**Size of schemas:** the size of the evaluated XML Schemas, i.e., the number of their elements and attributes, ranges from 12 to 645. The minimum, the maximum and the average size of the sources exploited for evaluating the other approaches, derived by [6], are reported in Table 2<sup>1</sup>. An analysis of this table shows that the sizes of the schemas evaluated by our approach are quite close to those of the sources examined by the other systems. The size of a test schema is relevant because it influences the quality of obtained results; in fact, as mentioned in [6], the bigger the input schemas are, the greater the search space for candidate pairs is and the lower the quality of obtained results will be.

The number of comparisons we have carried out for each domain are shown in the last column of Table 1.

### 5.3 Accuracy Measures exploited in our experimental tests

All accuracy measures proposed in [6] and computed during our experiments have been obtained according to the following general framework: (i) a set of experts has been asked to identify interschema properties existing among involved XML Schemas; (ii) interschema properties among the same XML Schemas have been determined by the approach to evaluate; (iii) the properties provided by the experts and those returned by the approach to test have been compared and accuracy measures have been computed.

The number of experts that have been involved in manually solving the match tasks is as follows: 6 for Biomedical Data, 3 for Project Management, 3 for Property Register, 4

for Industrial Companies, 4 for Universities, 2 for Airlines, 2 for Scientific Publications and 7 for Biological Data.

Let  $A$  be the set of properties provided by the experts and let  $C$  be the set of properties returned by the approach to evaluate; two basic accuracy measures are: (i) *Precision* (hereafter  $Pre$ ), that specifies the share of correct properties detected by the system among those it derived. It is defined as:  $Pre = \frac{|A \cap C|}{|C|}$ . (ii) *Recall* (hereafter  $Rec$ ), that indicates the share of correct properties detected by the system among those the experts provided. It is defined as:  $Rec = \frac{|A \cap C|}{|A|}$ . Precision and Recall are typical measures of Information Retrieval (see [25]). Both of them fall within the interval  $[0, 1]$ ; in the ideal case (i.e., when  $A \equiv C$ ) they are both equal to 1. It is worth noting that the set  $C$  of interschema properties our approach derives varies with the severity level; in order to make this evident, we shall use the symbol  $C(u)$  instead of  $C$ . However, as pointed out in [6], neither Precision nor Recall alone can accurately measure the quality of an interschema property extraction algorithm; in order to improve the result accuracy, it appears necessary to consider a joint measure of them. Two very popular measures satisfying these requirements are: (i) *F-Measure* [3, 25], that represents the harmonic mean between Precision and Recall. It is defined as:  $F\text{-Measure} = 2 \cdot \frac{Pre \cdot Rec}{Pre + Rec}$ . (ii) *Overall* [6, 18], that measures the post-match effort needed for adding false negatives and removing false positives from the set of properties returned by the system to evaluate. It is defined as:  $Overall = Rec \cdot (2 - \frac{1}{Pre})$ . F-Measure falls within the interval  $[0, 1]$  whereas Overall ranges between  $-\infty$  and 1; the higher F-Measure (resp., Overall) is, the better the accuracy of the tested approach will be.

### 5.4 Discussion of obtained results

As for the evaluation of Precision and Recall associated with our approach, we argued that, due to its philosophy and intrinsic structure, an increase of the severity level should have caused an increase of its Precision and a decrease of its Recall. This intuition is motivated by considering that  $C(u+1) \subseteq C(u)$  and that  $C(u+1)$  is obtained from  $C(u)$  by eliminating the weakest properties; this should cause  $C(u+1)$  to be more precise than  $C(u)$ . However, this filtering task could erroneously discard some valid properties; for this reason  $C(u+1)$  could have a smaller Recall than  $C(u)$ .

In order to verify this intuition and, possibly, to quantify it, we have applied our approach on our test Schemas and we have computed the Average Precision, the Average Recall, the Average F-Measure and the Average Overall at various severity levels. Obtained results are shown in Table 3. From the analysis of this table we can draw the following conclusions:

- As for the severity level 0, (i) Precision shows its lowest value; as a consequence, our approach returns some false positives; (ii) Recall assumes its highest

<sup>1</sup>The size of a relational source has been intended as the number of its relations and attributes.

Application context	Number of Schemas	Max Depth	Minimum, Average and Maximum Number of <i>x</i> -components	Minimum, Average and Maximum Number of complex elements	Total Number of Comparisons
Biomedical Data	11	8	15 - 26 - 38	4 - 8 - 16	55
Project Management	3	4	37 - 40 - 42	6 - 7 - 8	3
Property Register	2	4	64 - 70 - 75	14 - 14 - 14	1
Industrial Companies	5	4	23 - 28 - 46	6 - 8 - 9	10
Universities	5	5	15 - 17 - 19	3 - 4 - 5	10
Airlines	2	4	12 - 13 - 13	4 - 4 - 4	1
Scientific Publications	2	6	17 - 18 - 18	8 - 9 - 9	1
Biological Data	5	8	250 - 327 - 645	36 - 60 - 206	10

Table 1: Characteristics of the XML Schemas exploited for testing the performance of our approach

System	Typology of tested Schema	Number of Schemas	Minimum size of Schemas	Maximum size of Schemas	Average size of Schemas
Our system	XML	35	12	645	70
Autoplex & Automatch	Relational	15	-	-	-
COMA	XML	5	40	145	77
Cupid	XML	2	40	54	47
LSD	XML	24	14	66	-
GLUE	XML	3	34	333	143
SemInt	Relational	10	6	260	57
SF	XML	18	5	22	12

Table 2: Characteristics of the XML Schemas exploited by the other approaches for their evaluation activity

Property Typology	Average Precision	Average Recall	Average F-Measure	Average Overall
Severity Level 0	0.86	0.97	0.91	0.81
Severity Level 1	0.96	0.81	0.88	0.78
Severity Level 2	0.97	0.77	0.86	0.75
Severity Level 3	0.97	0.72	0.83	0.70

Table 3: Accuracy measures associated with our approach at various severity levels

value; as a consequence, our approach returns almost all valid properties or, in other words, it returns a very small number of false negatives.

- If the severity level is 1, (i) the set of properties returned by our approach contains a smaller number of false positives than the previous case; specifically, it is possible to observe that Precision increases to 0.96; (ii) Recall decreases of about 16% w.r.t. the previous case; in other words, a certain increase of false negatives can be observed.
- As for the severity level 2, (i) Precision slightly increases to 0.97; (ii) Recall decreases of about 5% w.r.t. the previous case.
- When the severity level is equal to 3, (i) Precision saturates at its highest value, i.e., 0.97; (ii) Recall presents the same trend as the previous case; specifically, a further decrease is observed.

All these experiments confirm our original intuition about the trend of Precision and Recall in presence of variations of the severity level.

From the examination of Table 3 we observe that passing from low to high severity levels causes an increase of the Precision and a corresponding decrease of the Recall. This behaviour is explained by considering that, at low severity levels, a user is willing to accept false positives if this allows him to obtain a complete set of similarities. On the

contrary, at high severity levels, a user is willing to receive an incomplete set of similarities by the system but he desires that proposed properties are (almost surely) correct.

Table 3 shows also the great importance of the severity level that provides our approach with a high flexibility. As a matter of fact, in real cases, there are many application contexts where having a high Recall is more important than achieving a high Precision; in these cases our approach can be applied with a severity level equal to 0. On the contrary, there are other situations where obtaining a high Precision is more relevant than having a high Recall; in these situations the user might: (i) obtain it automatically by setting an adequate, presumably high, severity level; in this way the automaticity of the approach is preserved but its Recall decreases; (ii) obtain it semi-automatically by setting a low severity level and by performing a further, deep, validation of obtained results; in this way the Recall of the approach is preserved but the time the user needs for validation sensibly increases.

After this, we have compared the accuracy of our approach w.r.t. that of the other approaches evaluated in [6]; obtained results are reported in Table 4. We point out that the accuracy measure of the other systems shown in that table have been directly derived from [6]. The only missing data regard Cupid; in fact, in [6], the Authors provide only a qualitative analysis of this system without specifying any quantitative value of its Precision, its Recall, its F-Measure and its Overall. However, a quantitative analysis of Cupid can be found in [13]; in that paper the Authors claim that, for the schemas considered by them, Cupid showed a Precision equal to 0.60, a Recall equal to 0.55, an F-Measure equal to 0.57 and an Overall equal to 0.18; in order to allow a more precise comparison, we must say that the Authors of [13] applied also COMA on the same sources considered for Cupid and, for these sources, they obtained a Precision equal to 0.82, a Recall equal to 0.75, an F-Measure equal to 0.78 and an Overall equal to 0.59.

From the analysis of Table 4 we can observe that: (i) at the severity level 0 the Precision of our approach is satisfactory, even if COMA presents a better value; at the severity level 1 our approach has the highest Precision; (ii) at the severity level 0 our approach shows the best Recall; on the contrary, at the severity level 1, the Recall of our approach significantly decreases; (iii) at the severity level 0 our approach presents, along with COMA, the highest values of F-Measure and Overall; both these two accuracy measures slightly decrease at the severity level 1.

As a conclusion, in our opinion, all these experiments agree on determining that the accuracy of our approach is extremely satisfactory and promising. In addition, our approach shows a great flexibility in that it can be adapted for obtaining the best Precision or the best Recall, according to the exigencies of the application context it is operating in. These results are even more relevant if we take into account that both measures and most of the test sources we have considered had been already uniformly exploited for evaluating a large variety of existing approaches.

We have also verified if the accuracy of our approach depends on the application domain which the test Schemas belong to. The results we have obtained are shown in Figure 1. From the analysis of this figure, it is possible to conclude that the accuracy of our approach is substantially independent of the application domain. As far as our experiments are concerned, we have obtained the best accuracy for the Property Register domain; here, Precision reaches its best value at the severity level 3 and is 0.99; Recall, F-Measure and Overall are maximum at the severity level 0 and are 0.99, 0.94 and 0.87, respectively. We have obtained the worst accuracy in the Biological domain; here, Precision is maximum at the severity level 3 and is 0.87; Recall, F-Measure and Overall reach their best values at the severity level 0 and are 0.87, 0.82 and 0.62, respectively.

## 5.5 Robustness analysis

### 5.5.1 Robustness against structural dissimilarities

XML is inherently hierarchical; it allows nested, possibly complex, structures to be exploited for representing a domain. As a consequence, two human experts might model the same reality by means of two XML Schemas characterized by deep structural dissimilarities. We have performed a robustness analysis of our approach, devoted to verify if it is resilient to structural dissimilarities. Before describing our experimental tests about this issue, we point out that the specific features of our approach make it intrinsically robust for the following two cases, that are very common in practice:

- If the typology of an x-component  $x_{1_j}$  of an XML Schema  $S_1$  changes from “simple element” to “attribute”, or vice versa, no modifications of the interschema properties involving x-components of  $S_1$  occur. This result directly derives from the definition of the function *veryclose* (see Section 3).

- If  $x_{1_j}$  and  $x'_{1_j}$  are two complex elements of the same XML Schema  $S_1$  such that  $x'_{1_j}$  is a sub-element of  $x_{1_j}$  and if  $S_1$  is modified in such a way that  $x'_{1_j}$  is no longer a sub-element of  $x_{1_j}$  but there exists a `keyref` relating  $x_{1_j}$  to  $x'_{1_j}$ , then no modifications of the interschema properties involving x-components of  $S_1$  occur. An analogous reasoning holds for the opposite change. This result directly derives from the definition of the function *close* (see Section 3).

There are further structural modifications that could influence the results of our approach and for which it is not intrinsically robust; for these cases an experimental measure of its robustness appears necessary. Two of the most common structural modifications are analyzed in the following.

**Flattening of x-components.** Consider Figure 2 illustrating two portions of XML Schemas representing persons. Specifically, in the first XML Schema, the concept “Person” is represented by means of a nested hierarchical structure; on the contrary, in the second XML Schema, the same concept is represented by means of a flat structure.

In order to determine the robustness of our approach against errors occurring owing to the flattening of x-components, for each pair of XML Schemas into consideration, we have progressively altered the structure of one of the XML Schemas by transforming a certain percentage of its x-components from a nested structure to a flat one. For each of these transformations, we have derived the interschema properties associated with the “modified” versions of the XML Schemas and we have computed the corresponding values of the accuracy measures. Specifically, we have considered five cases, corresponding to a percentage of flattened x-components (hereafter *FXP - Flattened X-component Percentage*) equal to: (a) 0%; (b) 7%; (c) 14%; (d) 21%; (e) 28%. The results we have obtained are shown in Figure 3.

From the analysis of this figure it is possible to observe that our approach shows a good robustness against increases of *FXP*. As a matter of fact, even if structural dissimilarities occur, the changes in the accuracy measures are generally quite small. In fact, the maximum decrement of the Average Precision (resp., Average Recall, Average F-Measure, Average Overall) w.r.t. case (a) is equal to 0.11 (resp., 0.16, 0.13, 0.24) and can be found at the severity level 1 (resp., 0, 2, 0). However, we stress that if the increases of *FXP* would be significantly greater than those considered above, the changes in the accuracy measures could be significant; this behaviour is correct since it guarantees that our approach shows the right degree of sensitivity to changes to the structure of involved XML Schemas.

**Exchange of nesting levels.** Consider Figure 4 illustrating two portions of XML Schemas representing catalogues. In the first XML Schema, a catalogue is organized by grouping involved models by brands and, then, by product

System	Precision	Recall	F-Measure	Overall
Our system (severity level 0)	0.86	0.97	0.91	0.81
Our system (severity level 1)	0.96	0.81	0.88	0.78
Autoplex & Automatch	0.84	0.82	0.82 & 0.72	0.66
COMA	0.93	0.89	0.90	0.82
Cupid	—	—	—	—
LSD	~ 0.80	0.80	~ 0.80	~ 0.60
GLUE	~ 0.80	0.80	~ 0.80	~ 0.60
SemInt	0.78	0.86	0.81	0.48
SF	—	—	—	~ 0.60

Table 4: Comparison of the accuracy of our approach w.r.t. that of the other approaches evaluated in [6]

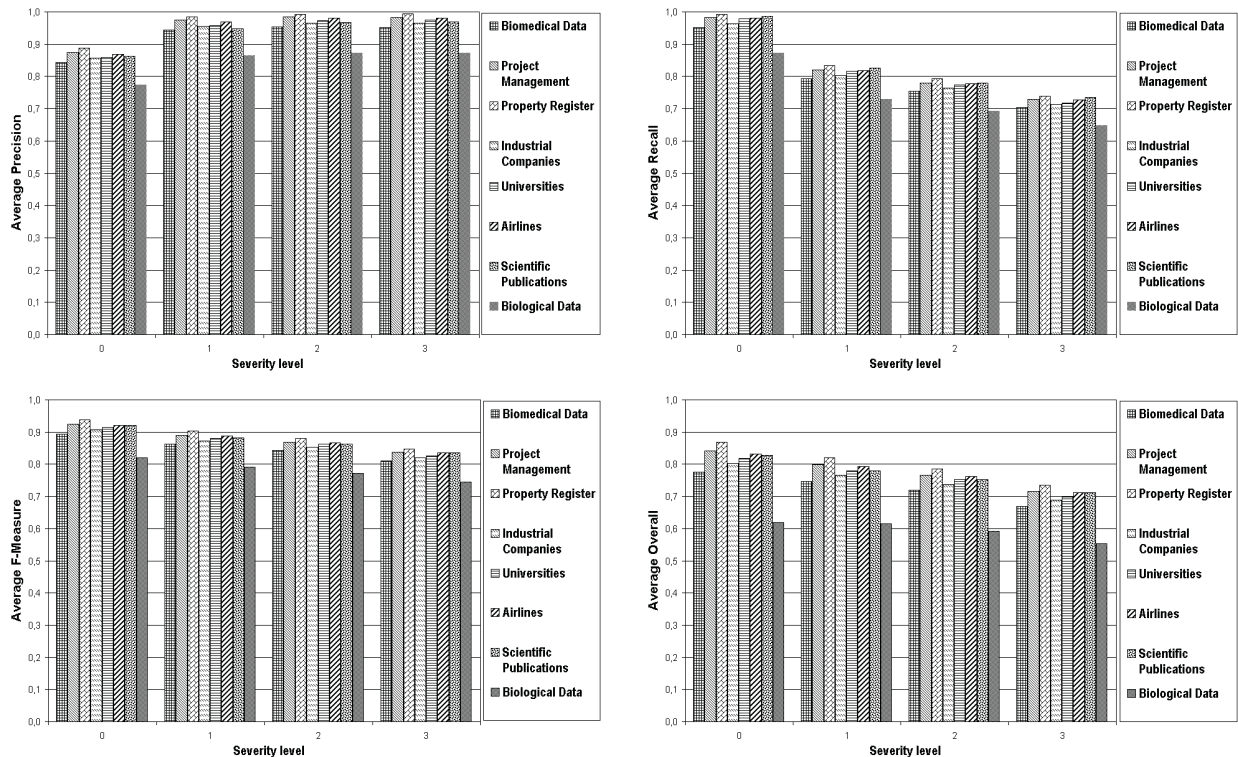


Figure 1: Average Precision, Average Recall, Average F-Measure, Average Overall of our approach in different domains

categories; on the contrary, in the second XML Schema, the same catalogue is organized by grouping involved models by product categories and, then, by brands. In this case we have that two x-components, namely “brand” and “product\_category” exchanged their nesting levels within the corresponding XML Schemas. Clearly, an exchange of nesting levels may occur only between complex elements.

Note that the exchange of nesting levels between two complex elements is not always “safe” from a semantic point of view. In fact, consider, again, the first XML Schema of Figure 4, and assume the nesting level of “catalogue” and “brand” to be exchanged; in this case, the semantics of the resulting XML Schema would be quite different w.r.t. that of the original XML Schema; in fact, the new XML Schema would represent a list of brands each of which associated with a separate catalogue of products. Therefore, as for the robustness of our approach in the management of this kind of structural modification, we could expect a decrease of performance w.r.t. the previous case because the semantic modifications produced by the ex-

change of nesting levels are deeper than those caused by the flattening of x-components.

Our approach is partially intrinsically robust against this kind of structural modification. In fact, our definition of neighborhood, which is the core of our interschema property extraction technique, puts in the same set all the x-components laying at a “distance” less than or equal to  $j$  from the component under consideration (see Definition 3.4). Now, consider an x-component  $x_S$  and assume that an exchange of nesting levels occurs between two of its sub-elements, say  $x'_S$  and  $x''_S$ , laying at distance  $j$  and  $j+1$  from  $x_S$ , respectively; this structural modification will imply some differences in  $nbh(x_S, j)$  but not in  $nbh(x_S, j+1)$ . Specifically,  $nbh(x_S, j)$  contains  $x'_S$  before the exchange of nesting levels whereas it contains  $x''_S$  after the exchange; by contrast,  $nbh(x_S, j+1)$  contains  $x'_S$  and  $x''_S$  both before and after the exchange. This implies that a possible error in the evaluation of the interschema properties involving  $x_S$  may occur only when the  $j^{th}$  neighborhood of  $x_S$  is considered; however, this error is not propagated through



```

<xs:element name="person">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="address"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="address">
  <xs:complexType>
    <xs:attribute name="city" type="xs:string"/>
    <xs:attribute name="state" type="xs:string"/>
    <xs:attribute name="country" type="xs:string"/>
    <xs:attribute name="zip" type="xs:string"/>
  </xs:complexType>
</xs:element>

<xs:element name="person">
  <xs:complexType>
    <xs:attribute name="first_name" type="xs:string"/>
    <xs:attribute name="last_name" type="xs:string"/>
    <xs:attribute name="gender" type="xs:string"/>
    <xs:attribute name="birthdate" type="xs:date"/>
    <xs:attribute name="city" type="xs:string"/>
    <xs:attribute name="state" type="xs:string"/>
    <xs:attribute name="country" type="xs:string"/>
    <xs:attribute name="zip" type="xs:string"/>
  </xs:complexType>
</xs:element>
    
```

Figure 2: Example of “nested” and “flat” structures

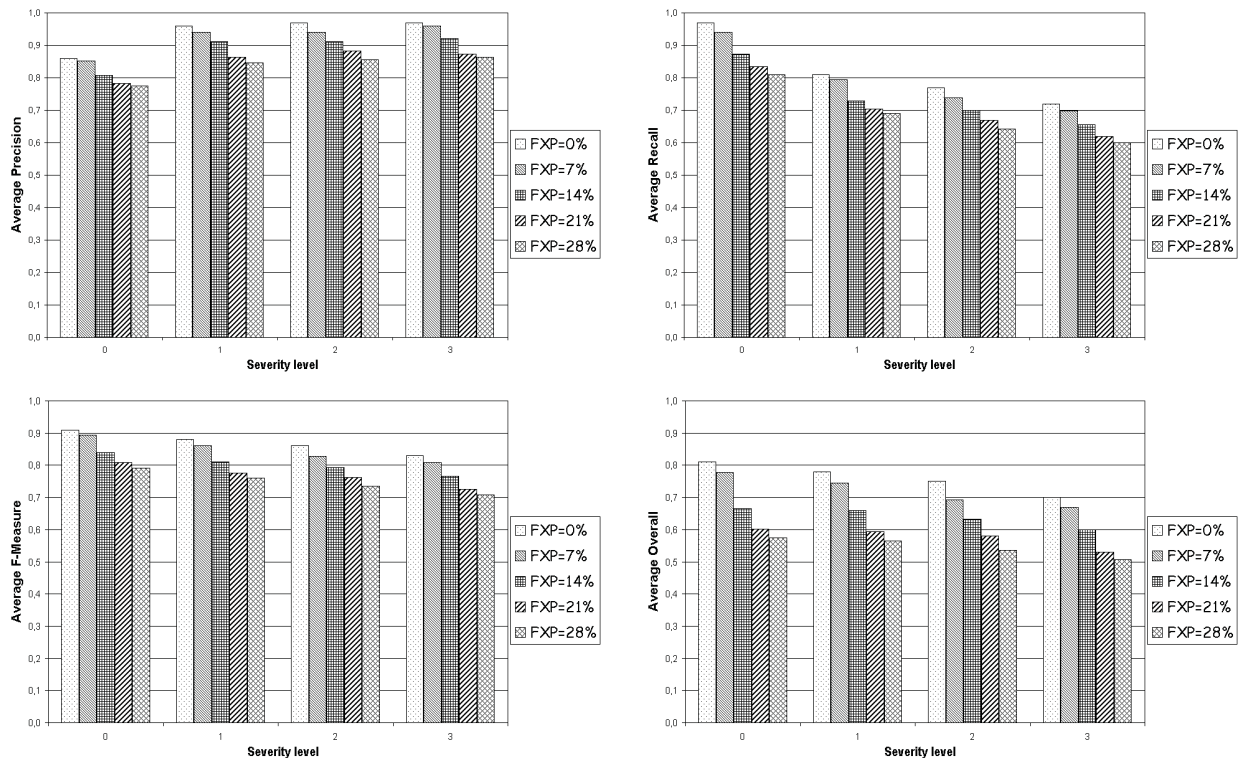


Figure 3: Average Precision, Average Recall, Average F-Measure, Average Overall for various values of *FXP*

the next neighborhoods. This important feature of our approach mitigates the possible problems arising from the structural modifications caused by the exchange of nesting levels.

In order to quantitatively evaluate the robustness of our approach against errors caused by the exchange of nesting levels, for each pair of XML Schemas into consideration, we have progressively altered the structure of one of the XML Schemas of the pair by exchanging the nesting level of a certain percentage of its x-components. For each of these transformations, we have derived the interschema properties associated with the “modified” versions of the XML Schemas and we have computed the corresponding values of the accuracy measures. Specifically, we have considered five cases, corresponding to a percentage of exchanged nesting levels (hereafter *ENP - Exchanged Nesting level Percentage*) equal to: (a) 0%; (b) 7%; (c) 14%;

(d) 21%; (e) 28%. The results that we have obtained are shown in Figure 5.

From the analysis of this figure it is possible to observe that the robustness of our approach against increases of *ENP* is satisfactory, even if, as expected, its overall performance is slightly worse than that obtained for the same percentage of *FXP*. In any case, the changes of the accuracy measures caused by an increase of *ENP* are generally acceptable. In fact, the maximum decrement of the Average Precision (resp., Average Recall, Average F-Measure, Average Overall) w.r.t. case (a) is equal to 0.17 (resp., 0.21, 0.18, 0.33) and can be found at the severity level 2 (resp., 0, 2, 0). However, analogously to the previous case, if the increases of *ENP* would be quite high, the variations of the semantics of the corresponding XML Schemas would be also significant and, consequently, the accuracy measures might significantly decrease; however, we point out

---

```

<xs:element name="catalogue">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="brand"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="brand">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="product_category"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="product_category">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="model"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="model">
  <xs:complexType>
    <xs:attribute name="detail" type="xs:string"/>
    <xs:attribute name="price" type="xs:string"/>
  </xs:complexType>
</xs:element>

```

```

<xs:element name="catalogue">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="product_category"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="product_category">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="brand"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="brand">
  <xs:complexType>
    <xs:sequence>
      <xs:element ref="model"/>
    </xs:sequence>
  </xs:complexType>
</xs:element>
<xs:element name="model">
  <xs:complexType>
    <xs:attribute name="detail" type="xs:string"/>
    <xs:attribute name="price" type="xs:string"/>
  </xs:complexType>
</xs:element>

```

---

Figure 4: Example of exchange of nesting levels

that this behaviour is desirable since it proves, again, that our approach shows a good degree of sensitivity against changes of the structure of involved XML Schemas.

### 5.5.2 Robustness against thesaurus errors

In this experiment we have tested the effects of errors and inaccuracies in the thesaurus received in input by our approach. Specifically, we have asked experts to validate the similarities contained in the input thesauruses and involving elements and attributes of the considered XML Schemas in such a way to remove any possible error.

After this, we have performed some variations on the corrected thesauruses and, for each of them, we have computed Average Precision, Average Recall, Average F-Measure and Average Overall of our system. Variations we have carried out on the correct thesauruses are: (a) 10% of correct similarities have been filtered out; (b) 20% of correct similarities have been filtered out; (c) 30% of correct similarities have been filtered out; (d) 50% of correct similarities have been filtered out; (e) 10% of wrong similarities have been added; (f) 20% of wrong similarities have been added; (g) 30% of wrong similarities have been added; (h) 50% of wrong similarities have been added.

Table 5 presents the values of the Average Precision, the Average Recall, the Average F-Measure and the Average Overall we have obtained for the extraction of the interschema properties at various severity levels. These results show that our system is quite robust w.r.t. errors and inaccuracies in the thesauruses provided in input. At the same time it shows a good sensitivity against errors because, if the correct similarities that are filtered out or the wrong similarities that are added are excessive, the system accuracy significantly decreases.

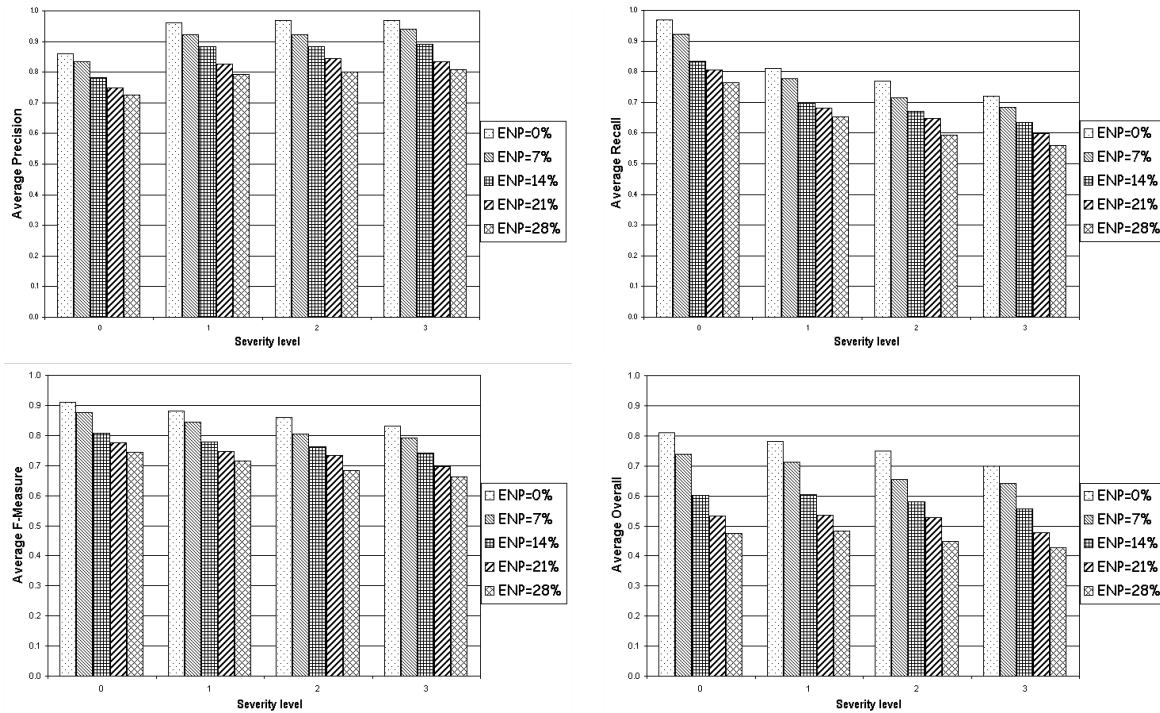
## 6 Comparison between our approach and the related ones illustrated in Section 2

In this section we compare our approach with the related ones already illustrated in Section 2.

**CGLUE.** The only similarity between our approach and CGLUE concerns the exploitation of *auxiliary information*; in particular, CGLUE uses the *training matches* (i.e., semantic matches provided by the users for training its learners) whereas our approach exploits a thesaurus. As for differences existing between them, we may observe that: (i) CGLUE exploits machine learning techniques, whereas our approach is based on graph matching algorithms; (ii) CGLUE is *generic* whereas our approach is specialized for XML sources; (iii) CGLUE is both *schema-based* and *instance-based*; as a consequence, it requires a deep analysis of data instances; by contrast, our approach is *schema-based*; (iv) CGLUE is *composite* in that it combines various algorithms for detecting semantic matches; by contrast, our approach is *hybrid*; (v) CGLUE was conceived for detecting *1:1*, *1:n* and *n:m* matchings, whereas our approach aims to derive *1:1 matchings*.

**Approach of [11].** The approach of [11] and ours share some similarities; specifically, (i) both of them are *hybrid*; (ii) both of them were conceived for detecting *1:1 matchings*. As for differences, we may observe that: (i) in order to carry out its tasks, the approach of [11] exploits fuzzy tools whereas our approach uses graph-based techniques; (ii) the approach of [11] is *generic*, i.e., it is not specialized for XML sources; (iii) the approach of [11] is *instance-based* whereas our approach is *schema-based*; (iv) the approach of [11] does not require *auxiliary information*.

**Cupid.** As for similarities between our approach and Cupid, we may notice that: (i) in both of them schema elements are matched in a pair-wise manner by means of

Figure 5: Average Precision, Average Recall, Average F-Measure and Average Overall for various values of  $ENP$ 

Case	Average Precision Severity Levels 0-1-2-3	Average Recall Severity Levels 0-1-2-3	Average F-Measure Severity Levels 0-1-2-3	Average Overall Severity Levels 0-1-2-3
No errors	0.86 - 0.96 - 0.97 - 0.97	0.97 - 0.81 - 0.77 - 0.72	0.91 - 0.88 - 0.86 - 0.83	0.81 - 0.78 - 0.75 - 0.70
(a)	0.86 - 0.96 - 0.97 - 0.97	0.92 - 0.77 - 0.73 - 0.68	0.89 - 0.86 - 0.83 - 0.80	0.77 - 0.74 - 0.71 - 0.66
(b)	0.86 - 0.96 - 0.97 - 0.97	0.86 - 0.72 - 0.68 - 0.64	0.86 - 0.82 - 0.80 - 0.77	0.72 - 0.69 - 0.66 - 0.62
(c)	0.87 - 0.97 - 0.98 - 0.98	0.77 - 0.64 - 0.61 - 0.57	0.82 - 0.77 - 0.75 - 0.72	0.65 - 0.62 - 0.60 - 0.56
(d)	0.87 - 0.97 - 0.98 - 0.98	0.68 - 0.57 - 0.54 - 0.50	0.76 - 0.71 - 0.69 - 0.66	0.57 - 0.55 - 0.53 - 0.49
(e)	0.82 - 0.91 - 0.92 - 0.92	0.97 - 0.81 - 0.77 - 0.72	0.89 - 0.86 - 0.84 - 0.81	0.75 - 0.73 - 0.71 - 0.66
(f)	0.76 - 0.85 - 0.86 - 0.86	0.97 - 0.81 - 0.77 - 0.72	0.85 - 0.83 - 0.81 - 0.78	0.67 - 0.67 - 0.64 - 0.60
(g)	0.68 - 0.76 - 0.77 - 0.77	0.98 - 0.81 - 0.77 - 0.72	0.80 - 0.79 - 0.77 - 0.75	0.52 - 0.56 - 0.54 - 0.51
(h)	0.60 - 0.67 - 0.68 - 0.68	0.98 - 0.82 - 0.78 - 0.73	0.75 - 0.74 - 0.72 - 0.70	0.33 - 0.42 - 0.41 - 0.38

Table 5: Variation of Precision, Recall, F-Measure and Overall w.r.t. possible errors in the input thesauruses

suitable similarity functions; (ii) both of them are *schema-based*; (iii) both of them are *hybrid*; (iv) both of them exploit a thesaurus as *auxiliary information*. As for differences, we may observe that: (i) Cupid is based on tree matching whereas our approach is based on graph matching; (ii) Cupid is capable of managing *generic* data sources whereas our approach has been developed for operating only on XML sources; (iii) Cupid is capable of extracting also *1:n matchings* whereas our approach has been conceived for deriving only *1:1 matchings*.

**MOMIS.** Some similarities exist between our approach and MOMIS; in fact: (i) both of them are *schema-based*; (ii) both of them are *hybrid*; (iii) both of them derive *1:1 matchings*; (iv) both of them exploit a thesaurus as *auxiliary information*. As for differences, we may observe that: (i) MOMIS is based on description logics whereas our approach is graph-based; (ii) MOMIS is *generic*; (iii) MOMIS has been conceived mainly for integration and

querying whereas our approach is specialized for inter-schema property extraction.

**Approach of [14].** There exist some similarities between the approach of [14] and ours; specifically, (i) both of them are *schema-based*; (ii) both of them are *hybrid*; (iii) both of them derive *1:1 matchings*. There are also important differences between the two approaches; specifically: (i) in order to perform matching activities, the approach of [14] adopts *statistical-based* techniques whereas our approach operates on graphs; (ii) the approach of [14] operates on databases that can be accessed through Web query interfaces; (iii) the approach of [14] does not exploit *auxiliary information*; (iii) the approach of [14] creates a hidden schema which is both capable of fully describing a domain and useful as a mediated schema; such a characteristic is not present in our approach; however, as claimed by the authors, this makes the approach of [14] to be exponential; as a consequence, the approach of [14] can be applied only if schema match-

ing is carried out off-line; on the contrary, our approach is much more light and can be applied both on-line and off-line.

**Approach of [4].** The main goal of the approach proposed in [4] is clearly different from that of our approach; in fact, the approach of [4] has been conceived to determine if data stored in an XML document approximatively conform to a DTD; by contrast, our approach aims to detect semantic similarities between two XML Schemas. Despite this substantial difference, we can observe that the approach of [4] and ours share some similarities. Specifically: (i) in both of them the analysis of structural properties of input data sources plays a key role; (ii) both of them clearly distinguish the roles played by simple and complex elements; (iii) both of them consider the constraints related to the occurrences of an element (e.g., if an element is optional or mandatory); (iv) both of them are specific for XML sources; (v) both of them are *hybrid*. As for the main differences between the two approaches, we observe that: (i) the approach of [4] is based on tree matching whereas our approach is based on graph matching; (ii) the approach of [4] is both *schema-based* and *instance-based*; (iii) the approach of [4] can extract *1:1*, *1:m* and *m:n* matchings; (iv) the approach of [4] does not exploit any *auxiliary information*.

**DIKE.** There are some similarities between our approach and DIKE; specifically, both of them: (i) are *graph-based*; (ii) are *schema-based*; (iii) are *hybrid*; (iv) exploit a thesaurus as *auxiliary information*.

However, there are also important differences between them; specifically: (i) DIKE operates on E/R schemas whereas our approach is graph-based. (ii) DIKE derives *1:1*, *1:n* and *m:n* matchings. (iii) The algorithms underlying DIKE rely on various thresholds and weights. (iv) DIKE does not consider a “severity” level that, on the contrary, plays a key role in our approach. (v) As far as the property derivation technique is concerned, DIKE and our system follow very different philosophies. As a matter of fact, DIKE exploits a sophisticated fixpoint computation strategy to derive interschema properties, whereas the approach we are presenting in this paper is simpler. (vi) Finally, the user intervention required by DIKE is heavier than that required by our approach since the former requires a tuning activity to be carried out for all thresholds and weights before the extraction process can start.

## 7 Conclusions

In this paper we have proposed an approach for the extraction of synonymies, hyponymies, overlappings and homonymies from a set of XML Schemas. We have shown that our approach is specialized for XML sources, is almost automatic, semantic and “light”; it derives all these properties in a uniform way and allows the choice of the

“severity” level against which the extraction task must be performed.

We have illustrated some experiments that we have carried out to test its performance and to compare its results with those achieved by other approaches. We have also examined various related approaches previously proposed in the literature and we have compared them with ours from various points of views.

In the future we plan to investigate various research issues related to those presented here. First, we plan to develop approaches for deriving other typologies of interschema properties. Specifically, we would like to derive complex knowledge patterns involving a large variety of concepts belonging to different XML Schemas; in this application context we plan to exploit data mining techniques. After this, we would like to define new approaches for exploiting the properties considered in this paper, as well as those we shall study in the future, in the various application contexts where interschema properties can generally play a key role.

Finally, we would like to put the system described here as a part of a more complex system whose purpose is the extraction of intensional knowledge from semantically heterogeneous XML sources and its exploitation for handling their interoperability.

## References

- [1] S. Bergamaschi, S. Castano, and M. Vincini. Semantic integration of semistructured and structured data sources. *SIGMOD Record*, 28(1):54–59, 1999.
- [2] J. Berlin and A. Motro. Autoplex: Automated discovery of content for virtual databases. In *Proc. of the International Conference on Cooperative Information Systems (CoopIS 2001)*, pages 108–122, Trento, Italy, 2001. Lecture Notes in Computer Science, Springer.
- [3] J. Berlin and A. Motro. Database schema matching using machine learning with feature selection. In *Proc. of the International Conference on Advanced Information Systems Engineering (CAiSE 2002)*, pages 452–466, Toronto, Canada, 2002. Lecture Notes in Computer Science, Springer.
- [4] E. Bertino, G. Guerrini, and M. Mesiti. A matching algorithm for measuring the structural similarity between an XML document and a DTD and its applications. *Information Systems*, 29(1):23–46, 2004.
- [5] P. De Meo, G. Quattrone, G. Terracina, and D. Ursino. Integration of XML Schemas at various “severity” levels. *Information Systems*, 31(6):397–434, 2006.
- [6] H. Do, S. Melnik, and E. Rahm. Comparison of schema matching evaluations. In *Proc. of the International Workshop on Web, Web-Services, and Database Systems*, pages 221–237, Erfurt, Germany, 2002. Lecture Notes in Computer Science, Springer.

- [7] H. Do and E. Rahm. COMA- a system for flexible combination of schema matching approaches. In *Proc. of the International Conference on Very Large Databases (VLDB 2002)*, pages 610–621, Hong Kong, China, 2002. VLDB Endowment.
- [8] A. Doan, P. Domingos, and A. Halevy. Reconciling schemas of disparate data sources: a machine-learning approach. In *Proc. of the ACM International Conference on Management of Data (SIGMOD 2001)*, pages 509–520, Santa Barbara, California, USA, 2001. ACM Press.
- [9] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy. Learning to match ontologies on the Semantic Web. *The International Journal on Very Large Databases*, 12(4):303–319, 2003.
- [10] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Learning to map between ontologies on the Semantic Web. In *Proc. of the ACM International Conference on World Wide Web (WWW 2002)*, pages 662–673, Honolulu, Hawaii, USA, 2002. ACM Press.
- [11] A. Gal, A. Anaby-Tavor, A. Trombetta, and D. Montesi. A framework for modeling and evaluating automatic semantic reconciliation. *The International Journal on Very Large Databases*, 14(1):50–67, 2005.
- [12] Z. Galil. Efficient algorithms for finding maximum matching in graphs. *ACM Computing Surveys*, 18:23–38, 1986.
- [13] F. Giunchiglia, P. Shvaiko, and M. Yatskevich. S-Match: an Algorithm and an Implementation of Semantic Matching. In *Proc. of the European Semantic Web Symposium (ESWS'04)*, pages 61–75, Heraklion, Crete, Greece, 2004. Springer, Lecture Notes in Computer Science.
- [14] B. He and K. Chen-Chuan Chang. Statistical schema matching across Web query interfaces. In *Proc. of the ACM International Conference on Management of Data (SIGMOD 2003)*, pages 217–228, San Diego, California, United States, 2003. ACM Press.
- [15] M.L. Lee, L.H. Yang, W. Hsu, and X. Yang. XClust: clustering XML schemas for effective integration. In *Proc. of the ACM International Conference on Information and Knowledge Management (CIKM 2002)*, pages 292–299, McLean, Virginia, USA, 2002. ACM Press.
- [16] W. Li and C. Clifton. SEMINT: A tool for identifying attribute correspondences in heterogeneous databases using neural networks. *Data and Knowledge Engineering*, 33(1):49–84, 2000.
- [17] J. Madhavan, P.A. Bernstein, and E. Rahm. Generic schema matching with Cupid. In *Proc. of the International Conference on Very Large Data Bases (VLDB 2001)*, pages 49–58, Roma, Italy, 2001. Morgan Kaufmann.
- [18] S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity Flooding: A versatile graph matching algorithm and its application to schema matching. In *Proc. of the IEEE International Conference on Data Engineering (ICDE 2002)*, pages 117–128, San Jose, California, USA, 2002. IEEE Computer Society Press.
- [19] A.G. Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41, 1995.
- [20] L. Palopoli, D. Saccà, G. Terracina, and D. Ursino. Uniform techniques for deriving similarities of objects and subschemes in heterogeneous databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(2):271–294, 2003.
- [21] L. Palopoli, G. Terracina, and D. Ursino. Experiences using DIKE, a system for supporting cooperative information system and data warehouse design. *Information Systems*, 28(7):835–865, 2003.
- [22] K. Passi, L. Lane, S.K. Madria, B.C. Sakamuri, M.K. Mohania, and S.S. Bhowmick. A model for XML Schema integration. In *Proc. of the International Conference on E-Commerce and Web Technologies (EC-Web 2002)*, pages 193–202, Aix-en-Provence, France, 2002. Lecture Notes in Computer Science, Springer.
- [23] E. Rahm and P.A. Bernstein. A survey of approaches to automatic schema matching. *VLDB Journal*, 10(4):334–350, 2001.
- [24] V. C. Storey. Understanding semantic relationships. *The International Journal on Very Large Databases*, 2(4):455–488, 1993.
- [25] C.J. Van Rijsbergen. *Information Retrieval*. Butterworth, 1979.

# Preliminary Numerical Experiments in Multiobjective Optimization of a Metallurgical Production Process

Bogdan Filipič and Tea Tušar  
 Department of Intelligent Systems  
 Jožef Stefan Institute  
 Jamova 39, SI-1000 Ljubljana, Slovenia  
 E-mail: bogdan.filipic@ijs.si, tea.tusar@ijs.si

Erkki Laitinen  
 Department of Mathematical Sciences  
 University of Oulu  
 P.O. Box 3000, FIN-90014 Oulu, Finland  
 E-mail: erkki.laitinen@oulu.fi

**Keywords:** continuous casting of steel, product quality, process parameters, multiobjective optimization, differential evolution

**Received:** November 3, 2006

*This paper reports on preliminary numerical experiments in optimizing coolant flows in continuous casting of steel with respect to multiple objectives. For this purpose, Differential Evolution for Multiobjective Optimization (DEMO) coupled with a reliable numerical simulator of the casting process was applied. The algorithm parameters were initially tuned to balance between the quality of the expected results and the computational cost of the optimization process. Afterwards, suitable sets of coolant flow settings were calculated under conflicting requirements for minimum temperature deviations and predefined core length in the caster. In contrast to solutions produced in single-objective optimization, approximation sets of Pareto optimal fronts obtained in multiobjective optimization provide more information to metallurgists and allow for better insight into the casting process.*

*Povzetek: Članek obravnava nastavljanje pretokov hladila v industrijskem kontinuiranem ulivanju jekla kot večkriterijski optimizacijski problem in ga rešuje z evolucionim algoritmom DEMO.*

## 1 Introduction

Production and processing of materials are nowadays under strong market-driven pressure for shortening the process development time, reducing experimental costs, improving material properties, and increasing productivity. In achieving these goals, numerical analysis is playing an increasingly important role. Material scientists and engineers actually consider empirical knowledge and computational approximation as the basis for material process design and control. Numerical simulators give insight into process evolution, allow for execution of virtual experiments and support manual optimization by trial and error. However, the optimization procedure can be automated by coupling a simulator with an optimization algorithm and introducing a quality function which allows for automatic assessment of the simulation results.

Continuous casting of steel is an example of a process to which novel computational approaches have been applied intensively over the last years to enhance product characteristics and minimize production costs. In this complex metallurgical process molten steel is cooled and shaped into

semi-manufactures. To cast high quality steel, it is important to properly control the metal flow and heat transfer during the process. They depend on numerous parameters, including the casting temperature, casting speed and coolant flows. Finding optimal values of process parameters is difficult as the number of possible parameter settings is high, the involved criteria are often conflicting, and parameter tuning through real-world experimentation is not feasible because of safety risk and high costs. Techniques applied to overcome these difficulties include knowledge-based techniques, neural networks, fuzzy logic and evolutionary computation. Nevertheless, the predominant optimization approach taken in the applied studies so far was to aggregate multiple criteria into a single cost value and solve the optimization problem empirically using the simulator-optimizer coupling.

In this paper we report on preliminary numerical experiments in optimizing secondary coolant flows on a steel casting machine with respect to multiple objectives and under technological constraints. The experiments were performed using a novel multiobjective optimization evolutionary algorithm, while in the underlying numerical sim-



ulations continuous casting of a selected steel grade under steady-state conditions was assumed. Through the obtained approximation sets of optimal solutions the plant engineers can get better insight into process behavior and parameter effects.

The paper outlines the related work, describes the optimization task and the multiobjective optimization approach, and reports on the performed numerical experiments and obtained results.

## 2 Related Work

Over the last years, several advanced computer techniques have been used in attempts to enhance the process performance and material properties in metallurgical production. Cheung and Garcia [3], for example, combine a numerical model of the process with an artificial intelligence heuristic search technique linked to a knowledge base to find parameters values that result in defect-free billet production. Chakraborti and coworkers [1] report that genetic algorithms have proved to be the most suitable for optimizing the settings of the continuous casting mold. They use a Pareto-converging genetic algorithm to solve a multiobjective problem of setting the casting velocity in the mold region. In a further study [2] relying on heat transfer modeling, genetic algorithms are used to determine the maximum casting speed and solidified shell thickness at the mold exit. Oduguwa and Roy [13] use a novel fuzzy fitness evaluation in evolutionary optimization and apply it in rod rolling optimization. They solve a multi-objective problem of optimal rod shape design.

Our approach to process parameter optimization in continuous casting of steel involves a numerical simulator of the casting process and various stochastic optimization techniques among which evolutionary algorithms play the key role. The initial version of the optimization system [11] was designed to search for process parameter values that would result in as high as possible quality of continuously cast steel. Based on empirical metallurgical criteria, it was able to deliver improved parameter settings that proved beneficial in practice. However, using a simple evolutionary algorithm, it spent thousands of process simulations to find high-quality solutions. As the time aspect is critical, the purpose of further exploration [9, 7] was to reduce the number of needed process simulations. These applied studies were all using the weighted-sum technique of aggregating multiple criteria into a scalar cost function. As opposed to that, in a recent work [10] an attempt was made to handle multiple criteria by means of evolutionary multiobjective optimization. Based on the initial findings, this paper refines the problem definition by introducing an additional technological constraint, justifies the algorithm settings by checking the algorithm performance metrics and analyzes the new numerical results.

## 3 Problem Description

In industrial continuous casting, liquid steel is poured into a bottomless mold which is cooled with internal water flow. The cooling in the mold extracts heat from the molten steel and initiates the formation of a solid shell. The shell formation is crucial for the support of the slab behind the mold exit. The slab then enters the secondary cooling area in which it is cooled by water sprays. The secondary cooling region is divided into cooling zones where the amount of the cooling water can be controlled separately.

We consider a casting machine with the secondary cooling area divided into nine zones. In each zone, cooling water is dispersed to the slab at the center and corner positions. Target temperatures are specified for the slab center and corner in every zone. Water flows should be tuned in such a way that the resulting slab surface temperatures match the target temperatures as closely as possible. From metallurgical practice this is known to reduce cracks and inhomogeneities in the structure of the cast steel. Formally, cost function  $c_1$  is introduced to measure deviations of actual temperatures from the target ones:

$$c_1 = \sum_{i=1}^{N_Z} |T_i^{\text{center}} - T_i^{\text{center}*}| + \sum_{i=1}^{N_Z} |T_i^{\text{corner}} - T_i^{\text{corner}*}|, \quad (1)$$

where  $N_Z$  denotes the number of zones,  $T_i^{\text{center}}$  and  $T_i^{\text{corner}}$  the slab center and corner temperatures in zone  $i$ , and  $T_i^{\text{center}*}$  and  $T_i^{\text{corner}*}$  the respective target temperatures in zone  $i$ .

There is also a requirement for core length,  $l^{\text{core}}$ , which is the distance between the mold exit and the point of complete solidification of the slab. The target value for the core length,  $l^{\text{core}*}$ , is prespecified, and the actual core length should be as close to it as possible. Shorter core length may result in unwanted deformations of the slab as it solidifies too early, while longer core length may threaten the process safety. We formally treat this requirement as cost function  $c_2$ :

$$c_2 = |l^{\text{core}} - l^{\text{core}*}|. \quad (2)$$

The optimization task is to minimize both  $c_1$  and  $c_2$  over possible cooling patterns (water flow settings). It is known that the two objectives are conflicting, hence it is reasonable to handle this optimization problem as a multiobjective one.

In search for solutions, water flows cannot be set arbitrarily, but according to the technological constraints. For each zone, minimum and maximum values are prescribed for the center and corner water flows. Moreover, to avoid unacceptable deviations of the core length from the target value, a hard constraint is imposed:  $c_2 \leq \Delta l_{\text{max}}^{\text{core}}$ . Candidate solutions not satisfying the water flow constraint and/or the core length constraint are considered infeasible.

A prerequisite for optimization of this process is an accurate numerical simulator, capable of calculating the temperature field in the slab as a function of process parameters and evaluating it with respect to cost functions (1) and (2).

For this purpose we used the mathematical model of the process with Finite Element Method (FEM) discretization of the temperature field and the corresponding nonlinear equations solved with relaxation iterative methods, already applied in previous single-objective optimization study of the casting process [8].

## 4 Multiobjective Optimization

### 4.1 Preliminaries

The multiobjective optimization problem (MOP) is defined as finding the minimum of the cost function  $\mathbf{c}$ :

$$\mathbf{c}: X \rightarrow Z$$

$$\mathbf{c}: (x_1, \dots, x_n) \mapsto (c_1(x_1, \dots, x_n), \dots, c_m(x_1, \dots, x_n)),$$

where  $X$  is an  $n$ -dimensional decision space, and  $Z \subseteq \mathbb{R}^m$  is an  $m$ -dimensional objective space ( $m \geq 2$ ). The objective vectors from  $Z$  can be partially ordered using the concept of *Pareto dominance*:  $\mathbf{z}^1$  dominates  $\mathbf{z}^2$  ( $\mathbf{z}^1 \prec \mathbf{z}^2$ ) iff  $\mathbf{z}^1$  is not worse than  $\mathbf{z}^2$  in all objectives and better in at least one objective. When the objectives are conflicting, there exists a set of optimal objective vectors called *Pareto optimal front*. Each vector from the Pareto optimal front represents a different trade-off between the objectives and without additional information no vector can be preferred to another.

With a multiobjective optimizer we search for an *approximation set* that approximates the Pareto optimal front as well as possible. When solving MOPs in practice it is often important to provide the user with a diverse choice of trade-offs. Therefore, beside including vectors close to the Pareto optimal front, the approximation set should also contain near-optimal vectors that are as distinct as possible.

### 4.2 The DEMO Algorithm

Finding a good approximation set in a single run requires a population-based method. Consequently, evolutionary algorithms have been frequently used as multiobjective optimizers [4]. Among them, the recently proposed Differential Evolution for Multiobjective Optimization (DEMO) [15] is applied in optimizing the described metallurgical process.

DEMO is based on Differential Evolution (DE) [14], an evolutionary algorithm for single-objective optimization that has proved to be very successful in solving numerical optimization problems. In DE, each solution is encoded as an  $n$ -dimensional vector. New solutions, also called candidates, are constructed using operations such as vector addition and scalar multiplication. After the creation of a candidate, the candidate is compared with its parent and the best of them remains in the population, while the other one is discarded.

Because the objective space in MOPs is multidimensional, DE needs to be modified to deal with multiple objectives. DEMO is a modification of DE with a particular

mechanism for deciding which solution should remain in the population. For each parent in the population, DEMO constructs the candidate solution using DE. If the candidate dominates the parent, it replaces the parent in the current population. If the parent dominates the candidate, the candidate is discarded. Otherwise, if the candidate and its parent are incomparable, the candidate is added to the population. After constructing candidates for each parent individual in the population, the population has possibly increased. In this case, it is truncated to the original size using nondominated sorting and crowding distance metric (as in NSGA-II [5]). These steps are repeated until a stopping criterion is met.

DEMO is a simple but powerful algorithm, presented in detail in [15]. From the three proposed algorithm variants, the elementary one, called DEMO/parent, is used in this work.

## 5 Optimization Experiments

### 5.1 Experimental Setup

Numerical experiments in multiobjective optimization of the casting process were performed for a selected steel grade with the slab cross-section of 1.70 m  $\times$  0.21 m. Candidate solutions were encoded as 18-dimensional real-valued vectors, representing water flow values at the center and corner positions in 9 zones of the secondary cooling area. Search intervals for cooling water flows at both center and corner positions in zones 1, 2 and 3 were between 0 and 50 m<sup>3</sup>/h, while in the zones 4–9 between 0 and 10 m<sup>3</sup>/h. Table 1 shows the prescribed target slab surface temperatures. The target value for the core length  $l^{\text{core*}}$  was 27 m, while its maximum deviation allowed  $\Delta l_{\text{max}}^{\text{core}}$  was 7 m.

Table 1: Target surface temperatures in °C.

Zone number	Center position	Corner position
1	1050	880
2	1040	870
3	980	810
4	970	800
5	960	790
6	950	780
7	940	770
8	930	760
9	920	750

Four instances of the optimization problem were used in experiments, differing in the casting speed. The casting speed reflects the conditions under which the process needs to be conducted and significantly affects the productivity and product quality. In each problem instance the speed was kept constant, but at a different value. The values used were: 1.2 m/min, 1.4 m/min, 1.6 m/min and 1.8 m/min.

DEMO was integrated with the numerical simulator of the casting process into an automated optimization environment. DEMO evolved sets of candidate solutions in search for a good approximation set, and the simulator served as a solution evaluator. Steady-state operation of the casting machine was assumed and optimization performed in the off-line manner.

The most limiting factor for experimental analysis is the computational complexity of the casting process simulation. A single simulator run takes about 40 seconds on a 1.8 GHz Pentium IV computer. In initial experimentation we found DEMO runs with 5000 solution evaluations (and therefore taking about 55 hours) well compromising between the execution time and solution quality. Further algorithm settings were also adopted according to the initial parameter tuning experiments [6] and were as follows: population size 50, number of generations 100, scaling factor 0.5 and crossover probability 0.05. These settings ensure highly acceptable algorithm performance and repeatability of the results as indicated by the hypervolume measure [16] and attainment surface plots [12] obtained over five test runs of the algorithm and shown in Figs. 1–2.

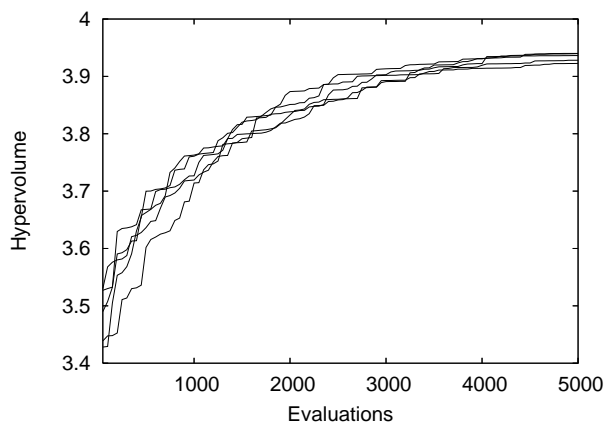


Figure 1: Hypervolume values in five test runs of the DEMO algorithm.

## 5.2 Results and Findings

The key result of this study were approximation sets of Pareto optimal fronts. Figure 3 shows the approximation sets found by DEMO for five casting speeds, ranging from 1.2 m/min to 1.8 m/min. Each set of nondominated solutions is the final result of a single DEMO run at a constant casting speed.

It can be observed that the two objectives are really conflicting in the sense that finding a minimum for one of them the optimization procedure fails to do so for the other and vice versa. It is also obvious that the casting speed has a decisive impact on the result. Moreover, the higher the casting speed, the more the two objectives can be met simultaneously. This corresponds with practical experience on the considered casting machine, where the process is easier to

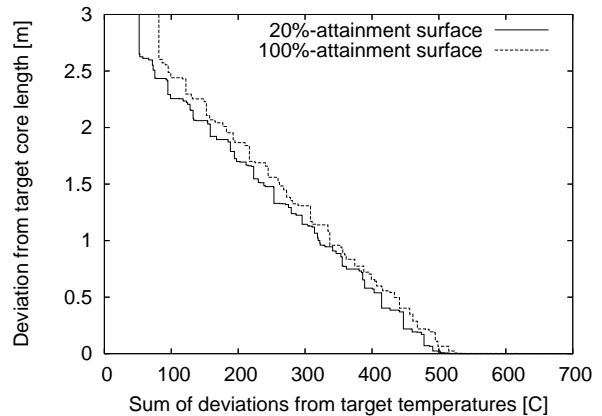


Figure 2: 20% and 100% attainment surfaces for the solutions found in five test runs of the DEMO algorithm.

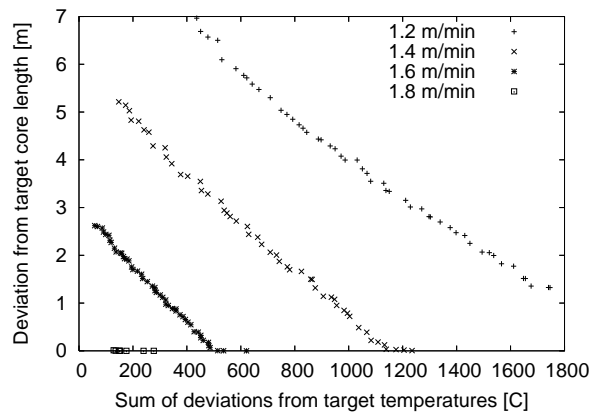


Figure 3: Nondominated solutions found with DEMO for different casting speeds.

control at the usual casting speed (1.6–1.8 m/min). Lower casting speed is clearly shown as disadvantageous and in practice it is only set exceptionally, for example, when a new batch of steel is awaited.

A detailed analysis of the solution properties also reveals that, in view of the objective  $c_1$ , the majority of actual surface temperatures are higher than the target temperatures, while regarding  $c_2$ , the actual core length is almost always shorter than the target value.

Looking into decision space, one can also observe certain regularities. In case of applying trade-off solutions from the middle of the approximation sets, the amount of coolant spent increases with the casting speed (see the left-hand side diagrams in Figs. 4–7). This is an expected result as higher casting speed implies more intense cooling. On the other hand, the distributions of temperature differences across the secondary cooling zones (right-hand side diagrams in Figs. 4–7) exhibit two characteristics. First, the target temperatures are much more difficult to achieve at the center than in the corner slab positions. Second, the differences at the center are rather non-uniform. While some are close to zero, others reach up to 200°C at lower cast-

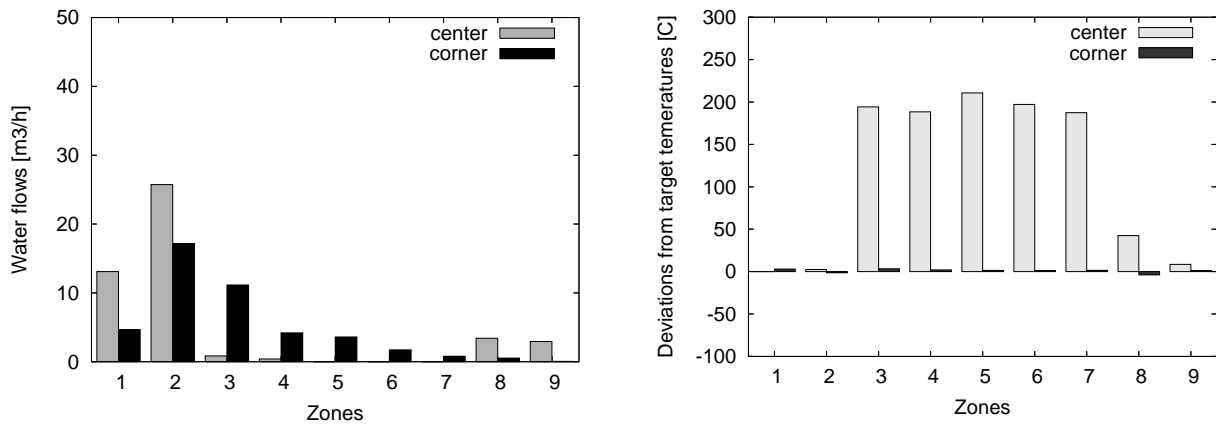


Figure 4: A trade-off solution from the middle of the approximation set for the casting speed speed of 1.2 m/min:  $c_1 = 1051^\circ\text{C}$ ,  $c_2 = 3.8$  m.

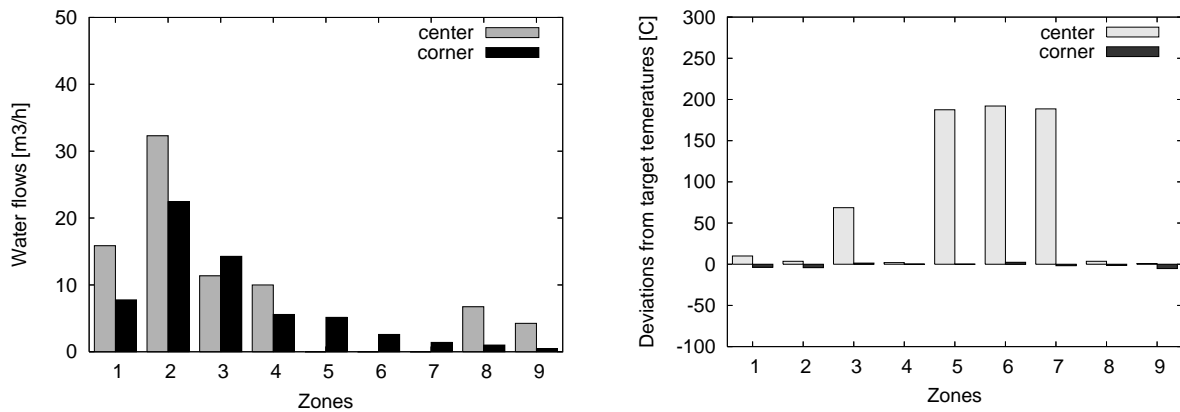


Figure 5: A trade-off solution from the middle of the approximation set for the casting speed speed of 1.4 m/min:  $c_1 = 677^\circ\text{C}$ ,  $c_2 = 2.2$  m.

ing speeds. Such a situation is not preferred in practice and calls for the reformulation of objective  $c_1$  in further calculations.

On the other hand, it is worth checking the extreme solutions from an approximation set at a given casting speed. Figures 8 and 9 clearly show how one objective is met at the expense of the other. None of these would normally be used in practice. Instead, a plant engineer would rather select a trade-off setting balancing between the two objectives.

## 6 Conclusion

Advanced manufacturing and processing of materials strongly rely on numerical analysis of the related processes made possible by powerful modeling and simulation software packages. To use them efficiently, an upgrade is needed towards process automatic optimization. The optimization environment studied in this paper consists of a numerical process simulator and an evolutionary multiob-

jective optimization algorithm. We illustrated the capabilities of this approach in process parameter optimization in continuous casting of steel. Solving this task successfully is a key to higher product quality.

In the preliminary study of optimizing 18 cooling water flows with respect to two objectives on an industrial casting machine the capabilities of the multiobjective problem treatment were shown. The analysis assumes steady-state process conditions, hence the results are not primarily intended for control purposes but rather for better understanding of the process and evaluation of the casting machine performance. The resulting approximation sets of Pareto optimal fronts indeed offer a more general view of the process properties. The results support some facts already known in practice and, at the same time, show critical points, such as the need to reformulate the temperature deviation criterion to ensure uniform distribution of temperature differences over the zones, and extend the optimization problem definition with an additional constraint. From the practical point of view, further studies will also explore

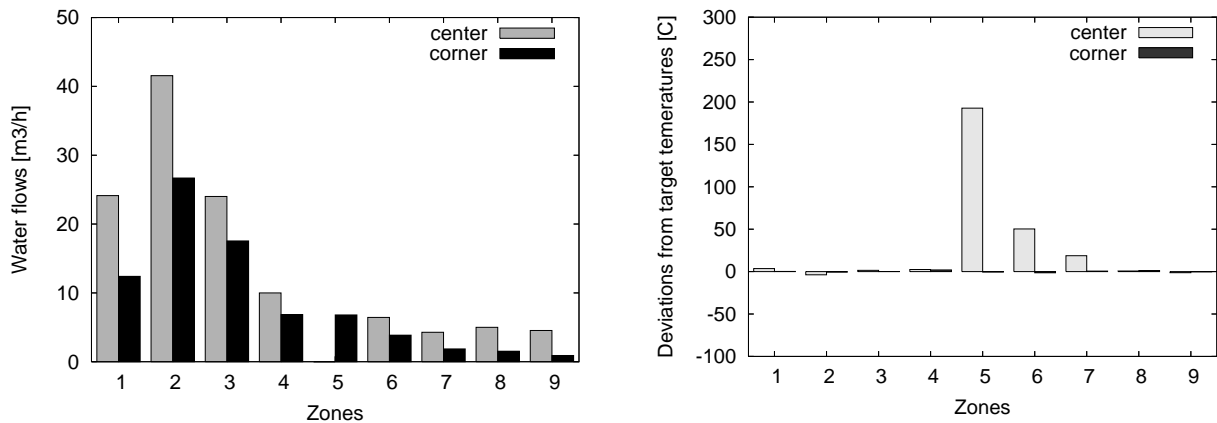


Figure 6: A trade-off solution from the middle of the approximation set for the casting speed speed of 1.6 m/min:  $c_1 = 281^\circ\text{C}$ ,  $c_2 = 1.3$  m.

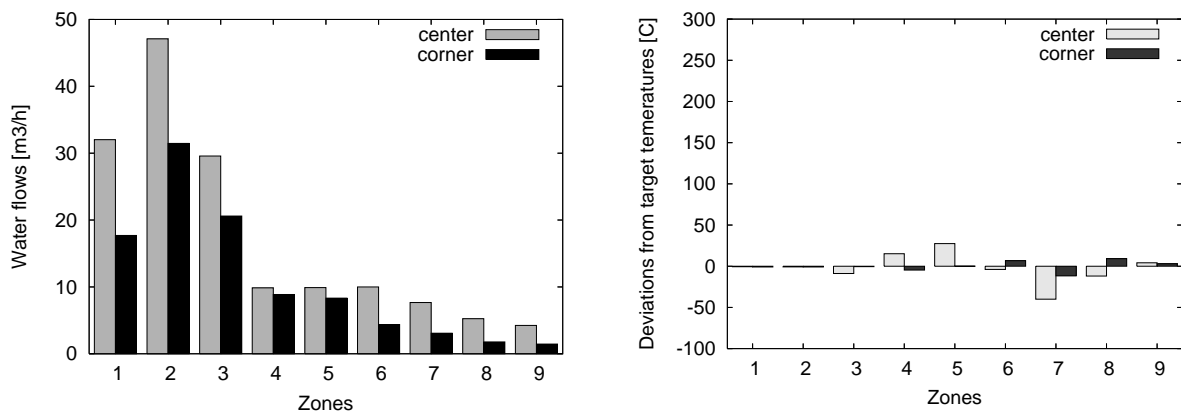


Figure 7: A trade-off solution from the middle of the approximation set for the casting speed speed of 1.8 m/min:  $c_1 = 151^\circ\text{C}$ ,  $c_2 = 0.0$  m.

how much the optimization results are affected by the factors that were kept constant so far, such as steel grade, slab geometry and casting machine characteristics.

## Acknowledgment

This work was supported by the Slovenian Research Agency and the Academy of Finland under the Slovenian-Finnish project BI-FI/04-05-009 *Numerical Optimization of Continuous Casting of Steel*, and by the Slovenian Research Agency under the Research Programme P2-0209 *Artificial Intelligence and Intelligent Systems*.

## References

- [1] N. Chakraborti, R. Kumara and D. Jain. A study of the continuous casting mold using a Pareto-converging genetic algorithm. *Applied Mathematical Modelling*, 25 (4): 287–297, 2001.
- [2] N. Chakraborti, R. S. P. Gupta and T. K. Tiwari. Optimisation of continuous casting process using genetic algorithms: studies of spray and radiation cooling regions. *Ironmaking and Steelmaking*, 30 (4): 273–278, 2003.
- [3] N. Cheung and A. Garcia. The use of a heuristic search technique for the optimization of quality of steel billets produced by continuous casting. *Engineering Applications of Artificial Intelligence*, 14 (2): 229–238, 2001.
- [4] K. Deb. *Multi-Objective Optimization Using Evolutionary Algorithms*. John Wiley & Sons, Chichester, UK, 2001.
- [5] K. Deb, A. Pratap, S. Agrawala and T. Meyarivan. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Transactions on Evolutionary Computation*, 6 (2): 182–197, 2002.

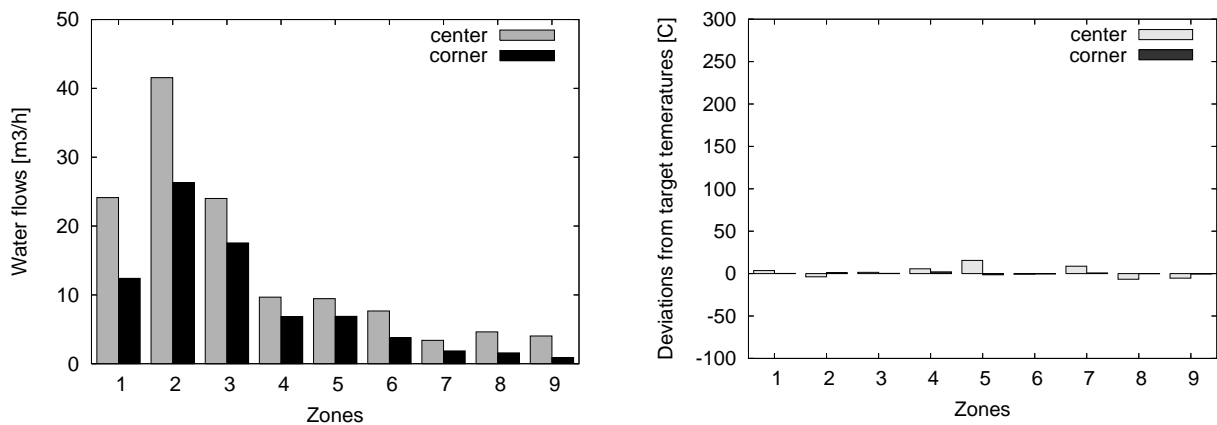


Figure 8: The leftmost solution from the approximation set for the casting speed speed of 1.6 m/min:  $c_1 = 58^\circ\text{C}$ ,  $c_2 = 2.6$  m.

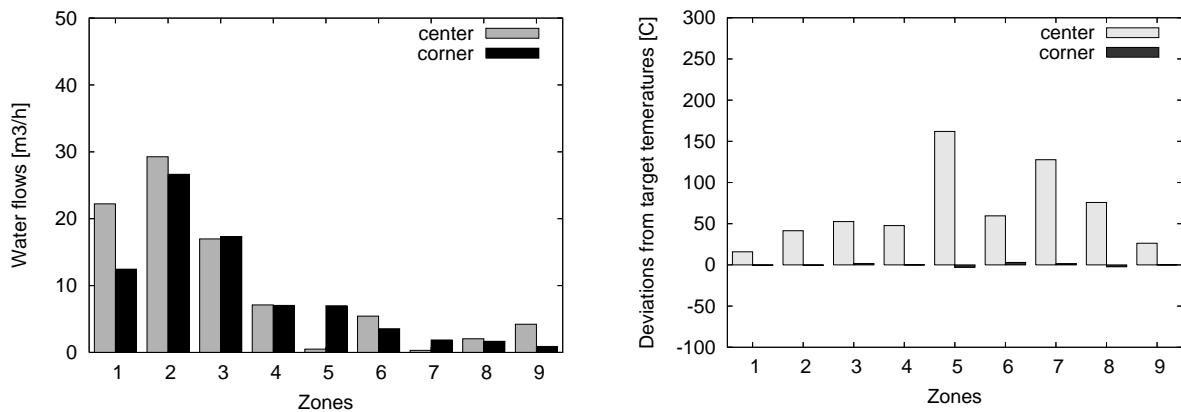


Figure 9: The rightmost solution from the approximation set for the casting speed speed of 1.6 m/min:  $c_1 = 620^\circ\text{C}$ ,  $c_2 = 0.0$  m.

- [6] M. Depolli, T. Tušara and B. Filipič. Tuning parameters of a multiobjective optimization evolutionary algorithm on an industrial problem. In B. Zajc and A. Trost (Eds.), *Proceedings of the Fifteenth International Electrotechnical and Computer Science Conference ERK 2006*, Vol. B, pages 95–98, Portorož, Slovenia, 2006. Slovenian Section IEEE, Ljubljana (in Slovenian).
- [7] B. Filipič. Efficient simulation-based optimization of process parameters in continuous casting of steel. In D. Büche, N. Hofmann (Eds.), *COST 526: Automatic Process Optimization in Materials Technology: First Invited Conference*, pages 193–198, Morschach, Switzerland, 2005.
- [8] B. Filipič and E. Laitinen. Model-based tuning of process parameters for steady-state steel casting. *Informatica*, 29 (4): 491–496, 2005.
- [9] B. Filipič and T. Robič. A comparative study of coolant flow optimization on a steel casting machine. In *Proceedings of the 2004 Congress on Evolutionary Computation CEC 2004*, Vol. 1, pages 569–573, Portland, OR, USA, 2004. IEEE, Piscataway.
- [10] B. Filipič, T. Tusar and E. Laitinen. Computer-assisted analysis of a metallurgical production process in view of multiple objectives. In B. Filipič, J. Šilc (Eds.), *Proceedings of the Second International Conference on Bioinspired Optimization Methods and their Applications BIOMA 2006*, pages 167–176. Jožef Stefan Institute, Ljubljana, Slovenia, 2006.
- [11] B. Filipič and B. Šarler. Evolving parameter settings for continuous casting of steel. In *Proceedings of the 6th European Conference on Intelligent Techniques and Soft Computing EUFIT'98*, Vol. 1, pages 444–449, Aachen, Germany, 1998. Verlag Mainz, Aachen.
- [12] C. M. Fonseca and P. J. Fleming. On the Performance Assessment and Comparison of Stochastic Multiobjective Optimizers. In W. Ebeling, I. Rechenberg, H.-P. Schwefel, H.-M. Voigt (Eds.), *Parallel Prob-*

- lem Solving from Nature PPSN IV, 4th International Conference*, Lecture Notes in Computer Science, Vol. 1141, pages 584–593, Berlin, 1996. Springer.
- [13] V. Oduguwa and R. Roy. Multiobjective optimization of rolling rod product design using meta-modeling approach. In *Proceedings of the Genetic and Evolutionary Computation Conference GECCO 2002*, pp. 1164–1171. Morgan Kaufmann, San Francisco, CA, 2002.
- [14] K. V. Price and R. Storn. Differential evolution – a simple evolution strategy for fast optimization. *Dr. Dobbs's Journal*, 22 (4): 18–24, 1997.
- [15] T. Robič and B. Filipič. DEMO: Differential evolution for multiobjective optimization. In C. A. Coello Coello, A. Hernandez Aguirre, E. Zitzler (Eds.): *Proceedings of the Third International Conference on Evolutionary Multi-Criterion Optimization EMO 2005*, pages 520–533, Guanajuato, Mexico, 2005. Lecture Notes in Computer Science, Vol. 3410, Springer, Berlin.
- [16] E. Zitzler, L. Thiele, M. Laumanns, C. M. Fonseca and V. G. da Fonseca. Performance assessment of multiobjective optimizers: An analysis and review. *IEEE Transactions on Evolutionary Computation*, 7 (2): 117–132, 2003.



## Development of a Hungarian Medical Dictation System

András Bánhalmi, Dénes Paczolay, László Tóth and András Kocsor<sup>†</sup>  
 Research Group on Artificial Intelligence  
 Hungarian Academy of Sciences and the University of Szeged  
 Aradi vértanúk tere 1, H-6720 Szeged  
 {banhalmi, pdenes, tothl, kocsor}@inf.u-szeged.hu

<sup>†</sup> Applied Intelligence Laboratory Ltd. and  
 Research Group on Artificial Intelligence NPC,  
 Petőfi Sgt. 43, H-6723 Szeged, Hungary

**Keywords:** speech recognition, dictation systems, 2D-cepstrum

**Received:** May 12, 2004

*This paper reviews the current state of a Hungarian project which seeks to create a speech recognition system for the dictation of thyroid gland medical reports. First, we present the MRBA speech corpus that was assembled to support the training of general-purpose Hungarian speech recognition systems. Then we describe the processing of medical reports that were collected to help the creation of domain-specific language models. At the acoustic modelling level we experimented with two techniques – a conventional HMM one and an ANN-based solution – which are both briefly described in the paper. Finally, we present the language modelling methodology currently applied in the system, and round off with recognition results on test data taken from four speakers. The scores show that on a somewhat restricted sub-domain of the task we are able to produce word accuracies well over 95%.*

*Povzetek: Prispevek predstavlja pregled trenutnega stanja madžarskega projekta, ki skuša vzpostaviti sistem razpoznavanja govora za narekovanje zdravniških izvidov na temo žleze ščitnice.*

### 1 Introduction: state of the art and goals of the project

Automating the dictation of texts is one of the main applications of speech recognition. Mainly because of the huge training corpora, the increased processor speeds and the refined search techniques dictation systems have reached such a level of sophistication that the commercial products now offer sufficiently good accuracy even for arbitrary normal-pace fluent speech [12]. Experience tells us, however, that for a really good performance it is still worth applying some tricks like an initial speaker enrollment process where the machine can adapt to the voice of the speaker, or the restriction of the dictation topic to some specific (e.g. medical or legal) domain. Such dictation systems already exist for the biggest languages, but the situation for those languages that can offer only a small market is not as good. For Hungarian at the present time there exists no general-purpose large vocabulary continuous speech recognizer (LVCSR). Among the university publications even papers that deal with continuous speech recognition are hard to find, and these give results only for restricted vocabularies [15]. Although on the industrial side Philips have adapted its SpeechMagic system to two special application domains in Hungarian, it is sold at a price that is affordable for only the largest institutes [9]. The experts

usually cite two main reasons for the lack of Hungarian LVCSR systems. First, there are no sufficiently large, publicly available speech databases that would allow the training of reliable phone models. The second reason is the special difficulties of language modelling that arise due to the highly agglutinative nature of Hungarian.

In 2004 the Research Group on Artificial Intelligence at the University of Szeged and the Laboratory of Speech Acoustics of the Budapest University of Technology and Economics began a project with the aim of collecting and/or creating the basic resources needed for the construction of a continuous dictation system for Hungarian. The project lasted for three years (2004-2006), and was financially supported by the national fund IKTA-056/2003. For the acoustic modelling part, the project included the collection and annotation of a large speech corpus of phonetically rich sentences. As regards the language modelling part, we restricted the target domain to the dictation of some limited types of medical reports. Although this clearly led to a significant reduction compared to a general dictation task, we chose this application area with the intent of assessing the capabilities of our acoustic and language modelling technologies. Depending on the findings, later we hope to extend the system to more general dictation domains. This is why the language resources were chosen to be domain-specific, while the acoustic database contains quite general,

domain-independent recordings.

Although both participating teams used the same speech database to train their acoustic models, they focused on two different dictation tasks and experimented with their own acoustic and language modelling technologies. The team at the University of Szeged focused on the task of the dictation of thyroid scintigraphy medical reports, while the Budapest team dealt with gastroenterology reports. This paper just describes the research and development efforts of the Szeged team. The interested reader can find a survey of the research done by the Laboratory of Speech Acoustics in [16].

## 2 Speech and language resources

In the first phase of the project we designed, assembled and annotated a speech database called the MRBA corpus (the abbreviation stands for the "Hungarian Reference Speech Database") [16]. Our goal was to create a database that allows the training of general-purpose dictation systems which run on personal computers in office environments and operate with continuous, read speech. The contents of the database were designed by the Laboratory of Speech Acoustics. As a starting point, they took a large (1.6 MB) text corpus and after automatic phonetic transcription they created phone, diphone and triphone statistics from it. Then they selected 1992 different sentences and 1992 different words in such a way that 98.8% of the most frequent diphones had at least one occurrence in them. These sentences and words were recorded from 332 speakers, each reading 12 sentences and 12 words. Thus all sentences and words have two recordings in the speech corpus. Both teams participated in the collection of the recordings, which was carried out in four big cities, mostly at universities labs, offices and home environments. In the database the ratio of male and female speakers is 57.5% to 42.5%. About one-third of the speakers were between 16 to 30 years of age, the rest being evenly distributed among the remaining age groups. Both home PCs and laptops were used to make the recordings, and the microphones and sound cards of course varied as well. The sound files were cleaned and annotated at the Laboratory of Speech Acoustics, while the Research Group on Artificial Intelligence manually segmented and labelled one third of the files at the phone level. This part of the corpus is intended to support the initialization of phone models prior to training on the whole corpus.

Besides the general-purpose MRBA corpus, we also collected recordings that are specific for the target domain, namely thyroid scintigraphy medical reports. From these recordings 20–20 reports read aloud by 4 persons were used as test data in the experiments reported here.

For the construction of the domain-specific language models, we got 9231 written medical reports from the Department of Nuclear Medicine of the University of Szeged. These thyroid scintigraphy reports were written and stored

using various software packages that were employed at the department during 1998 to 2004. So first of all we had to convert all the reports to a common format, followed by several steps of routine error correction. Each report consists of 7 fields: header (name, ID number etc. of the patient), clinical observations, request of the referral doctor, a summary of previous examinations (if any), the findings of this examination, a one-sentence summary, and a signature. From the corpus we omitted the first and the last, person-specific fields, for the sake of personal data privacy. Then we discarded those reports that were incomplete such as those that had missing fields. This way only 8546 reports were kept, which, on average, contained 11 sentences and 6 words per sentence. The next step was to remove any typographical errors from the database, of which there were surprisingly many (the most frequent words occurred in 10–15 mistyped forms). A special problem was that of unifying the Latin terms, many of which are allowed to be written both with a Latin or a Hungarian spelling in medical texts (for example *therapia* vs. *terápia*). The abbreviations also had to be resolved. The corpus we got after these steps contained approximately 2500 different word forms (excluding numbers and dates), so we were confronted with a medium-sized vocabulary dictation task.

## 3 The user interface

Our GUI was really designed with the goal of serving many users on the same computer. The other main design aspect was to combine simplicity with good functionality. With our program only a microphone and a text editor (Microsoft Word or a similar word processing program) are needed for dictating medical reports.

Every user has one or more profiles containing all the special information characterizing his or her voice for a given language and vocabulary. The language models and the acoustic core modules can be installed separately, and the system can optionally adapt to the individual characteristics of the users. The user interface basically consists of a toolbar at the top of the desktop. Using the toolbar all the main functionalities related to the initial parameter settings can be accessed, such as choosing a specific user, choosing the actual task and selecting the output window (Fig. 1). Other functionalities can only be accessed from the actual text editor. The most important of these features could be that the user can ask the speech recognition system for other possible variants of the recognized sentences in cases where he/she discovers the recognized word or sentence to be incorrect.

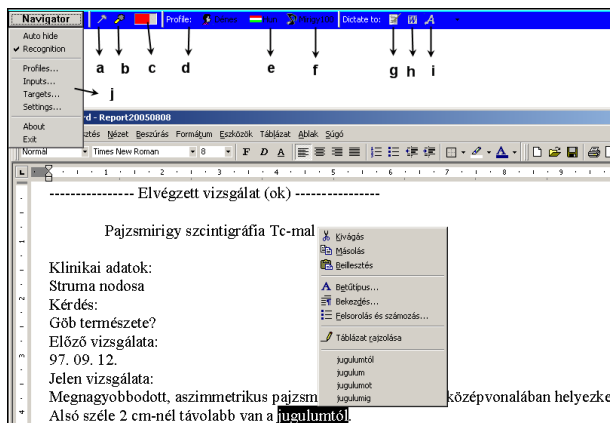


Figure 1: Functions of the graphical user interface: a) Enable or disable auto hiding of the main toolbar. b) Start or stop the recognition procedure. The user can suspend the dictation at any time, and can continue later. c) Volume display bar. The volume of the microphone input can be checked here. d) Choosing a specific user. The user can be selected from the list of existing users. e) Choosing the actual language. The language assigned to the current user can be chosen from a listbox. f) Choosing the actual grammar. Any available grammar can be chosen with just one click. g) Selecting the internal text editor. The recognized text will appear in the internal smart text editor. h) Selecting the Microsoft Word plugin for output. i) Selecting the window of the active application. With this function the user can dictate into any MS Windows-based application like MS Excel or MS Outlook. j) The main menu for managing the user profiles. The functions presented above can be accessed from here.

#### 4 Acoustic modelling I: HMM phone models over MFCC features

At the level of acoustic modelling we have been experimenting with two quite different technologies. One of these is a quite conventional Hidden Markov Model (HMM) decoder that works with the usual mel-frequency cepstral coefficient (MFCC) features [4]. More precisely, 13 coefficients are extracted from 25 msec frames, along with their  $\Delta$  and  $\Delta\Delta$  values, at a rate of 100 frames/sec. The phone models applied have the usual 3-state left-to-right topology. Hungarian has the special property that almost all phones have a short and a long counterpart, and their difference is phonologically relevant (i. e. there are word pairs that differ only in the duration of one phone – for example ‘tör’–‘tőr’ or ‘szál’–‘száll’) [14]. However, it is known that such minimal word pairs are relatively rare [14], and inspecting the vocabulary of our specific dictation task we found no such words. Hence most of the long/short consonant labels were fused, and this way we worked with just 44 phone classes. One phone model was associated with each

of these classes, that is we applied monophone modelling and this far no context-dependent models were tested in the system. The decoder built on these HMM phone models performs a combination of Viterbi and multi-stack decoding [4]. For speed efficiency it contains several built-in pruning criteria. First, it applies beam pruning, so only the hypotheses with a score no worse than the best score minus a threshold are kept. Second, the number of hypotheses extended at every time point is limited, corresponding to multi-stack decoding with a stack size constraint. The maximal evaluated phone duration can also be fixed. With the proper choice of these parameters the decoder on a typical PC runs faster than real-time on the medical dictation task.

#### 5 Acoustic modelling II: HMM/ANN phone models over 2D-cepstrum features

Our alternative, more experimental acoustic model employs the HMM/ANN hybrid technology [2]. The basic difference between this and the standard HMM scheme is that here the emission probabilities are modelled by Artificial Neural Networks (ANNs) instead of the conventional Gaussian mixtures (GMM). In the simplest configuration one can train the neural net over the usual 39 MFCC coefficients – whose result can serve as a baseline for comparison with the conventional HMM. However, ANNs seem to be more capable of modelling the observation context than the GMM technology, so the hybrid models are usually trained over longer time windows. The easiest way of doing this is to specify a couple of neighboring frames as input to the net: in a typical arrangement 4 neighboring frames are used on both sides of the actual frame [2]. Another option is to apply some kind of transformation on the data block of several neighboring frames. Knowing that the modulation components play an important role in human speech perception, performing a frequency analysis over the feature trajectories seems reasonable. When this analysis is applied to the cepstral coefficients, the resulting feature set is usually referred to as the 2D-cepstrum [6]. Research shows that most of the linguistic information is in the modulation frequency components between 1 and 16 Hz, especially between 2 and 10 Hz. This means that not all of the components of a frequency analysis have to be retained, and so the 2D-cepstrum offers a compact representation of a longer temporal context.

In the experiments we tried to find the smallest feature set that would give the best recognition results. Running the whole recognition test for each parameter setting would have required too much time so, as a quick indicator of the efficiency of a feature set we used the frame-level classification score. Hence the values given in the following tables are frame-level accuracy values measured on a held-out data set of 20% of the training data.

First of all we tried to extend the data of the ‘target’ frame by neighboring frames, without applying any transformation. The results shown in Table 1 indicate that training on more than 5 neighboring frames significantly increased the number of features and hidden neurons (and also significantly raised the training time) without bringing any real improvement in the score.

Obs. size	Hidden neurons	Frame accuracy
1 frames	150	64.16%
3 frames	200	67.51%
5 frames	250	68.67%
7 frames	300	68.81%
9 frames	350	68.76%

Table 1: The effect of varying the observation context size.

In the experiments with the 2D-cepstrum we first tried to find the optimal size of the temporal window. Hence we varied the size of the DFT analysis between 8, 16, 32, and 64, always keeping the first and second components<sup>1</sup> (both the real and the imaginary parts), and combined these with the static MFCC coefficients. The results displayed in Table 2 indicate that the optimum is somewhere between 16 and 32 (corresponding to 160 and 320 milliseconds). This is smaller than the 400 ms value found optimal in [6] and the 310 ms value reported in [13], but this might depend on the amount of training data available (a larger database would cover more of the possible variations and hence would allow a larger window size). Of course, one could also experiment with combining various window sizes as Kanedera did [6], but we did not run such multi-resolution tests.

DFT size	Hidden neurons	Frame accuracy
8	200	64.63%
16	200	67.60%
32	200	67.01%
64	200	64.75%

Table 2: Frame-level results at various DFT sizes.

As the next step we examined whether it was worth retaining more components. In the case of the 16-point DFT we kept 3 components, while for the 32-point DFT we tried retaining 5 components (the highest center frequency being 18.75 Hz and 15.625 Hz, respectively). The results (see Table 3) show that the higher modulation frequency components are less useful, which accords with what is known about the importance of the various modulation frequencies.

Finally, we tried varying the type of transformation applied. Motlíček reported that there is no need to keep both the real and imaginary parts of the DFT coefficients; using

DFT Size	Components	H. n.	Frame acc.
16	1, 2, 3	250	68.40%
32	1, 2, 3, 4, 5	300	70.64%

Table 3: Frame-level results with more DFT components.

just one of them is sufficient. Also, he obtained a similar performance when replacing the complex DFT with the DCT [10]. Our findings agree more with those of Kanedera [6], that is we obtained slightly worse results with these modifications (see Table 4). Hence we opted for the complex DFT, using both the real and imaginary coefficients. One advantage of the complex DFT over the DCT might be that when only some of its coefficients are required (as in our case), it can be very efficiently computed using a recursive formula [5].

Transform	H. neurons	Frame accuracy
DFT Re + Im	300	70.64%
DFT Re only	220	65.81%
DCT	220	68.00%

Table 4: The effect of varying the transformation type.

## 6 Domain-specific language modelling

A special difficulty of creating language models for Hungarian is the highly agglutinative [3] nature of the language. This means that most words are formed by joining several morphemes together, and those modifications of the meaning that other languages express e.g. by pronouns or prepositions in Hungarian are handled by affixes (for example ‘in my house’ is ‘házamban’) [7]. Because of this, in a large vocabulary modelling task the application of a morphologic analyzer/generator seems inevitable. First, simply listing and storing all the possible word forms would be almost impossible (e.g. an average noun can have about 700 inflected forms). Second, if we simply handled all these inflected forms as different words, then achieving a certain coverage rate in Hungarian would require a text about 5 times bigger than that in German and 20 times bigger than that in English [11]. Hence the training of conventional  $n$ -gram models would require significantly larger corpora in Hungarian than in English, or even in German. A possible solution might be to train the  $n$ -grams over morphemes instead of word forms, but then again the handling of the morphology would be necessary.

Though decent morphological tools exist now for Hungarian, in our medical dictation system we preferred to avoid the complications incurred by morphology. In fact, the restricted vocabulary is one of the reasons why we opted for the medical dictation task. For, as we men-

<sup>1</sup>The DC offset being indexed as the zeroth component.

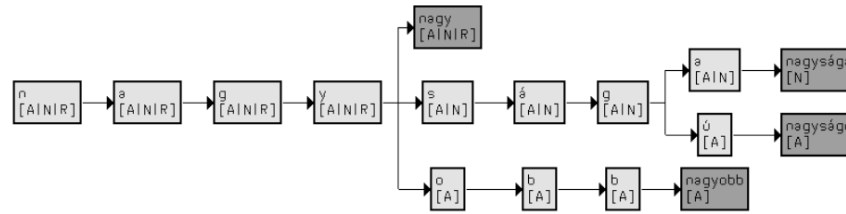


Figure 2: Prefix tree for some Hungarian words with their MSD code. At the branches of the tree the grammar model can generate the probability of a word based on the word  $n$ -gram and also based on the class  $n$ -gram.

tioned earlier, the thyroid gland medical reports contain only about 2500 different word forms. Although these many words could be easily managed even by a simple list ('linear lexicon'), we organized the words into a lexical tree where the common prefixes of the lexical entries are shared. Apart from storage reduction advantages, this representation also speeds up decoding, as it eliminates redundant acoustic evaluations [4]. A prefix tree representation is probably far more useful for agglutinative languages than for English because of the many inflected forms of the same stem.

The limited size of the vocabulary and the highly restricted (i.e. low-perplexity) nature of the sentences used in the reports allowed us to create very efficient  $n$ -grams. Moreover, we did not really have to worry about out-of-vocabulary words, since we had all the reports from the previous six years, so the risk of encountering unknown words during usage seemed minimal. The system currently applies 3-grams by default, but it is able to 'back off' to smaller  $n$ -grams (in the worst case to a small  $\epsilon$  constant) when necessary. During the evaluation of the  $n$ -grams the system applies a language model lookahead technique. This means that the language model returns its scores as early as possible, not just at word endings. For this reason the lexical trees are stored in a factored form, so that when several words share a common prefix, the maximum of their probabilities is associated with that prefix [4]. These techniques allow a more efficient pruning of the search space.

Besides word  $n$ -grams we also experimented with constructing class  $n$ -grams. For this purpose the words were grouped into classes according to their parts-of-speech category. The words were categorized using the POS tagger software developed at our university [8]. This software associates one or more MSD (morpho-syntactic description) code with the words, and we constructed the class  $n$ -grams over these codes. With the help of the class  $n$ -grams the language model can be made more robust in those cases when the word  $n$ -gram encounters an unknown word, so it practically performs a kind of language model smoothing. In previous experiments we found that the application of the language model lookahead technique and class  $n$ -grams brought about a 30% decrease in the word error rate when it was applied in combination with our HMM-based fast decoder [1]. Figure 2 shows an example of a prefix tree

storing four words, along with their MSD codes.

## 7 Experimental results and discussion

For testing purposes we recorded 20-20 medical reports from 2 male and 2 female speakers. The language model applied in the tests was constructed based on just 500 reports instead of all the 8546 we had collected. This subset contained almost all the sentence types that occur in the reports, so this restriction mostly reduced the dictionary by removing a lot of rarely occurring words (e.g. dates and disease names). Besides the HMM decoder we tested the HMM/ANN hybrid system in three configurations: the net being trained on one frame of data, on five neighboring frames, and on the best 2D-cepstrum feature set (static MFCC features plus 5 modulation components using a 32-point DFT with both *Re* and *Im* parts). The results are listed in Table 5 below. Comparing the first two lines, we see that when using the classic MFCC features the HMM and the HMM/ANN system performed quite similarly on the male speakers. For some reason, however, the HMM system did not like the set of female voices. The remaining rows of the table show that extending the net's input with an observation context – either by neighboring frames or by modulation features – brought only very modest improvements over the baseline results. We think the reason for this is that in the current arrangement the recognizer relies very strongly on the language model, thanks to the high predictability of the sentences. We suspect that the improvement in the acoustic modelling will be better seen in the scores when we apply the system to a linguistically less restricted domain. Pure phone recognition tests (i.e. recognition experiments with no language model support) that could verify this conjecture are just under development.

## 8 Conclusions

This paper reported the current state of a Hungarian project for the automated dictation of medical reports. We described the acoustic and linguistic training data collected and the current state of development in both the acoustic and linguistic modelling areas. Recognition results were

Model Type	Feature Set	Male 1	Male 2	Female 1	Female 2
HMM	MFCC + $\Delta$ + $\Delta\Delta$	97.75%	98.22%	93.40%	93.39%
HMM/ANN	MFCC + $\Delta$ + $\Delta\Delta$	97.65%	97.37%	96.78%	96.91%
HMM/ANN	5-frames * (MFCC + $\Delta$ + $\Delta\Delta$ )	97.65%	97.74%	96.67%	98.05%
HMM/ANN	MFCC + 5 Mod. Comp. (Re + Im)	97.88%	97.83%	96.86%	96.42%

Table 5: Word recognition accuracies of the various models and feature sets.

also given over a somewhat restricted subset of the full domain. For the next step we plan to extend the vocabulary and language model to cover all the available data, and then to test the system over other dictation domains as well. Our preliminary results indicate that for tasks over larger vocabularies several further improvements will be required. On the acoustic modelling side we intend to implement speaker adaptation and context-dependent models within the HMM system. We also plan to continue our research on observation context modelling within the HMM/ANN system. Finally, the language model will also need to be improved in many respects, especially when handling certain special features like dates and abbreviations.

## References

- [1] A. Bánhalmi, A. Kocsor, and D. Paczolay. 2005. Supporting a Hungarian dictation system with novel language models (in Hungarian). In: *Proc. of the 3rd Hungarian Conf. on Computational Linguistics*, pp. 337–347.
- [2] H. Bourlard and N. Morgan. 1994. *Connectionist Speech Recognition – A Hybrid Approach*. Kluwer Academic.
- [3] D. Crystal. 2003. *A Dictionary of Linguistics and Phonetics*. Blackwell Publishing.
- [4] X. Huang, A. Acero, and H.-W. Hon. 2001. *Spoken Language Processing*. Prentice Hall.
- [5] E. Jacobsen and R. Lyons. 2004. An update to the sliding DFT. *IEEE Signal Processing Magazine*, 21(1):110–111.
- [6] N. Kanedera, H. Hermansky, and T. Arai. 1998. Desired characteristics of modulation spectrum for robust automatic speech recognition. In: *Proc. ICASSP'98*, pp. 613–616.
- [7] A. Kornai. 1994. *On Hungarian morphology*. Hungarian Academy of Sciences.
- [8] A. Kuba, A. Hóczy, and J. Csirik. 2004. POS tagging of Hungarian with combined statistical and rule-based methods. In: *Proc. TSD 2004*, pp. 113–121.
- [9] Medisoft. 2004. [www.medisoftspeech.hu](http://www.medisoftspeech.hu)
- [10] P. Motlíček. 2003. *Modeling of Spectra and Temporal Trajectories in Speech Processing*. Ph.D. Thesis, Brno University of Technology.
- [11] G. Németh and Cs. Zainkó. 2001. Word unit based multilingual comparative analysis of text corpora. In: *Proc. Eurospeech 2001*, pp. 2035–2038.
- [12] Nuance. 2007. <http://www.nuance.co.uk/naturallyspeaking/>
- [13] P. Schwarz, P. Matějka, and J. Černocký. 2003. Recognition of phoneme strings using TRAP technique. In: *Proc. Eurospeech 2003*, pp. 825–828.
- [14] P. Siptár, M. Törkenczy. 2000. *The phonology of Hungarian*. Oxford University Press.
- [15] M. Szarvas and S. Furui. 2002. Finite-state transducer based Hungarian LVCSR with explicit modeling of phonological changes. In: *Proc. ICSLP 2002*, pp. 1297–1300.
- [16] K. Vicsi, A. Kocsor, Cs. Teleki, and L. Tóth. 2004. Hungarian speech database for computer-using environments in offices (in Hungarian). In: *Proc. 2nd Hungarian Conf. on Computational Linguistics*, pp. 315–318.
- [17] K. Vicsi, Sz. Velkei, Gy. Szaszák, G. Borostyán, G. Gordos. 2006. Speech recognizer for preparing medical reports – Development experiences of a Hungarian speaker independent continuous speech recognizer. *Híradástechnika, Vol. 61, No. 7*, pp. 22–27.

## JOŽEF STEFAN INSTITUTE

*Jožef Stefan (1835-1893) was one of the most prominent physicists of the 19th century. Born to Slovene parents, he obtained his Ph.D. at Vienna University, where he was later Director of the Physics Institute, Vice-President of the Vienna Academy of Sciences and a member of several scientific institutions in Europe. Stefan explored many areas in hydrodynamics, optics, acoustics, electricity, magnetism and the kinetic theory of gases. Among other things, he originated the law that the total radiation from a black body is proportional to the 4th power of its absolute temperature, known as the Stefan–Boltzmann law.*

The Jožef Stefan Institute (JSI) is the leading independent scientific research institution in Slovenia, covering a broad spectrum of fundamental and applied research in the fields of physics, chemistry and biochemistry, electronics and information science, nuclear science technology, energy research and environmental science.

The Jožef Stefan Institute (JSI) is a research organisation for pure and applied research in the natural sciences and technology. Both are closely interconnected in research departments composed of different task teams. Emphasis in basic research is given to the development and education of young scientists, while applied research and development serve for the transfer of advanced knowledge, contributing to the development of the national economy and society in general.

At present the Institute, with a total of about 800 staff, has 600 researchers, about 250 of whom are postgraduates, nearly 400 of whom have doctorates (Ph.D.), and around 200 of whom have permanent professorships or temporary teaching assignments at the Universities.

In view of its activities and status, the JSI plays the role of a national institute, complementing the role of the universities and bridging the gap between basic science and applications.

Research at the JSI includes the following major fields: physics; chemistry; electronics, informatics and computer sciences; biochemistry; ecology; reactor technology; applied mathematics. Most of the activities are more or less closely connected to information sciences, in particular computer sciences, artificial intelligence, language and speech technologies, computer-aided design, computer architectures, biocybernetics and robotics, computer automation and control, professional electronics, digital communications and networks, and applied mathematics.

The Institute is located in Ljubljana, the capital of the independent state of Slovenia (or S $\heartsuit$ nia). The capital today is considered a crossroad between East, West and Mediter-

anean Europe, offering excellent productive capabilities and solid business opportunities, with strong international connections. Ljubljana is connected to important centers such as Prague, Budapest, Vienna, Zagreb, Milan, Rome, Monaco, Nice, Bern and Munich, all within a radius of 600 km.

From the Jožef Stefan Institute, the Technology park "Ljubljana" has been proposed as part of the national strategy for technological development to foster synergies between research and industry, to promote joint ventures between university bodies, research institutes and innovative industry, to act as an incubator for high-tech initiatives and to accelerate the development cycle of innovative products.

Part of the Institute was reorganized into several high-tech units supported by and connected within the Technology park at the Jožef Stefan Institute, established as the beginning of a regional Technology park "Ljubljana". The project was developed at a particularly historical moment, characterized by the process of state reorganisation, privatisation and private initiative. The national Technology Park is a shareholding company hosting an independent venture-capital institution.

The promoters and operational entities of the project are the Republic of Slovenia, Ministry of Higher Education, Science and Technology and the Jožef Stefan Institute. The framework of the operation also includes the University of Ljubljana, the National Institute of Chemistry, the Institute for Electronics and Vacuum Technology and the Institute for Materials and Construction Research among others. In addition, the project is supported by the Ministry of the Economy, the National Chamber of Economy and the City of Ljubljana.

Jožef Stefan Institute  
Jamova 39, 1000 Ljubljana, Slovenia  
Tel.: +386 1 4773 900, Fax.: +386 1 251 93 85  
WWW: <http://www.ijs.si>  
E-mail: [matjaz.gams@ijs.si](mailto:matjaz.gams@ijs.si)  
Public relations: Polona Strnad



**INFORMATICA**  
**AN INTERNATIONAL JOURNAL OF COMPUTING AND INFORMATICS**  
**INVITATION, COOPERATION**

**Submissions and Refereeing**

Please submit an email with the manuscript to one of the editors from the Editorial Board or to the Managing Editor. At least two referees outside the author's country will examine it, and they are invited to make as many remarks as possible from typing errors to global philosophical disagreements. The chosen editor will send the author the obtained reviews. If the paper is accepted, the editor will also send an email to the managing editor. The executive board will inform the author that the paper has been accepted, and the author will send the paper to the managing editor. The paper will be published within one year of receipt of email with the text in Informatica MS Word format or Informatica L<sup>A</sup>T<sub>E</sub>X format and figures in .eps format. Style and examples of papers can be obtained from <http://www.informatica.si>. Opinions, news, calls for conferences, calls for papers, etc. should be sent directly to the managing editor.

Since 1977, Informatica has been a major Slovenian scientific journal of computing and informatics, including telecommunications, automation and other related areas. In its 16th year (more than ten years ago) it became truly international, although it still remains connected to Central Europe. The basic aim of Informatica is to impose intellectual values (science, engineering) in a distributed organisation.

Informatica is a journal primarily covering the European computer science and informatics community - scientific and educational as well as technical, commercial and industrial. Its basic aim is to enhance communications between different European structures on the basis of equal rights and international refereeing. It publishes scientific papers accepted by at least two referees outside the author's country. In addition, it contains information about conferences, opinions, critical examinations of existing publications and news. Finally, major practical achievements and innovations in the computer and information industry are presented through commercial publications as well as through independent evaluations.

Editing and refereeing are distributed. Each editor can conduct the refereeing process by appointing two new referees or referees from the Board of Referees or Editorial Board. Referees should not be from the author's country. If new referees are appointed, their names will appear in the Refereeing Board.

Informatica is free of charge for major scientific, educational and governmental institutions. Others should subscribe (see the last page of Informatica).

**QUESTIONNAIRE**

Send Informatica free of charge

Yes, we subscribe

Please, complete the order form and send it to Dr. Drago Torkar, Informatica, Institut Jožef Stefan, Jamova 39, 1000 Ljubljana, Slovenia. E-mail: [drago.torkar@ijs.si](mailto:drago.torkar@ijs.si)

**ORDER FORM – INFORMATICA**

Name: .....	Office Address and Telephone (optional): .....
Title and Profession (optional): .....	.....
.....	E-mail Address (optional): .....
Home Address and Telephone (optional): .....	.....
.....	Signature and Date: .....

## Informatica WWW:

<http://www.informatica.si/>

### Referees:

Witold Abramowicz, David Abramson, Adel Adi, Kenneth Aizawa, Suad Alagić, Mohamad Alam, Dia Ali, Alan Aliu, Richard Amoroso, John Anderson, Hans-Jurgen Appelrath, Iván Araujo, Vladimir Bajič, Michel Barbeau, Grzegorz Bartoszewicz, Catriel Beerli, Daniel Beech, Fevzi Belli, Simon Beloglavec, Sondes Bennisri, Francesco Bergadano, Istvan Berkeley, Azer Bestavros, Andraž Bežek, Balaji Bharadwaj, Ralph Bisland, Jacek Blazewicz, Laszlo Boeszöereményi, Damjan Bojadžijev, Jeff Bone, Ivan Bratko, Pavel Brazdil, Bostjan Brumen, Jerzy Brzezinski, Marian Bubak, Davide Bugali, Troy Bull, Sabin Corneliu Buraga, Leslie Burkholder, Frada Burstein, Wojciech Buszkowski, Rajkumar Bvyya, Giacomo Cabri, Netiva Caftori, Patricia Carando, Robert Catral, Jason Ceddia, Ryszard Choras, Wojciech Cellary, Wojciech Chybowski, Andrzej Ciepiewski, Vic Ciesielski, Mel Ó Cinnéide, David Cliff, Maria Cobb, Jean-Pierre Corriveau, Travis Craig, Noel Craske, Matthew Crocker, Tadeusz Czachorski, Milan Češka, Honghua Dai, Bart de Decker, Deborah Dent, Andrej Dobnikar, Sait Dogru, Peter Dolog, Georg Dorfner, Ludoslaw Drelichowski, Matija Drobnič, Maciej Drozdowski, Marek Druzdzel, Marjan Družovec, Jozo Dujmović, Pavol Ďuriš, Amnon Eden, Johann Eder, Hesham El-Rewini, Darrell Ferguson, Warren Fergusson, David Flater, Pierre Flener, Wojciech Fliegner, Vladimir A. Fomichov, Terrence Forgarty, Hans Fraaije, Stan Franklin, Violetta Galant, Hugo de Garis, Eugeniusz Gatnar, Grant Gayed, James Geller, Michael Georgiopolus, Michael Gertz, Jan Goliński, Janusz Gorski, Georg Gottlob, David Green, Herbert Groiss, Jozsef Gyorkos, Marten Haglind, Abdelwahab Hamou-Lhadj, Inman Harvey, Jaak Henno, Marjan Hericko, Henry Hexmoor, Elke Hochmueller, Jack Hodges, John-Paul Hosom, Doug Howe, Rod Howell, Tomáš Hruška, Don Huch, Simone Fischer-Huebner, Zbigniew Huzar, Alexey Ippa, Hannu Jaakkola, Sushil Jajodia, Ryszard Jakubowski, Piotr Jedrzejowicz, A. Milton Jenkins, Eric Johnson, Polina Jordanova, Djani Juričič, Marko Juvancic, Sabhash Kak, Li-Shan Kang, Ivan Kapustok, Orlando Karam, Roland Kaschek, Jacek Kierzenka, Jan Kniat, Stavros Kokkotos, Fabio Kon, Kevin Korb, Gilad Koren, Andrej Krajnc, Henryk Krawczyk, Ben Kroese, Zbyszko Krolikowski, Benjamin Kuipers, Matjaž Kukar, Aarre Laakso, Sofiane Labidi, Les Labuschagne, Ivan Lah, Phil Laplante, Bud Lawson, Herbert Leitold, Ulrike Leopold-Wildburger, Timothy C. Lethbridge, Joseph Y-T. Leung, Barry Levine, Xuefeng Li, Alexander Linkevich, Raymond Lister, Doug Locke, Peter Lockeman, Vincenzo Loia, Matija Lokar, Jason Lowder, Kim Teng Lua, Ann Macintosh, Bernardo Magnini, Andrzej Małachowski, Peter Marcer, Andrzej Marciniak, Witold Marciszewski, Vladimir Marik, Jacek Martinek, Tomasz Maruszewski, Florian Matthes, Daniel Memmi, Timothy Menzies, Dieter Merkl, Zbigniew Michalewicz, Armin R. Mikler, Gautam Mitra, Roland Mittermeir, Madhav Moganti, Reinhard Moller, Tadeusz Morzy, Daniel Mossé, John Mueller, Jari Multisilta, Hari Narayanan, Jerzy Nawrocki, Rance Necaie, Elzbieta Niedzielska, Marian Niedq'zwiadziński, Jaroslav Nieplocha, Oscar Nierstrasz, Roumen Nikolov, Mark Nissen, Jerzy Nogieć, Stefano Nolfi, Franc Novak, Antoni Nowakowski, Adam Nowicki, Tadeusz Nowicki, Daniel Olejar, Hubert Österle, Wojciech Olejniczak, Jerzy Olszewski, Cherry Owen, Mieczyslaw Owoc, Tadeusz Pankowski, Jens Penberg, William C. Perkins, Warren Persons, Mitja Peruš, Fred Petry, Stephen Pike, Niki Pissinou, Aleksander Pivk, Ullin Place, Peter Planinšec, Gabika Polčicová, Gustav Pomberger, James Pomykalski, Tomas E. Potok, Dimithu Prasanna, Gary Preckshot, Dejan Rakovič, Cveta Razdevšek Pučko, Ke Qiu, Michael Quinn, Gerald Quirchmayer, Vojislav D. Radonjic, Luc de Raedt, Ewaryst Rafajlowicz, Sita Ramakrishnan, Kai Rannenberg, Wolf Rauch, Peter Rechenberg, Felix Redmill, James Edward Ries, David Robertson, Marko Robnik, Colette Rolland, Wilhelm Rossak, Ingrid Russel, A.S.M. Sajeev, Kimmo Salmenjoki, Pierangela Samarati, Bo Sanden, P. G. Sarang, Vivek Sarin, Iztok Savnik, Ichiro Satoh, Walter Schempp, Wolfgang Schreiner, Guenter Schmidt, Heinz Schmidt, Dennis Sewer, Zhongzhi Shi, Mária Smolárová, Carine Souveyet, William Spears, Hartmut Stadler, Stanislaw Stanek, Olivero Stock, Janusz Stokłosa, Przemysław Stpicyński, Andrej Stritar, Maciej Stroinski, Leon Strous, Ron Sun, Tomasz Szmuc, Zdzislaw Szyjewski, Jure Šilc, Metod Škarja, Jiří Šlechta, Chew Lim Tan, Zahir Tari, Jurij Tasič, Gheorge Tecuci, Piotr Teczynski, Stephanie Teufel, Ken Tindell, A Min Tjoa, Drago Torkar, Vladimir Tomic, Wieslaw Traczyk, Denis Trček, Roman Trobec, Marek Tudruj, Andrej Ule, Amjad Umar, Andrzej Urbanski, Marko Uršič, Tadeusz Usowicz, Romana Vajde Horvat, Elisabeth Valentine, Kanonkluk Vanapipat, Alexander P. Vazhenin, Jan Verschuren, Zygmunt Vetulani, Olivier de Vel, Didier Vojtisek, Valentino Vranić, Jozef Vyskoc, Eugene Wallingford, Matthew Warren, John Weckert, Michael Weiss, Tatjana Welzer, Lee White, Gerhard Widmer, Stefan Wrobel, Stanislaw Wrycza, Tatyana Yakhno, Janusz Zalewski, Damir Zazula, Yanchun Zhang, Ales Zivkovic, Zonling Zhou, Robert Zorc, Anton P. Železnikar

# *Informatica*

## An International Journal of Computing and Informatics

Web edition of Informatica may be accessed at: <http://www.informatica.si>.

**Subscription Information** Informatica (ISSN 0350-5596) is published four times a year in Spring, Summer, Autumn, and Winter (4 issues per year) by the Slovene Society Informatika, Vožarski pot 12, 1000 Ljubljana, Slovenia.

The subscription rate for 2007 (Volume 31) is

- 60 EUR for institutions,
- 30 EUR for individuals, and
- 15 EUR for students

Claims for missing issues will be honored free of charge within six months after the publication date of the issue.

Typesetting: Borut Žnidar.

Printing: Dikplast Kregar Ivan s.p., Kotna ulica 5, 3000 Celje.

Orders may be placed by email ([drago.torkar@ijs.si](mailto:drago.torkar@ijs.si)), telephone (+386 1 477 3900) or fax (+386 1 251 93 85). The payment should be made to our bank account no.: 02083-0013014662 at NLB d.d., 1520 Ljubljana, Trg republike 2, Slovenija, IBAN no.: SI56020830013014662, SWIFT Code: LJBASI2X.

Informatica is published by Slovene Society Informatika (president Niko Schlamberger) in cooperation with the following societies (and contact persons):

Robotics Society of Slovenia (Jadran Lenarčič)

Slovene Society for Pattern Recognition (Franjo Pernuš)

Slovenian Artificial Intelligence Society; Cognitive Science Society (Matjaž Gams)

Slovenian Society of Mathematicians, Physicists and Astronomers (Bojan Mohar)

Automatic Control Society of Slovenia (Borut Zupančič)

Slovenian Association of Technical and Natural Sciences / Engineering Academy of Slovenia (Igor Grabec)

ACM Slovenia (Dunja Mladenič)

Informatica is surveyed by: Citeseer, COBISS, Compendex, Computer & Information Systems Abstracts, Computer Database, Computer Science Index, Current Mathematical Publications, DBLP Computer Science Bibliography, Directory of Open Access Journals, InfoTrac OneFile, Inspec, Linguistic and Language Behaviour Abstracts, Mathematical Reviews, MatSciNet, MatSci on SilverPlatter, Scopus, Zentralblatt Math
--

*The issuing of the Informatica journal is financially supported by the Ministry of Higher Education, Science and Technology, Trg OF 13, 1000 Ljubljana, Slovenia.*

# *Informatica*

An International Journal of Computing and Informatics

Introduction	M. McPherson, P. Isaias A. Umar	<b>139</b> <b>141</b>
Intelligent Decision Support for Architecture and Integration of Next Generation Enterprises		
Computational Trust in Web Content Quality: A Comparative Evaluation on the Wikipedia Project	P. Dondio, S. Barrett	<b>151</b>
Mobile Ticket Control System with RFID Cards for Administering Annual Secret Elections of University Committees	H. Weghorn, H. P. Großmann, D. Hellwig, C. K. Ratih, A. Schmeiser, H. Hutschenreiter S. Lukosch	<b>161</b> <b>167</b>
Facilitating Shared Knowledge Construction in Collaborative Learning		
A Fuzzy Classification Model for Online Customers	A. Meier, N. Werro	<b>175</b>
Mobile Location-Based Gaming as Driver for Location-Based Services (LBS) - Exemplified by Mobile Hunters	J. Lonthoff, E. Ortner	<b>183</b>
A Group Learning Management Method for Intelligent Tutoring Systems	E. Pozzebon, J. Cardoso, G. Bittencourt, C. Hanachi	<b>191</b>
Reliability, Availability and Security of Wireless Networks in the Community	C. Maple, G. Williams, Y. Yue	<b>201</b>
A Model and Framework for Online Security Benchmarking	G. Pye, M. J. Warren	<b>209</b>
<hr/> End of Special Issue / Start of normal papers <hr/>		
An Approach to Extracting Interschema Properties from XML Schemas at Various "Severity" Levels	P. De Meo, G. Quattrone, G. Terracina, D. Ursino	<b>217</b>
Preliminary Numerical Experiments in Multiobjective Optimization of a Metallurgical Production Process	B. Filipič, T. Tušar, E. Laitinen	<b>233</b>
Development of a Hungarian Medical Dictation System	A. Bánhalmi, D. Paczolay, L. Tóth, A. Kocsor	<b>241</b>

