ICDS 2011

The Fifth International Conference on Digital Society

ASPEN 2011 Workshop

The First International Workshop on Advances in IT-Service Process Engineering

IPDS 2011 Workshop

The First International Workshop for Innovative Methods for Intrusion Prevention and Detection Systems

February 23-28, 2011 - Gosier

Guadeloupe, France

ICDS 2011 Editors

Lasse Berntzen, Vestfold University College - Tønsberg, Norway

Åsa Smedberg, DSV, Stockholm University/KTH, Sweden

Adolfo Villafiorita, Fondazione Bruno Kessler, Italy

Ted Szymanski, McMaster University, Canada

David Day, University of Derby, UK

ICDS 2011 and Workshops ASPEN 2011, IPDS 2011

Foreword

The Fifth International Conference on Digital Society [ICDS 2011], held between February 23-28, 2011 in Gosier, Guadeloupe, France, continued a series of international events covering a large spectrum of topics related to advanced networking, applications, and systems technologies in a digital society.

Nowadays, most of the economic activities and business models are driven by the unprecedented evolution of theories and technologies. The impregnation of these achievements into our society is present everywhere, and it is only question of user education and business models optimization towards a digital society.

Digital devices conquer from kitchen to space vessels most of the functionality commonly performed by human beings. Telecommunications, advanced computation, miniaturization, and high speed devices make tele-presence easy. Wireless and mobility allow ubiquitous systems to be developed. Progress in image processing and exchanging facilitate e-health and virtual doctor teams for patient surgeries.

Naturally, issues on how to monitor, control and manage these systems become crucial to guarantee user privacy and safety. Not only devices, but also special software features must be enforced and guaranteed in a digital society.

The variety of the systems and applications and the heterogeneous nature of information and knowledge representation require special technologies to capture, manage, store, preserve, interpret and deliver the content and documents related to a particular target.

Progress in cognitive science, knowledge acquisition, representation, and processing helped to deal with imprecise, uncertain or incomplete information. Management of geographical and temporal information becomes a challenge, in terms of volume, speed, semantic, decision, and delivery.

Information technologies allow optimization in searching an interpreting data, yet special constraints imposed by the digital society require on-demand, ethics, and legal aspects, as well as user privacy and safety.

Nowadays, there is notable progress in designing and deploying information and organizational management systems, experts systems, tutoring systems, decision support systems, and in general, industrial systems.

The progress in difference domains, such as image processing, wireless communications, computer vision, cardiology, and information storage and management assure a virtual team to access online to the latest achievements.

Processing medical data benefits now from advanced techniques for color imaging, visualization of multi-dimensional projections, Internet imaging localization archiving and as well as from high resolution of medical devices.

Collecting, storing, and handling patient data requires robust processing systems, safe communications and storage, and easy and authenticated online access.

National and cross-national governments' decisions for using the digital advances require e-Government activities on developmental trends, adoption, architecture, transformation, barrier removals, and global success factors. There are challenges for government efficiency in using these technologies such as e-Voting, eHealth record cards, citizen identity digital cards, citizen-centric services, social e-financing projects, and so on.

ICDS 2011 also included:

- ASPEN 2011, The First International Workshop on Advances in IT-Service Process Engineering

- IPDS 2011, The First International Workshop for Innovative Methods for Intrusion Prevention and Detection Systems

We take here the opportunity to warmly thank all the members of the ICDS 2011 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to ICDS 2011. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the ICDS 2011 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that ICDS 2011 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the field of digital society.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the beautiful surroundings of Gosier, Guadeloupe, France.

ICDS 2011 Chairs

Lasse Berntzen, Vestfold University College - Tønsberg, Norway
Åsa Smedberg, DSV, Stockholm University/KTH, Sweden
Freimut Bodendorf, University of Erlangen, Germany
Ted Szymanski, McMaster University, Canada
Adolfo Villafiorita, Fondazione Bruno Kessler, Italy
Aljosa Pasic, Atos Origin, Spain

ASPEN 2011 Workshop Chairs

Christian Bartsch, FZI - Karlsruhe , Germany
Marco Mevius, HTWG Konstanz, Germany

IPDS 2011 Workshop Chairs

David Day, University of Derby, UK
Jianxin Li, Beihang University, China

ICDS 2011 and Workshops ASPEN 2011, IPDS 2011

Committee

ICDS 2011 Advisory Committee

Lasse Berntzen, Vestfold University College - Tønsberg, Norway
Åsa Smedberg, DSV, Stockholm University/KTH, Sweden
Freimut Bodendorf, University of Erlangen, Germany
Ted Szymanski, McMaster University, Canada
Adolfo Villafiorita, Fondazione Bruno Kessler, Italy
Aljosa Pasic, Atos Origin, Spain

ICDS 2011 Technical Program Committee

Shadi Aljawarneh, Isra University, Jordan
Giner Alor Hernández, Instituto Tecnológico de Orizaba-Veracruz, México
Ermeson Andrade, Universidade Federal de Pernambuco (UFPE), Brazil
Anis Ben Arbia, ISITC H. Sousse Tunisia
Pasquale Ardimento, University of Bari, Italy
Gil ad Ariely, Lauder School of Government Diplomacy and Strategy, Interdisciplinary Center Herzliya (IDC), Israel
Ezendu Ariwa, London Metropolitan University, UK
Mehmet N. Aydin, Isik University, Turkey
Gilbert Babin, HEC Montréal, Canada
Kambiz Badie, Iran Telecom Research Center & University of Tehran, Iran
Lasse Berntzen, Vestfold University College - Tønsberg, Norway
Aljosa Jerman Blazic, SETCCE - Ljubljana, Slovenia
Marco Block-Berlitz, Mediadesign Hochschule- Berlin, Germany
Freimut Bodendorf, University of Erlangen, Germany
Nicola Boffoli, University of Bari, Italy
Mahmoud Boufaida, Mentouri University of Constantine, Algeria
Mahmoud Brahimi, University of Msila, Algeria
Luis M. Camarinha-Matos, New University of Lisbon, Portugal
Vlatko Ceric, University of Zagreb, Croatia
Matthew Ilkwon Cho, National Information Society Agency-Seoul, Korea
Yul Chu, University of Texas- Pan American, USA
David Day, University of Derby UK
Gert-Jan de Vreede, University of Nebraska at Omaha, USA
Andrei Alexandru Enescu, Politehnic University of Bucharest, Romania
Karla Felix Navarro, University of Technology - Sydney, Australia
Alejandro Fernández, Universidad Nacional de La Plata, Argentina
Robert Forster, Edgemount Solutions, USA
Shauneen Furlong, Territorial Communications Ltd.-Ottawa, Canada / University of Liverpool, UK
Panos Hahamis, University of Westminster - London, UK
Norbert Heidenbluth, Ulm University, Germany
Mikko Heikkinen, TKK Helsinki University of Technology, Finland

Edward Jaser, Princess Sumaya University for Technology - Amman, Jordan
Gyorgy Kalman, UNIK, Norway
Atsushi Kanai, Hosei University, Japan
Károly Kondorosi, Budapest University of Technology and Economics (BME), Hungary
Christian Kop, University of Klagenfurt, Austria
Radek Kuchta, Brno University of Technology, Czech Republic
Andrew Kusiak, The University of Iowa, USA
Antti Lahtela, University of Eastern Finland, Finland
Man-Sze Li, IC Focus Limited-London, UK
Maryam Tayefeh Mahmoudi, Iran Telecom Research Center, Iran
Peter Mikulecký, University of Hradec Kralove, Czech Republic
Jörg Müller, TU-Clausthal, Germany
Soulakshmee Devi Nagowah, University of Mauritius, Mauritius
Sang-Kyun Noh, The Attached Institute of ETRI, Korea
Daniel O'Leary, University of Southern California, USA
Gerard Parr, University of Ulster, UK
Aljosa Pasic, Atos Origin, Spain
Jyrki Penttinen, Nokia Siemens Networks Spain / Helsinki University ofTechnology, Finland
Augustin Prodan, Iuliu Hatieganu University, Cluj-Napoca, Romania
Drogkaris Prokopios, University of the Aegean, Greece
Juha Puustjärvi, Helsinki University of Technology, Finland
Sanguthevar Rajasekaran, University of Connecticut, USA
Karim Mohammed Rezaul, Glyndwr University - Wrexham, UK
Jarogniew Rykowski, Poznan University of Economics, Poland
Bastien Sasseville, Université du Québec à Rimouski, Canada
Rainer Schmidt, Aalen University, Germany
Thorsten Schöler, Hochschule Augsburg, Deutschland
Dimitrios Serpanos, ISI/R.C. Athena & Univ. of Patras,, Greece
Pushpendra B. Singh, Techizen - Noida, India
Åsa Smedberg, DSV, Stockholm University/KTH, Sweden
Ted Szymanski, McMaster University, Canada
Steffen Thiel, Furtwangen University of Applied Sciences, Germany
Ioan Toma, STI, Austria
Adolfo Villafiorita, Fondazione Bruno Kessler, Italy
Tobias Wegner, Technische Universitaet Dortmund, Germany
Komminist Weldemariam, Center for Scientific and Technological Research / Bruno Kessler Foundation, Italy
Qishi Wu, University of Memphis, USA
Yijian Wu, Fudan University - Shanghai, China
Xiaoli Yang, Purdue University - Calumet, USA

ASPEN Workshop Chairs

Christian Bartsch, FZI - Karlsruhe , Germany
Marco Mevius, HTWG Konstanz, Germany

ASPEN 2011 Technical Program Committee

Claudio Bartolini, HP Labs Research, USA
Kamal Bhattacharya, IBM Research - India, Bangalore
Carsten Gentner, ENBW Energie Baden Wuerttemberg, Germany
Hans-Peter Fröschle, IT Service Management Forum (itSMF), Germany
Dieter Hertweck, Heilbronn University, Germany
Khaled Nagi, Alexandria University, Egypt
Martin Necasky, Charles University in Prague, Czech Republic
Christian Pichler, Research Studios, Austria
Christoph Rentrop, HTWG Konstanz, Germany
Florian Rosenberg, CSIRO ICT Centre, Australia
Gerhard Satzger, Karlsruhe Service Research Institute (KSRI), KIT, Germany
Andreas Schmietendorf, Berlin School of Economics and Law, Germany
Rainer Schmidt, HTW Aalen, Germany
Swen Schneider, Frankfurt University of Applied Sciences, Germany
Thomas Setzer, TUM, Informatics Department, Germany
Larisa Shwartz, IBM T. J. Watson Research Center, USA
Stefan Tai, Karlsruhe Service Research Institute (KSRI), KIT, Germany
Christopher Ward, IBM T. J. Watson Research Center, USA

IPDS Workshop Chairs

David Day, University of Derby, UK
Jianxin Li, Beihang University, China

IPDS 2011 Technical Program Committee

David Day, University of Derby UK
George Exarchakos, Eindhoven University of Technology, The Netherlands
Geofrey Falzon, STMicroelectronics (Malta) Ltd, Malta
Richard Hill, University of Derby, UK
Harjinder Lallie, University of Derby, UK
Jianxin Li, Beihang University, China
Lu Liu, University of Derby, UK
Simon Polovina, Sheffield Hallam University, UK
Ann Reddipogu, ManyWorlds UK Ltd, UK
Dharam Shadija, Sheffield Hallam Univeristy, UK
Lei Shu, Osaka University, Japan
Ashley Thomas, SecureWorks, USA
Richard Thorne, Victoria University, Australia
Dave Voorhis, University of Derby, UK
Zhengxu Zhao, Shijiazhuang Tiedao University, China
Zhangbing Zhou, Telecom SudParis, France
Rongbo Zhu, South Centre University for Nationalities, China

Copyright Information

Table of Contents

Exploring an IT Service Change Management Process: A Case Study

Marko Jäntti
University of Eastern Finland
School of Computing, Software Engineering Research Group
P.O.B 1627, 70211 Kuopio, Finland
Email: firstname.lastname@uef.fi

Merja Kainulainen
Istekki Oy
P.O.B 1777, Kuopio, Finland
Email: firstname.lastname@istekki.fi

***Abstract*—IT service providers need effective methods for managing change requests and changes regarding provided IT services and the IT infrastructure. However, many IT service providers consider the implementation of a service-oriented change management process as a difficult task. This challenge has led us to examine the research problem: How to improve a change management process based on IT service management practices? The main contribution of this study is 1) to explore how an IT service provider organization in Finland has implemented a change management process, and 2) to identify what types of challenges are related to the transition process from traditional change management to service-oriented change management. Data for this study were collected by using a case study research method.**

***Keywords*-change management; request for change; IT service**

I. Introduction

Many IT service provider organizations are living under continuous change. Changes have effects on organizational structures, provided IT services, software products and systems, technologies, data networks, documentation and processes. A documented change management process plays an important role because changes need to be recorded, evaluated, authorized, prioritized, planned, tested, implemented, documented and reviewed in a controlled manner [1].

Improvement of change management is usually started from three main reasons. First, an organization may identify quality problems in change management activities, such as information on changes is never logged into the customer support tool, unauthorized changes are performed, or there is no defined process for handling changes that cause several different types of interpretations how changes can be handled. Second, a key customer may require that an IT provider should improve change management practices. Informing IT people on customer's requirements regarding change management is very effective way to motivate workers to follow change management procedures. Third, an organization may be interested in adopting a standard or a process framework that requires establishment or improvement of the documented change management process.

In this paper, we present results from a case study where the improvement of change management was triggered by the adoption of the ISO/IEC 20 000 service management standard. The ISO/IEC 20 000 standard is aligned with IT Infrastructure Library best practices and consists of two parts: Part 1: Specification for service management (Shall requirements) [2], and Part 2: Code of practice for service management (Should requirements) [3].

IT organizations that aim to achieve ISO/IEC 20 00 shall very likely exploit the most widely used IT service management framework IT Infrastructure Library (ITIL) as guidelines in the process improvement work. ITIL is a collection of best practices for defining, designing, implementing, managing and monitoring IT services and IT service management processes [4]. ITIL defines service management as "the implementation and management of quality IT services that meet the needs of the business". The service management section of the ITIL version 2 consists of two parts: Service Delivery and Service Support (including change management). In the ITIL version 3, change management is part of the Service Transition publication [1]. The main objective of the change management process is to ensure that standardised methods and procedures are used for efficient and prompt handling of all changes [5].

A. Related work

Much has been written about software maintenance and managing changes and defects. In the software engineering literature, change management is mainly focused on tracking changes to the technical artifacts (software components) [6]. Information on a change is captured by a software change order. Service-oriented change management covers both changes to technical configuration items (servers, desktops, software components) and changes to services and the service infrastructure. Lientz and Swanson [7] categorize software maintenance into four categories: adaptive, perfective, corrective and preventive.

Change management seems to be related to all of those four perspectives covering changes in the software environment, changes due to new user requirements, changes due to fixing errors, and changes that are made to prevent problems in the future. Implementation of changes as corrective actions can be seen as one way to prevent defects [8]. Requests for change are typical outputs from problem management and defect management process [9]. Bennett [10] states that

ISBN: 978-1-61208-116-8

when software enters the servicing (software maturity) stage, only small tactical changes (patches, code changes) are possible. Wallmueller [11] considers change management activities as a subgroup of configuration management. He states that configuration control phase includes inputting change requests into the development process, controlling the processing of changes and tracing the changes to their closure.

Within service science and services computing, IT change management is part of the services operation phase in the services lifecycle [12]. There is a wide number of IT service management studies. Hochstein, Zarnekow and Brenner [13] have studied ITIL as a common-practice reference model in three case studies. Tan, Cater-Steel and Toleman [14] have identified six success factors in ITIL implementations: senior management support, project champion, relationships with tool vendors, change in corporate culture, project governance and execution and realisation of benefits. Additionally, Lahtela, Jäntti and Kaukola [15] have explored implementing an ITIL-based IT service management measurement system.

Pollard and Cater-Steel [16] report four interesting challenges in ITIL implementation: People do not understand that one person can hold many ITIL roles (hats), engaging the right people to make changes, gaining support from technical staff and measuring the ROI of ITIL implementation. Kapella [17] has presented a framework for incident management and problem management where known errors from problem management process should be implemented via change management process. Niessink and van Vliet have explored software maintenance from a service perspective [18] and reported that IT support organizations seem to have problems in the interface between incident management and problem management. Duffy and Denison [19] have presented a conceptual model of ITIL impacts on IT services.

Surprisingly few studies have dealt with IT service change management. Wickboldt et al. have studied IT change management from the risk analysis viewpoint [20]. They report that all requests for change must be submitted to the Change Advisory Board (CAB) to be analyzed, approved, and scheduled. We propose that a change manager should be able to preview low impact RFCs and authorize them instead of a formal CAB meeting in order to make handling the change more effective.

The ITIL change management process consists of the following activities (see Fig. 1). An organization must define separate procedures for handling standard changes and urgent (emergency) changes.

Change management is also clearly visible in the Control Objectives for IT and Related Technology (COBIT) framework. Manage Changes is one of the Acquire and Implement (AI) processes [21].

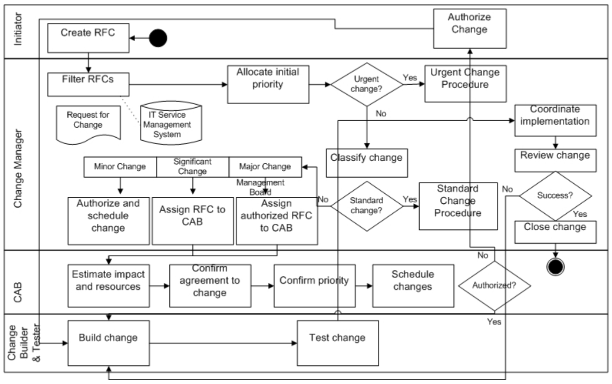


Figure 1. The change management process (adapted from [5])

B. Our Contribution

Surprisingly few studies have dealt with change management from IT service management perspective. This case study belongs to the results of KISMET (Keys to IT Service Management and Effective Transition of Services) research project at the University of Eastern Finland. A part of the research work was conducted during our previous research project MaISSI (Managing IT Services and Service Implementation) research project. The main contribution of this paper is to

- to explore how an IT service provider organization from healthcare domain in Finland has implemented a change management process, and
- to identify what types of challenges are related to the change management.

The results of this study are valuable for IT service provider organizations that are planning to implement IT service change management process or improve existing change management activities. The remainder of the paper is organized as follows. In Section II the research problem and research methods of this study are described. In Section III, we explore the change management process of the case organization. Section IV is the analysis. Finally, the discussion and the conclusions are given in Section V.

II. Research Questions & Methodology

A case study research method was used to answer the research problem: How to improve a change management process based on IT service management practices? Benbasat and Goldstein [22] report that a case study research is a viable research strategy to answer “how” and “why” research questions. According to Yin [23] a case study is “an empirical inquiry that investigates a contemporary phenomenon within its real-life context”. Additionally, a case study is defined as “a research strategy which focuses on understanding the dynamics present with single settings” [24]. The case study method was used to explore the current state of change management process in the case organization and to identify process-related challenges.

The case study design included the following questions that were monitored during the study:

- How requests for change (RFCs) are handled?
- How the initial review of RFCs has been organized?
- How changes are approved for implementation?
- How changes are planned?
- How changes are implemented?
- How changes are monitored or handled after the implementation?
- What types of challenges are related to change management?

A. Case Organization and Data Collection Methods

The case organization is one of the KISMET research project's industrial partners. KISMET research team aims to help IT service providers in the adoption of IT service management methods and processes. Figure 2 describes the context of this study.

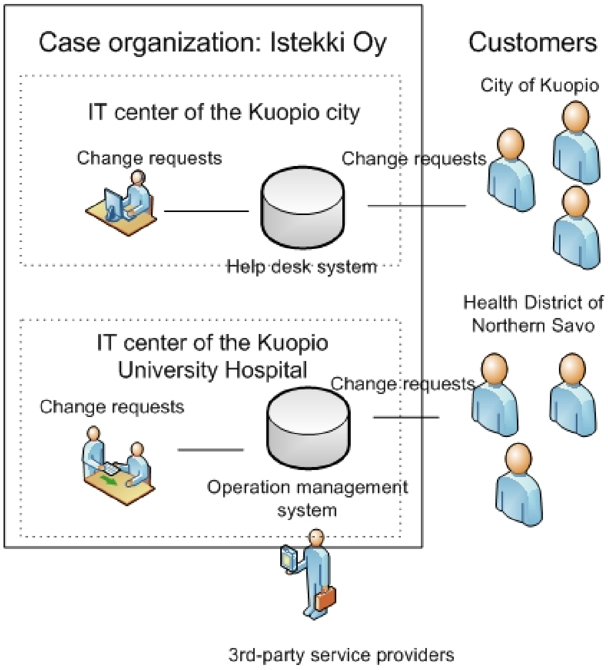

Case organization: Istekki Oy
Customers
IT center of the Kuopio city
City of Kuopio
Change requests
Change requests
Help desk system
Health District of Northern Savo
IT center of the Kuopio University Hospital
Change requests
Change requests
Operation management system
3rd-party service providers

Figure 2. The case study context

Our case organization Istekki is an IT service provider company with 160 employees. Istekki provides IT and medical technology services to the city of Kuopio and Hospital District of Northern Savo. The case organization started ISO 20000 project around two years ago. The following data collection methods were used in the study:

- Participative observation
  - June 30, 2009: a meeting with a customer support tool development team
  - August 25, 2009: a 'support process' training session in the case organized by a research team
  - September 15, 2009: a change management meeting (incident, problem and change manager, business area manager)
- July 29 - August 12, 2010: theme interviews with 10 persons
- A case study writeup
- Internal documentation (a project plan of the standardization project, a system specification of the customer support tool)

Theme interviews were the main data collection method in this study. The following interview questions were selected with a customer manager and a change manager: what is the current state of change management, what types of challenges are related to change management, which metrics are used within change management process, are you familiar with ISO 20000 standard and ITIL framework? Interviewees (two per unit) were selected randomly from the following units: service desk, application support services, on site support services and IT technology services.

B. Data Analysis Method

In data analysis, we used a within-case analysis method that examines a case carefully as a stand-alone entity [24]. Interviews were recorded and the most important findings were stored in an Excel file. After the interview phase, results were analyzed and reported as a case study writeup by a computer science student. The case study database was created to ensure the traceability between data sources, meetings and findings. The case study database included memos from meetings with a case organization.

III. IT Service Change Management Process

The main goal of this study was to explore the change management process in the case organization and identify the challenges regarding the current state of the process and the challenges related to the transition from traditional change management to service oriented change management.

A. Change Request Procedures

There are different types of change requests in different units. Interviewees had received change requests related to software, hardware, procedures and instructions. They had received change requests from service desk, customers, boss, colleagues, from their own working unit or they had identified a change target by themselves. Change requests usually come by phone, email or ticket systems (operational management system or help desk system). Many answers indicated that requests from customers would go through a service desk that assigns the requests to specialist teams if necessary. However, it is very common that customers contact specialists directly. There are different procedures for recording change requests because there are no unified instructions. Change requests are recorded in ticket systems, in personal documents and some requests are not recorded at all.

Additionally, the quality of change recording varies a lot in the case organization. A part of the change requests include a very detailed description of a change while other change requests are written in a very general level. The quality of recording depends on the size of the change. Large changes are recorded more detailed than small changes.

B. *Initial Assessment of Requests for Change*

According to case study observations, RFCs are not classified systematically in the initial assessment phase. The classification is often based on the persons' own interpretations. Several interviewees felt that they do not have instructions how to classify RFCs and how they should react to certain types of changes. Although there is no systematic classification, the reaction time to problems in the customers' critical information systems is shorter than other problems. In some cases, a manager had defined which change requests are urgent.

The service desk tool includes some classification rules for change impact: makes user's work difficult, user cannot continue work, major incident. In case of an urgent change the reaction time is shorter and there will be more resources to solve this kind of changes. When an urgent change is needed, the RFC will not receive detailed assessment but a change is implemented as soon as possible. Small changes can be implemented quite freely. Thus, the person who implements the change is also responsible for the change. The scope of the change defines how detailed impact analysis is carried out for the RFC. For major changes one can estimate costs, service downtime, resources and risks. Change impact is often estimated with colleagues or with own team which results in a better overview of the issue.

C. *Change Approval*

The type of the RFC defines who is able to approve the change. Change approval can be made by a customer, a customer and a specialist together, a system main user, a service manager, own team, own manager, information system manager, a specialist or an RFC initiator. Information on the decision whether the change is accepted is usually not recorded in the service management system. Decision is typically informed by email or by phone. In some cases, all stakeholders had not received information. If the RFC is rejected, the decision is often not recorded but the initiator of RFC is informed. If the RFC is recorded in the service desk tool, the rejected RFC shall be closed and the reason for rejection can be added to the RFC record. Interviewees hoped that decision making on change approval should be flexible enough in order to avoid unnecessary waiting.

D. *Planning of Change Implementation*

Several interviewees stated that planning of the change implementation is carried out carefully but there are challenges in documenting. For large changes, deailed planning of implementation is carried out including resources, schedule, implementation method, implementation steps and effects. There are no templates for plans. Thus, their structure and content varies between teams. Implementation plans are located in email, shared folders or in own documents. There is no agreed place for planning documentation archives.

Implemenation plans are not created for small changes. Some interviewees felt that in some cases it is waste of time to log changes to the system. However, other interviewees thought that planning should be detailed and it would be good to document also the small changes. It was noted that every person who is responsible for change implementation should take account in the backup and restoration plan in planning the implementation. Backup and restoration plans are typically not recorded anywhere or plans exist in the private folders.

E. *Implementation of Change*

Ínterview ´results indicate that there are several different methods to implement changes. Many interviewees stated that information on implemented changes is located in different places: personal documents, email messages, back office systems (for example, a hardware register), operation management tools, web sites and program code. There is no unified procedure how and where information on change implementations should be recorded. Bigger changes include a short description what was done and when. If necessary, information on a change is sent to appropriate stakeholders.

F. *Post-Implementation Monitoring*

The level of post-implementation monitoring of change depends on the type of the change and how active is the person who implemented the change. Quite often, there is no time for monitoring or monitoring is not possible. In these cases, post-implementation monitoring means that a customer contacts when a problen occurs.

After major changes the situation is strictly monitored and controlled and if necessary one can add resources. In case of a bigger change, a service can be monitored by technical monitoring tools. In some cases, IT provider can review the change with a customer whether a change has been successfully implemented. If the change had unexpected effects, the service can be restored to the initial state. In all cases, the restoration cannot be done. Large, failed changes should be reported to a manager level.

IV. ANALYSIS

Based on the case study results, the following strenghts were identified regarding the case organization's change management process:

- There is a change manager role with documented responsibilities
- A description of the change management process exists
- Change management and maintenance work is carried out carefully by specialists
- There are two tools that support recording RFCs and change details in a basic level
- There are defined workflows for different types of changes: small changes, emergency changes, major changes and development ideas.
- There is a list of planned metrics for change management.

In contrast to strenghts, the following key challenges can be identified in the case organization's change management:

- Lack of unified change management methods
- Lack of centralized point for change requests
- Poor documentation of how to record changes
- Lack of resources (more time should be allocated to improvement of the change management procedures)
- The service desk tool does not provide enough support for change management at the moment
- The issues defined in the theoretical process description are not applied in practice
- Knowledge sharing problems on planned or implemented changes
- Change requests come as a surprise which causes challenges for resource planning.
- It is difficult to identify the connection between a new incident and a recently made change.
- Sometimes IT people have to carry out changes that are not well planned.
- Unclear responsibilities of change management.
- The organization has two locations and two business areas which causes challenges in decision making regarding changes.

Many answers indicated that maintenance and change work is carried out carefully and in a professional way in the case organization. Additionally, the service desk tools support recording service requests and part of the RFCs are logged into the tools. The organization has planned and documented a change management process based on ISO 20 000 and ITIL concepts. The process description includes a description of the change management goals, roles, responsibilities, activities, the lifecycle of the change (with change status options), the content of the change record, process metrics and process relationships to other ISO 20 000 processes. The organizations has a change manager that is responsible for improving the process. However, the introduction of the process in the practice seems to be a big challenge in the near future. According to our observations, a change advisory board has been in a pilot use for some products.

The major challenges are related to guidance how changes should be recorded, classified and documented. There are people who do not understand the difference between service requests and change requests. This is clearly a challenge that is related to the transition process from traditional change management to service-oriented change management. Thus, these support requests are handled in the same way although they have different workflows. The process description has been made in one organizational unit and it does not address well the needs of another unit. Knowledge sharing problems on planned or implemented changes refers to the point that all the RFC/change-related information should be stored in the service desk tool instead of private files to improve the data availability. According to some interviewees, the tool does not support the change management enough, for example, in producing measurement reports.

Besides identified challenges, this study resulted in several useful process improvement suggestions. More resources should be allocated for improving change management procedures. More attention should be placed in informing IT people of change management practices and responsibilities and the existence of a process description. Informing could be done, for example, in weekly team meetings. The change management process description should be updated in the cooperation between representatives from both organizational units. Perhaps, in the future, they could have only one tool for managing changes. If the business domains of two units are very different, two or more change managers might be needed.

V. Discussion and Conclusion

This study aimed to answer the following research problem: How to improve a change management process based on IT service management practices? The main contribution of this study was to present 1) to explore how an IT service provider organization in Finland has implemented a change management process, and 2) to identify what types of challenges are related to change management. Data for this study were collected using a case study method.

First, we described the change management practices of the case organization. We focused on exploring change request procedures, initial assessment of requests for change, change approval, planning of change implementation, implementation of change, and post-implementation. Second, we presented challenges related to change management. Key challenges were related to lack of instructions how to record and classify RFCs, lack of unified change management methods, poor understanding of the difference between service requests and change requests, and knowledge sharing of change-related information.

There are several limitations to this study. First, data were collected from one IT service provider organization by using qualitative research methods such as interviews. More IT people, process managers and customers could have been interviewed. Second, the case organization was not randomly selected but selected from the partner pool of the MaISSI project. Thus, the selection method was the convenience sampling method. Third, the case study does not allow us to generalize our research results. However, we can use our results to expand the theory of IT service change management. Further research is needed to examine introduction of change management tools, processes and methods in IT service companies.

Acknowledgment

This paper is based on research in KISMET (Keys to IT Service Management and Effective Transition of Services) and MaISSI (Managing IT Services and Service Implementation) projects funded by the National Technology Agency TEKES, European Regional Development Fund (ERDF), and industrial partners.

References

[1] Office of Government Commerce, *ITIL Service Transition*. The Stationary Office, UK, 2007.

[2] ISOIEC20000, *IT Service Management, Part 1: Specification for service management*. ISO/IEC JTC1/SC7 Secretariat, 2005.

[3] ISOIEC20000b, *IT Service Management, Part 2: Code of practice for service management*. ISO/IEC JTC1/SC7 Secretariat, 2005.

[4] Office of Government Commerce, *ITIL Service Operation*. The Stationary Office, UK, 2007.

[5] OGC, *ITIL Service Support*. The Stationary Office, UK, 2002.

[6] W. Royce, *Software Project Management: A Unified Framework*. Addison-Wesley, 1998.

[7] B. P. Lientz and E. B. Swanson, *Software Maintenance Management*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1980.

[8] R. G. Mays, C. L. Jones, G. J. Holloway, and D. P. Studinski, "Experiences with defect prevention," *IBM Syst. J.*, vol. 29, no. 1, pp. 4–32, 1990.

[9] Quality Assurance Institute, "A software defect management process," Research Report number 8, 1995.

[10] K. H. Bennett and V. T. Rajlich, "Software maintenance and evolution: a roadmap," in *ICSE '00: Proceedings of the Conference on The Future of Software Engineering*. New York, NY, USA: ACM Press, 2000, pp. 73–87.

[11] E. Wallmueller, *Software quality assurance: A practical approach*. Prentice Hall International, 1994.

[12] L.-J. Zhang, J. Zhang, and H. Cai, *Services Computing*. Tsinghua University Press, Beijing and Springer-Verlag GmbH Berlin Heidelberg, 2007.

[13] A. Hochstein, R. Zarnekow, and W. Brenner, "Itil as common practice reference model for it service management: Formal assessment and implications for practice," in *EEE '05: Proceedings of the 2005 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'05) on e-Technology, e-Commerce and e-Service*. Washington, DC, USA: IEEE Computer Society, 2005, pp. 704–710.

[14] W.-G. Tan, A. Cater-Steel, and M. Toleman, "Implementing it service management: A case study focussing on critical success factors," *Journal of Computer Information Systems*, vol. 50, no. 2, 2009.

[15] A. Lahtela, M. Jäntti, and J. Kaukola, "Implementing an itil-based it service management measurement system," in *Proceedings of the 4th International Conference on Digital Society*. St. Maarten, Netherlands Antilles: IEEE Computer Society, February 2010, pp. 249–254.

[16] C. Pollard and A. Cater-Steel, "Justifications, strategies, and critical success factors in successful itil implementations in u.s. and australian companies: An exploratory study," *Information Systems Management*, vol. 26, no. 2, pp. 164–175, 2009.

[17] V. Kapella, "A framework for incident and problem management," International Network Services whitepaper, 2003.

[18] F. Niessink and H. van Vliet, "Software maintenance from a service perspective," *Journal of Software Maintenance*, vol. 12, no. 2, pp. 103–120, March/April 2000.

[19] K. Duffy and B. Denison, "Using itil to improve it services," in *AMCIS08: Proceedings of the FourteenthAmerican Conference on Information Systems 2008*. Toronto, Canada: Assocation for Information Systems, 2008.

[20] J. A. Wickboldt, G. S. Machado, W. L. da Costa Cordeiro, R. C. Lunardi, A. D. dos Santos, F. G. Andreis, C. B. Both, L. Z. Granville, L. P. Gaspary, C. Bartolini, and D. Trastour, "A solution to support risk analysis on it change management," in *IM'09: Proceedings of the 11th IFIP/IEEE international conference on Symposium on Integrated Network Management*. Piscataway, NJ, USA: IEEE Press, 2009, pp. 445–452.

[21] COBIT 4.1, *Control Objectives for Information and related Technology: COBIT 4.1*. IT Governance Institute, 2007.

[22] I. Benbasat, D. K. Goldstein, and M. Mead, "The case research strategy in studies of information systems," *MIS Q.*, vol. 11, no. 3, pp. 369–386, 1987.

[23] R. Yin, *Case Study Research: Design and Methods*. Beverly Hills, CA: Sage Publishing, 1994.

[24] K. Eisenhardt, "Building theories from case study research," *Academy of Management Review*, vol. 14, pp. 532–550, 1989.

Explicating technological and organizational interfaces of modular IT service components to support the process of IT service composition

Sebastian Dudek
Institute of Information Management
University of St. Gallen
St. Gallen, Switzerland
sebastian.dudek@unisg.ch

Falk Uebernickel
Institute of Information Management
University of St. Gallen
St. Gallen, Switzerland
falk.uebernickel@unisg.ch

Walter Brenner
Institute of Information Management
University of St. Gallen
St. Gallen, Switzerland
walter.brenner@unisg.ch

***Abstract*— IT organizations and IT service providers decompose monolithic IT services in IT service components for reaching a higher degree of reusability and therefore realizing higher economies of scales. Existing concepts of modularity are being adapted when creating these IT service components. In the area of software intensive systems modularity is a widely-used design principle to implement reusable and combinable elements. The elements (e.g. classes in object-oriented programming or services of a service-oriented architecture) commonly integrate the data and its behavior as well as supply interfaces for communication between them. Typically these elements are considered from a technological point of view regarding the possibilities to be composed. But IT service components are slightly different: they can be described by both technological and organizational aspects. The following paper examines both aspects of IT services and its components and demonstrates possibilities to integrate these facts in the description of IT service components. This approach is useful to decrease the effort for retrieving the necessary information in the process of IT service composition and therefore leads to less time-to-market, higher degree of reusability and higher quality of the delivered IT service.**

Keywords-IT service, IT service component, IT service composition, IT service engineering, IT service component description

I. INTRODUCTION

An increasing competitive environment in the IT industry leads to an enormous cost pressure for the delivery of IT services but concurrently IT organizations and IT service providers should also deliver high-quality, customer oriented IT services [31]. This arising trade-off is often faced by modularizing IT services or rather the components of which IT service are composed (e.g. [12,19]). The idea of modularization is adapted by former developed concepts e.g. from object-oriented programming or the service-oriented architecture. The approaches aim at loose coupled, high cohesive elements respectively components that only communicate via well-defined interfaces and therefore ensure the possibility of local modification without the need to change surrounding modules. In more mature industries these principle are also known as modular product or service architecture and are basic beliefs in the context of mass-customization [8,9,27,28]. Several publications deal with the idea to transfer either IT-centric or mass-customization approaches to the engineering or design of IT services [4,12]. The objective is obvious: to create modular IT service components that can be variously combined, locally modified and reused to increase the degree of standardization and hence realize economies of scale. These IT service components cover technological and organizational aspects, e.g. the delivery of a server system includes the server itself, all necessary production processes to get the server up and running and the corresponding business unit(s) must be assigned that operate the production processes [12]. As the example illustrates, a pure technological point of view is not sufficient to fully characterize an IT service component. Subsequently, mechanisms in creating, combining and describing IT service components should not only focus on technological aspects but on organizational issues, too. The possibility to combine two IT service components is accordingly dependent on both – the technological and organizational – characteristics of each IT service component and critical to reach a higher degree of reusability. Currently, the process of proofing the compatibility and effects when combining two IT service components is often operated by several business units and sparse structured.

We propose that a need exists to structure and to explicate corresponding information / knowledge to support the process of composing new IT services. This assumption is also based on research in the area of software engineering that cautions about the ‘modularity crises’. Overhage states that common development processes focus on creating modules but do not concentrate on reusing / composing them [26]. Hence the development processes must be adjusted and information about modules must be explicated to browse through the set of existing elements and finally find the appropriate one. We argue that such challenges also exist in the IT service industry when composing IT services out of IT service components.

The following paper will not focus on the process of IT service composition itself but on needed information during the process. We assume that process reference models that describe the process of IT service (de)composition (e.g. [19,23]) are already available but we do not deny that additional information could be necessary when further examination of the process of composing IT services is conducted.

We position this publication in the second stage of the information lifecycle suggested by [3]: (1) information retrieval, (2) information structuring and storing, (3) information maintenance, (4) information usage and enrichment, (5) information spreading and (6) information disposal. The shown artifact will provide a means to structure and store information about IT service components that are relevant when combining two or more IT service components to one IT service. On the one hand it focuses on properties of IT service components that influence the compatibility to other IT service components, technologically as well as organizational. On the other hand (data) structures are presented that are feasible for structuring and storing information about characteristics of IT service components.

The paper is divided in six parts: the next chapter summarizes the research methodology and related work. Following definitions of modularity, IT service product and IT service components we introduce an IT service component and process typology. Chapter 5 depicts the requirements and restrictions in the process of composition and examines compatibility of two IT service components. After applying the approach on a real life example we give a brief preview of future researches.

II. Related Work and Research Methodology

Related work in this area of research concentrates on service engineering and modularization in general [7,9,14], deriving IT services and particularly on methods and requirements needed for the creation of modular IT service components [12,19], packaging IT service components to IT services [22,23] or analyzing the necessity for a consistent service data management – partly focus on the IT service design stage [10,16,18,19,20,30]. An application-centric approach for describing and combining components is presented in [26]. Many authors highlight the advantages (e.g. standardization, reusability, economies of scale) arising from modular built components but only a few approaches provide a specific set of properties for describing and structuring IT service components. Mostly a technological hierarchy is proposed as exclusive sorting criterion [19]. Some work deal with the modeling of process interfaces between IT service components [11].

The research was conducted with leading IT service providers in Europe. We choose the design science research methodology [21] to construct the later presented IT artifacts. Concrete findings were gained during a research project with one IT service provider that offers IT services in each level of the IT technology stack (from telecommunication up to process related IT services). In the project we analyzed the existing description and structuring of IT service components. IT service components are typically specified in two ways: firstly a composite attribute (called short-description) encompass all defining properties of IT service components and secondly long descriptions exist in continuous text documents that are created and maintained with text processing software and enrich the information stored in the composite attributes. Hence the long descriptions are not able to be evaluated by applications that support the IT service engineer. We normalized (according to [13]) the composite attribute “short-description” and analyzed the continuous text descriptions to identify IT service components, its properties and its interdependencies.

III. IT Service Product and modular IT Service Components

Following ideas of the product architecture Burr defines service architecture as the decomposition of a service in components including the definition of technological and organizational interfaces between components [9]. It is also possible to adapt the concepts of service architecture to IT services [12].

As stated previously modularity is discussed in a lot of scientific contributions from different angles. Commonly, modules can be characterized by a few properties. The first characteristic of modules is that they are derived by decomposition of an IT service, i.e. each module contributes a part to the entire IT service. Secondly, modular IT services are built of nested hierarchies, i.e. one module can be used to compose another module and vice versa. Thirdly, the underlying design principles of modules are loose coupling and high cohesion, i.e. strongly-related elements should be merged to a module. The components only communicate via well-defined interfaces. These principles ensure reusability and possibilities of local modifications without changing surrounding components [12].

In the following we concentrate on the property of (nested) hierarchies of IT services. We therefore distinguish between atomic IT service components and compositions of IT service components. Atomic IT service components can be defined from two different points of view: firstly, the focused IT service component cannot truly be further decomposed (e.g. infrastructure IT services). Secondly, the considered IT organization or unit does not have knowledge about the further decomposition of the IT service component. This can be observed if the IT organization or unit sources some parts of their provided and delivered IT services. The emerging trend to bundle IT service components to IT services leads to “productized” IT services, labeled as IT service products [7,22,23]. It is noteworthy here that IT service products exist on every level of the IT technology stack (Figure 1), dependent on the considered customer. Hence IT service and IT service component are used equivalently, whereas an IT service product is a sellable (composition of) IT service component(s).

IV. IT Service Typology and Process Typology for Describing and Deriving IT Service Components

Before we propose our approach of supporting IT service composition we clarify our semantic understanding of IT service components.

Figure 1 illustrates an overview of so-called abstract IT service components. Abstract IT service components represent the templates for specific instances of IT service components. Each white rectangle represents an abstract IT service component and each grey rectangle groups several abstract IT service components. The four rows on the left side arrange all abstract IT service components along the IT technology stack. The two columns on the right side group abstract IT service components that cannot be assigned to a specific stage of the IT technology stack. These service are commonly referred as "professional services" [19] and consist of activities mainly operated by humans.

Various properties are assigned to every abstract IT service[1] (e.g. the database service is further described by database system capacity, database storage capacity, supported database languages, database conceptual model and so on). The groups of properties[2] are (partly based on [19]):

- Function – encompasses properties that further describe the functional aspects of IT service components (e.g. operating system, database model)
- Security – encompasses security relevant properties (e.g. end-to-end encryption or physical access restrictions to the data center)
- Quality – groups properties that further specify the quality aspects of IT service components. It is noteworthy here both technological quality (e.g. availability, reliability) and organizational quality (e.g. delivery time, support time) are considered.
- Capacity – stores the (technological and organizational) capacity requirements of IT service components (e.g. database capacity in TPC-C, capacity of a specified business unit)
- Qualification – describes the demand of necessary qualification of people (e.g. specified by complexity of the task, level of knowledge,)
- Customer integration – specifies if and how the customer is integrated in the delivery of the IT service component (e.g. integration of customer managed systems, jointly operated processes)
- Organizational structure – sets the delivering business unit and possible points of consumption (e.g. produced at the data center in Arizona, delivered at the office in Chicago)

This mechanism is adapted by concepts of object-oriented programming where abstract classes ensure a higher degree of reusability and define a common interface for all derived subclasses [2]. Moreover abstract classes assign general properties and state which properties must be overwritten. In our approach abstract IT service components – and entire hierarchies of them – are used to specify common behavior of derived IT service components of one specific IT service component type and its specification. The set and hierarchies of abstract IT service components and its assigned properties creates a typology for IT service components. We assume that these properties can be used to define the interface of IT service components, too. This mechanism is shown below.

An abstract IT service component is instantiated by specifying the necessary, assigned properties. For example, an abstract IT service component "database service" is instantiated by setting the database system capacity to x tpmC[3], database storage capacity to y Gbyte and so on. The result is an instance of the abstract IT service component with certain behavior. These instances are called IT services or rather IT service components.

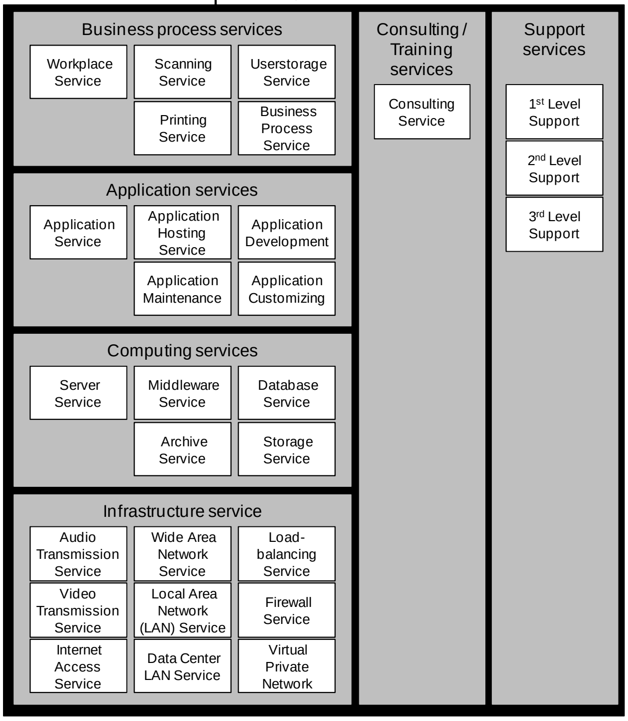


Figure 1: Typology of abstract IT service (components)

The typology (Figure 1) of abstract IT service components concentrates primarily on technological aspects of IT service products although we also integrated the "professional services" as non-technology oriented services. The underlying classification criterion is "required skills" of the delivering business unit. It is obvious that a database administrator generally needs different skills in contrast to a network administrator.

In addition to technological characteristics we previously mentioned that organizational aspects must be considered as well. We therefore introduce a process typology to provide a specification framework. These processes are fairly structured and typically operated in IT organizations while the IT service components are delivered: install/ setup/ register, move, add, change/ upgrade, remove/ dispose, test, backup, repair/ recover [1,6,17,25].

It is noteworthy here that one or more processes of one or more process types can be defined for each instantiated IT service component, i.e. there can be an install and a move process for an IT service component and there can be several

[1] Because of space restrictions we only present groups of properties. Each group consists of many properties.

[2] Common economic properties (e.g. tariffs, price models) are not listed

[3] tpmC is a database benchmark unit (www.tpc.org)

install processes (=process variants) for one IT service component. As IT service components include technological and organizational aspects both a different type of abstract IT service component (technological perspective) and a different process or process variant (organizational perspective) leads possibly to a new IT service component. The service engineer decides whether a new IT service component is created. This chapter has primarily focused on atomic IT service components and how they can be derived. The typologies and properties are used to structure the entire set of IT service components. The next chapter focuses on composite IT service components.

V. Component-Based Construction / Design of IT Service Products

The component-based construction or design of IT service products relies on two design principles: the composition of IT service components and the substitution of IT service components [26]. Adjustments of the internal structure of IT service components are only allowed on copies of the considered IT service components. This restriction bases upon the black-box principle of modularity where IT service components hide their internal structure [2,9,12,14]. We do not deny that several reuse mechanisms can also be applied (e.g. as described in [5]) at the design stage of new IT service components but this contribution focuses on the sole composition of yet developed IT service components, its description and its structured storing.

Precondition for applying the design principles (composition and substitution) is conformity or compatibility of the interfaces of two or more IT service components to be combined. In the case of composition this is fulfilled if a component A entirely or partly provides the demanded services of component B. Furthermore, component A must own a supplier interface that matches or is conform to the demand interface of component B [26]. The composed IT service product finally consists of IT service components where demands are partly or fully satisfied by provided services of other IT service components that were combined together. If not all demands are satisfied two possibilities exist to handle the situation: the demand can be fulfilled by increasing the quantity of several IT service components (e.g. 100 Gbyte of storage are required, the used IT service component only provides 50 Gbyte of storage) or the IT service product inherits the remaining demands. This consequently leads to dependencies between IT service products. The second solution is called “factoring out” of IT service components and allows extra potential of reuse. The inherited demands are treated in the same way as demands between IT service components [29]. The demands on IT service product level are fulfilled by preliminary ordered IT service products, e.g. before the installation of a database system the installation of all necessary network and storage services is required. It is therefore inevitable to manage the customer’s preceding bought IT service products in a so-called installed base [6].

An IT service component A can be substituted by another IT service component B if both components are of the same type. Type conformity persists if component B provides equal or more services and requests fewer services than component A. Moreover, component B must supply a specialized interface for its provided services and a generalized interface for its demanded service(s) [29].

IT service component types as well as interface types are specified in help with the previously illustrated typologies and properties. Specifying the interface in depth will restrict and reduce the number of suitable IT service components. For instance, if an application service component requires data encryption an applicable network service component must provide corresponding techniques. All “non-encrypting” network service components cannot be chosen.

As stated before the interface of IT service components must also cover organizational issues. According to Corsten et al. we assume that modularization of IT service components results in the necessity of coordination [14]. Coordination implies the alignment of activities to achieve a common target whereas objective and social causes can be considered. Social causes (asymmetry of information and complementary personal targets) are not in scope of the current contribution. Objective causes are further distinguished in resource, target and output interdependencies [15]. Resource interdependencies emerge if two or more production processes of IT service components demand the same, limited resource, e.g. a database server instance or high qualified employees. Although this is indeed determined in the design stage of IT service products it is typically object of capacity management processes [24]. This topic is addressed by the properties of capacity assigned to IT service components. Target interdependencies are observed if the contribution to profit of production processes of IT service components is interdependent of the parameterization of other production processes. This commonly leads to cost and profit effects and does not further restrict the technological or organizational compatibility of two IT service components. We therefore do not focus on this type of interdependencies. Output interdependencies exist if the output of production processes of two or more IT service components are interdependent. This type of interdependencies leads to several restrictions concerning (1) the execution order of production processes (chronological with / without given sequence or parallel), (2) the compatibility of two production processes (mutual required, mutual exclusive) and (3) possible process variants. We therefore propose to enhance the previously presented definition of interfaces to cope with output interdependencies of IT service components.

First of all we enhance the interface definition to address the first restriction. Each interface is therefore typed with the corresponding process execution restriction, e.g. the demand interface “network service” of a “server service” is typed as “in advance” in order to ensure that an IP address is preliminarily reserved when the “server service” is deployed. For instance a demand interface is typed as “parallel” if an “application service” requires x Gbyte of database storage but the execution sequence of the install processes is negligible as long as all install processes are finalized when the user wants to consume the IT service product.

Secondly we introduce a "*constraining interface*" that describes other IT service components that cannot be composed with the considered IT service component, e.g. it is notpossible to combine a "network attached storage" (a further specified IT service component of type "storage service" in Figure 1) with a "shared server service" (of type "server service"). This addresses the restriction "mutual exclusive" of two production processes whereas the restriction "mutual required" can already be represented by the previously adapted mechanisms of supply and demand interfaces.

The third restriction is addressed by two mechanisms: firstly process properties that determine which variant is chosen are handled in the same way as technological properties, i.e. a demanding IT service component sets the property to its corresponding values. This is often done by requesting a specific quantity or quality. Secondly the decision is delegated to the IT service product and either resolved by establishing interdependencies between IT service products or it must be specified by the service engineer (in the design stage) or the service consumer (when ordering the IT service product). The interdependencies are handled with concepts of variant configuration.

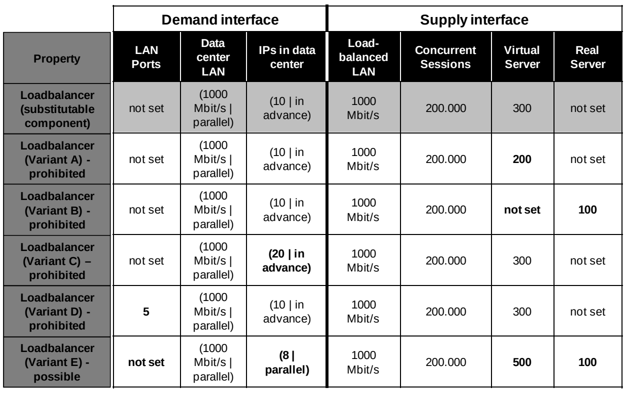
VI. Application of the Proposed Approach on a Company Example

Due to space restrictions, we focus in the following example on the interface specification of the presented approach. It is self-evident that the IT service component typology (Figure 1) and the presented process typology can be used to structure the set of IT service components. The following example is based on data of a leading IT service provider in Europe. For reasons of confidentiality we alienated some quantity and quality data and simplified the interface descriptions to avoid waste of space.

Figure 2 gives an insight how we applied the proposed approach to the interface description of the IT service component of an IT service provider. The presented examples are different variants of a "loadbalancer service" that possess a demand interface and a supply interface. The second row describes possible properties how the interfaces are defined (e.g. LAN Ports, Data center LAN and so on). The substitutability of the defined IT service components is evaluated in regard to the first specified IT service component ("Loadbalancer (substitutable component)"). Each property can be specified with technological and/ or organizational values, e.g. the required "IPs in data center" are characterized by the "technological required quantity" and whether the process of "IP registration" must be operated "in advance" or "parallel" to the installation of the "Loadbalancer service".

The following enumeration explains the evaluation:

- Variant A – substitution prohibited; Reason(s): supplying less service ("virtual server")
- Variant B – substitution prohibited; Reason(s): supply interface is generalization ("not set virtual server") although offering more service and specialized supply interface ("real server")
- Variant C – substitution prohibited: Reason(s): demanding more service ("IPs in data center")
- Variant D – substitution prohibited; Reason(s): demanding more service and demand interface is specialized ("LAN Ports")
- Variant E – substitution possible; Reason(s): demanding less service and less organizational restriction ("IPs in data center", amount and parallel execution of processes possible); supplying more service ("virtual server", "real server"); supply interface is specialized ("real server")



	Demand interface			Supply interface			
Property	LAN Ports	Data center LAN	IPs in data center	Load-balanced LAN	Concurrent Sessions	Virtual Server	Real Server
Loadbalancer (substitutable component)	not set	(1000 Mbit/s | parallel)	(10 | in advance)	1000 Mbit/s	200.000	300	not set
Loadbalancer (Variant A) - prohibited	not set	(1000 Mbit/s | parallel)	(10 | in advance)	1000 Mbit/s	200.000	**200**	not set
Loadbalancer (Variant B) - prohibited	not set	(1000 Mbit/s | parallel)	(10 | in advance)	1000 Mbit/s	200.000	**not set**	**100**
Loadbalancer (Variant C) – prohibited	not set	(1000 Mbit/s | parallel)	**(20 | in advance)**	1000 Mbit/s	200.000	300	not set
Loadbalancer (Variant D) - prohibited	**5**	(1000 Mbit/s | parallel)	(10 | in advance)	1000 Mbit/s	200.000	300	not set
Loadbalancer (Variant E) - possible	**not set**	(1000 Mbit/s | parallel)	**(8 | parallel)**	1000 Mbit/s	200.000	**500**	**100**

Figure 2: Interface defintion and substitutability of IT service components

As illustrated, the approach can be used for the specification of the technological and organizational interface as well as the evaluation of substitutability of two IT service components. "Constraining interfaces" are specified in the same way, e.g. a demand interface of a "network attached storage" possess the property "customer assignment" with the value "shared | mutual exclusive". This parameterization implies that all IT service components that define "shared customer assignment" in their supply interface cannot be combined.

In Figure 2 we combined technological and organizational aspects in one interface. Another possible solution is to split the interface and assign two interfaces to each IT service component.

VII. Conclusion, Benefits, Restrictions and Further Research

As response to the trade-off between customer-orientation and standardization IT organizations and IT service providers decompose their monolithic IT services in modular IT service components. The IT service components are afterwards used to compose and configure IT service products. Customer-specific variants, especially in the area of process-oriented IT service products, leads to an increasing number of IT service components. To realize potentials of standardization and reusability the set of IT service components must be structured to support the IT service engineer in the process of composition.

In this context the presented approach addressed mainly two challenges: the description/ derivation of IT service components and the interdependencies between two or more

IT service components. Both aspects were examined regarding technological and organizational characteristics. The approach can be used to structure the entire set of IT service components. This was also demonstrated on a real world example.

Implementing such an approach will enormously decrease the time-to-market of new IT service products and the time-to-delivery of configurable IT service products because the effort for coordination and communication declines. Furthermore, the quality of IT service products will increase, especially of new IT service products because each demand of an IT service component is explicated and must be satisfied. This leads to a predictable behavior of the composed IT service product.

The approach must be evaluated and enhanced in a real company environment to gain further information on structuring IT service components and its interfaces. The huge number of IT service components can only be managed by corresponding IT systems. A next step will be to design and implement the presented mechanisms.

REFERENCES

[1] Bailey, J., Kandogan, E., Haber, E., and Maglio, P.P. Activity-based management of IT service delivery. Proceedings of the 2007 symposium on Computer human interaction for the management of information technology - CHIMIT ’07, (2007), 5.

[2] Balzert, H. Lehrbuch der Software-Technik - Software Entwicklung. Heidelberg, 2001.

[3] Bodendorf, F. Daten- und Wissensmanagement. 2005.

[4] Brocke, H., Uebernickel, F., and Brenner, W. Mass customizing IT-service agreements - towards individualized on-demand services. Proceedings of the 18th European Conference on Information Systems (ECIS 2010), (2010).

[5] Brocke, H., Uebernickel, F., and Brenner, W. Reuse-Mechanisms for Mass Customizing IT-Service Agreements. Proceedings of the Sixteenth Americas Conference on Information Systems, (2010).

[6] Brocke, H. Managing the Current Customization of Process Related IT-Services. Proceedings of the Hawaii International Conference on System Sciences (HICSS-43).

[7] Bullinger, H.-J., Fähnrich, K.-P., and Meiren, T. Service engineering - methodical development of new service products. International Journal of Production Economics 85, 3 (2003), 275-287.

[8] Bullinger, H.-J. Service Engineering - Entwicklung und Gestaltung innovativer Dienstleistungen. In Service Engineering. 2006, 3-18.

[9] Burr, W. Chancen und Risiken der Modularisierung von Dienstleistungen aus betriebswirtschaftlicher Sicht. In Konzepte für das Service Engineering. Springer, 2005, 17–44.

[10] Böhmann, T. and Krcmar, H. Servicedatenmanagement für modulare Dienstleistungen. In Betriebliche Tertiarisierung. Duv, 2004, 149-178.

[11] Böhmann, T., Loser, K.-U., and Krcmar, H. Modellierung von Prozessschnittstellen modularer Servicearchitekturen. In Konzepte für das Service Engineering. 2005, 167 - 186.

[12] Böhmann, T. Modularisierung von IT-Dienstleistungen: eine Methode für das Service-Engineering. Gaber Edition Wissenschaft, 2004.

[13] Codd, E.F. A relational model of data for large shared data banks. Communications of the ACM 13, 6 (1970), 377-387.

[14] Corsten, H., Dresch, K.-M., and Gössinger, R. Gestaltung modularer Dienstleistungsproduktion. In Wertschöpfungsprozesse bei Dienstleistungen. Springer, 2007, 95–117.

[15] Ewert, R. and Wagenhofer, A. Interne Unternehmensrechnung. Springer Berlin Heidelberg, Berlin, Heidelberg, 2008.

[16] Fogl, F., Winkler, T., Böhmann, T., and Krcmar, H. MoSES – Baukastensystem für modulare Dienstleistungen Ebenen und Beteiligte. In Konzepte für das Service Engineering. 2005, 85-100.

[17] Garschhammer, M., Hauck, R., Hegering, H.-G., et al. Towards generic service management concepts a service model based approach. 2001 IEEE/IFIP International Symposium on Integrated Network Management Proceedings. Integrated Network Management VII. Integrated Management Strategies for the New Millennium (Cat. No.01EX470), , 719-732.

[18] Grawe, T. and Fähnrich, K.-P. Wissensgestützte Konfiguration komponentenbasierter IT-Dienstleistungen in Wertschöpfungsnetzen. In Content- und Wissensmanagement. 2003, 139-147.

[19] Grawe, T. and Fähnrich, K.-P. Service Engineering bei IT-Dienstleistern. In Entwicklung IT-basierter Dienstleistungen. 2008, 281-301.

[20] Herrmann, K., Klein, R., and The, T.-S. Computer Aided Service Engineering – Konzeption eines Service Engineering Tools. In Service Engineering. 2006, 649-678.

[21] Hevner, A.R., March, S.T., Park, J., and Ram, S. Design science in information systems research. Mis Quarterly 28, 1 (2004), 75-105.

[22] Kaitovaara, P. Increasing Business-Relevancy to the IT Service Product with the Support of Packaging of IT Services. 2001.

[23] Nieminen, P. and Auer, T. Packaging of IT services. 190 (1998).

[24] Office of Governement Commerce. IT Infrastructure Library - ITIL V3 (Service Strategy, Service Design, Service Transistion, Service Operations, Continous Service Improvement). 2007.

[25] Oliveira, F., Nagaraja, K., Bachwani, R., Bianchini, R., Martin, R.P., and Nguyen, T.D. Understanding and validating database system administration. Proceedings of the 2006 USENIX Annual Technical Conferenc, (2006), 213-228.

[26] Overhage, S. Vereinheitlichte Spezifikation von Komponenten: Grundlagen, UNSCOM Spezifikationsrahmen und Anwendung. 2006. http://deposit.ddb.de/cgi-bin/dokserv?idn=983172935&amp;dok_var=d1&amp;dok_ext=pdf&amp;filename=983172935.pdf.

[27] Piller, F. and Waringer, D. Modularisierung in der Automobilindustrie: neue Formen und Prinzipien: modular sourcing, Plattformkonzept und Fertigungssegmentierung als Mittel des Komplexitatsmanagements. 1999.

[28] Pine, B.J. and Davis, S. Mass Customization: The new frontier in business competition. 1993.

[29] Seco, J.C. and Caires, L. A basic model of typed components. ECOOP 2000—Object-Oriented Programming, (2000), 108-128.

[30] Shimomura, Y., Sakao, T., Sundin, E., and Lindahl, M. A design process model and a computer tool for service design. ASME 2007 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference, Las Vegas, NV, (2007), 1-13.

[31] Zarnekow, R., Brenner, W., and Pilgram, U. Integrated Information Management. 2006.

Driving Service Agility:

A longitudinal case in Yell

Dr Sharm Manwani
Henley Business School
University of Reading
UK
sharm.manwani@henley.com

Robert Carr
Head of Information Services
Yell
Reading, UK

Abstract. **Organizations require effective service management in order to meet business service levels and reduce costs in the operation of information systems. There is a growing body of knowledge that describes the rationale and the outcome of these experiences. These cases indicate that the capabilities and processes of the organization are important factors in achieving success. Our review of the literature considers both the hard and soft factors such as service processes and trust in service partners. These factors are explored through a longitudinal case study designed to provide insights into how the environment sets the parameters for service management. The selected case analyses the organization changes to its service management approaches during a period of several years. Results are discussed from both practitioner and theoretical viewpoints with proposals for further research.**

Keywords-component; service management, frameworks, agility

I. SERVICE MANAGEMENT INTRODUCTION

Services are becoming increasingly pervasive in organizations and society, driven partly by technologies exploiting the internet, mobile computing and enhanced security software (1). Information technology is also changing the nature of work in services organizations yet there are many challenges for IT service providers that make it difficult to gain trust from their customers. In the implementation stage, training and trials help to deliver a better service to customers; support needs to be maintained in the operational stage. This focus on service combined with building relationships is most likely to engender trust (2).

A broader approach is required to effectively deliver the whole package of IT services to internal customers. There has been significant support for improving service processes, exemplified by the growing use of the ITIL (IT Infrastructure Library) framework which provides support for IT Service Management. ITIL version 2 distinguished between service delivery and support while version 3 has expanded the scope into a lifecycle approach that starts with the service strategy. In version 2, service delivery has a focus on the end user as the customer whereas service support considers the business organization as the customer covering issues such as finance and overall availability (3).

Table 1 shows the target objectives of the processes within service delivery and support.

Table 1: ITIL Version 2 (adapted by authors)

Process	**Goal**
Service level mgt.	Create and monitor service level agreements
Financial management	Budget and monitor the financial aspects
Capacity management	Match IT capacity to agreed service levels
IT Service continuity mgt	Manage risks that impact IT services.
Availability mgt.	Define & monitor availability of IT services.
Service desk	Address user queries and issues
Incident management	Capture and monitor incidents to resolution
Problem management	Address underlying problems causing incidents
Change management	Manage approved changes to the service
Release management	Implement new releases of the system
Configuration mgt.	Track components of the IT services

One of the key service management decisions for an organization is how to source its capability. Three strategic intents can be articulated for appropriate sourcing: firstly improving the performance of IS, secondly exploiting IT assets and thirdly using IT to achieve better business results (4). A survey of organizations found over 30% cited cost savings as a reason for outsourcing, and nearly 20% wanted access to skills (5). On the benefits side, over 10% cited service quality, and over 6% mentioned skills and knowledge as the areas with which they were most satisfied.

Although key, this does not address the need for agility, defined as the ability to deal with fast business change (6). The core question addressed by this research is how to enhance service agility to satisfy business demands. Based on this question of service management, the next section proposes a research method and then describes key elements of the case prior to an evaluation discussion and conclusions.

ISBN: 978-1-61208-116-8

II. RESEARCH METHOD

A. *An Explorative Method and Longitudinal Case*

Our aim is to apply a longitudinal perspective as an aid to understanding the service agility approaches in an organization. Given the exploratory nature of this research, we selected a single case study to aim for a depth of understanding.

"The critical questions seem to be ones of information management strategy….one way of doing such managerial research is through case studies. They allow multi-disciplinary, integrative enquiry…Longitudinal case studies therefore could be valuable if theory development is in part making sense of firms' actions." (7)

We believe that selecting an information-intensive organization as a case study supports this positioning of an integrative enquiry that provides a small step towards theory development.

B. *Selection of Yell as Subject Organisation*

The following description provides a historical view of Yell's journey as an organization.

'Yell began life in 1966 as a 'Yellow Pages' section in the Brighton telephone directory. Yellow Pages, as part of BT, grew to become the UK's leading provider of classified directory advertising and associated services. In April 2000 the Yellow Pages division of BT became Yell and in June 2001 Yell was purchased from BT by a consortium of private equity investors. In July 2003 Yell was listed on the London Stock Exchange and became Yell Group plc. Although we are best known for our printed Yellow Pages directories, we offer an integrated portfolio of printed, online and phone-based products and services. The printed Yellow Pages directories, our Yell.com website and our operator-assisted telephone information service 118 24 7, ensure people have access to the information they need whenever and wherever they want it, and provide advertisers with access to high quality sales leads. To find out more about our products and services, please see the Yell UK product pages. Whilst our company has been built on the strength of our brands, people are the heart of our business and as at March 2009 we employed more than 5,300 people in the UK. From our head office in Reading, Sales head office in Slough and seven main offices in key locations, highly-trained, professional teams work to provide a world-class service to our advertisers and users throughout the country.' (8)

Given this strong business focus on providing a world-class service, we consider that Yell is an ideal case for a service management study. Further, Yell's Information Services (IS) group can be regarded as an exemplar by virtue of winning the Computing 2004 IT Department of the Year award (9)

C. *Data Sources*

Since this is a longitudinal case, we deem that the perspective of the second author is an important factor. The second author was a senior manager in the IT group throughout the period of the case study. This assists in generating deep insights through an action research perspective which collects data on business problems in an organization setting and helps to address the shortage of practitioner-focused research (10).

The research has two main checkpoint periods. The historical information presented in this case relates to Yell's winning submission for Computing's IT Department of the Year award. Objective documented evidence was presented for this award that was independently assessed by experienced IT practitioner judges.

During late 2009 and early 2010, the second author commissioned a Service Management review which was undertaken by the first author.

The terms of reference covered the following scope (11). 'Yell operates in a competitive marketplace which with changing technologies is more uncertain than in previous years. In response, Yell is undertaking a major business transformation which has major implications for IS, both in terms of enabling key programmes and its internal operating model. This will require IS to be more agile in order to deliver tangible business benefits in a faster and more flexible manner. The review has the goal of evaluating the requirements of the path towards agility and the capability to meet these requirements. The capability review will cover the hard and soft elements.'

The second phase field research involved six interviews: three with senior IT service managers and three with business stakeholders of the IS groups. Alongside this there were three focus groups held with different teams in the IT service management groups. Additionally access was provided to key documents such as service level agreements and service performance reports.

Three main questions were addressed by the research.

1. What does agility mean for service delivery?
2. What are the issues related to service agility?
3. How can constraints be removed and agility improved?

III. YELL'S SERVICE MANAGEMENT JOURNEY

A. *Periods of Change*

Yell's market, business and IT challenges evolved during the first decade of the new millennium. In hindsight, we can identify two major periods that drive the key service management approaches and results

B. *Best-in-class processes: 2002-2007*

During this first period, Yell was operating in a buoyant marketplace. Hence it had the financial standing to make major investments in Yell's infrastructure and processes. Yell's submission for the Computing award confirmed this. 'IS provides the operations and CTI (computer-telephony integration) infrastructure for these on-line services, with best-in-class availability. Results did not happen overnight. Industry standard (ITIL) processes have evolved since 1998 and IS is now rated best in class in service management. Processes underpin measurable and improved service level agreements. Yell's commitment is emphasized by all senior IS managers gaining at least the Foundation Certificate.'

The results of this approach are exemplified by Figure 1 which shows the top 20 ITIL performers based on audits conducted by an independent organization.

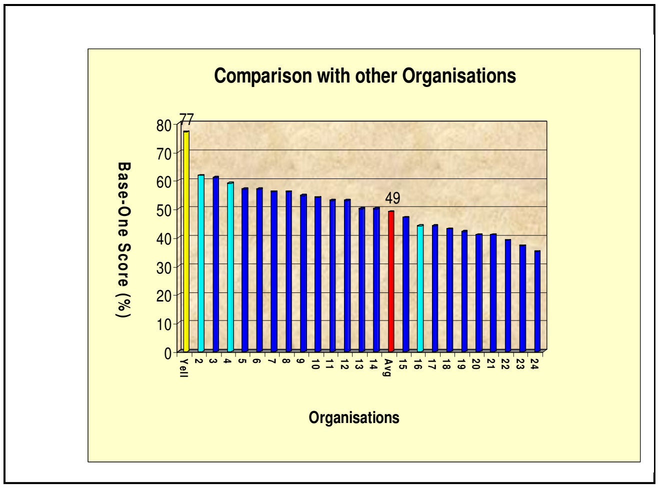


Figure 1: Comparison of ITIL process capability

The chart highlights that the Yell IS ITIL process rating of 77% was 13% higher than the next highest performer.

Also included in the submission were service performance reports showing that operational availability was consistently close to 100%. While service management was not the only factor in Yell winning the prestigious IT Department award, it was certainly a major contributor.

During the latter part of this period. the IS organization looked to improve service without increasing costs, moving towards an IS-Lite model. This was first proposed in a Gartner research report and promised greater IS agility and cost efficiency. Linked to this goal, Yell was able to leverage its mature processes when selecting outsourced partners. As a result, Yell significantly reduced their fixed costs moving to a larger proportion of variable costs. In summary, IS was able to achieve a consistently high performance due to the maturity of its processes in line with ITIL. This process competence was important in the next phase with greater demands for agility and cost reduction.

C. Responding to Market Changes: 2008-2010

During this second period, it became increasingly clear that Yell's service management needed to respond to a changing marketplace. There was an evolution from a position where the printed book was central to the product proposition to one where the internet became a growing service offering.

A key implication of this change was more challenging service pressures geared to meeting on-line service levels for customers. Critically, it involved moving from a stable and profitable print model to a very competitive new on-line model. This dual pressure for a more responsive service with lower cost investment necessitated a different approach, hence the focus of the review on service agility. The findings from this review follow.

1) What does agility mean for service delivery?

One definition of IT agility is the ability of a firm to adapt its IT capabilities to market changes (12). The interviews and workshops with IT service managers in Yell provided a variety of definitions for agility. At the basic level, it was seen as speed or reaction and quick delivery from Yell IS services to problems and new demands from internal and external customers.

2) What were the issues with service agility?

One key issue identified by the business stakeholders was how to measure service performance. Traditionally, this is supported through creating and monitoring service level agreements. These may be formal contracts in the case of external suppliers although this is not a substitute for effective relationships (13).

The Service Level Agreements (SLAs) in Yell typically measured overall availability of the services over a period of time, such as a month. A few minutes downtime at the end of the month would mean that the service level was met, but when this happened at a critical financial month-end process, the impact was highly significant.

The customer services manager made another point about measuring performance. She highlighted that having the system available for use was not enough if the response time was too slow or if the quality of data in the system was poor. This latter issue is potentially a major headache for IT service departments since data ownership can be unclear, and this was partly the case in Yell. It was also felt that tangible measures of performance did not address the perceptions of users about the responsiveness of the IT service personnel.

3) How can agility be improved?

Through the interviews and the focus groups, three opportunity areas were identified: people, process and technology. The people perspective included the use of flexible resources (servers, people, network etc) and processes to allow headroom to enable reaction to change. Flexibility is enhanced by having multi-skilled individuals, not operating in silos. Skills are not enough; people need to demonstrate the right sort of behaviors including a 'can-do attitude.

From a process perspective, this needs to apply across all the areas: the capturing and prioritization of demand, the scheduling and resourcing of work, and the evaluation of performance. Just as systems development has extended structured processes into agile processes, the same applies to service management. It is the appropriate mix of structured and agile that ensures effective service delivery. For example, pre-planned maintenance windows, where work can be slotted in, help to enable timely systems enhancements.

It is also important to make best use of the technology to enhance agility. This might mean having hardware that provides room to increase transaction. However in today's world the processing power is more likely to be outsourced and this was true in Yell with an exploration of the benefits of cloud computing.

Standardization is sometimes seen as the antithesis of agility since it limits the variety of available hardware and applications. However Yell's service management head argued that having a standard platform made the addition of new applications and services much easier. A similar argument was put forward for simplifying technologies whenever possible.

D. Rethinking Service Level Agreements (SLA)

It was recognized by the Yell service management team that the existing SLA format was not designed to meet the customer perception of usability nor was it appropriate to encourage service agility. The document was several pages long partly because it repeated many of the core infrastructure service levels rather than focusing on the specific service. Furthermore it only had availability measures and did not highlight what needed to be done to improve the service.

A new format was created as summarized in Table 2 which contains sample data for one of the services. This distinguished clearly between a comprehensive but concise description of the service and an analysis of performance from a customer perspective for review with the business owner.

Table 2: Revised service level agreement

AREA	**GOAL/COMMENTS**
Description	Provision of application to sell adverts
Business Owner	Telesales Manager
IS Accountability	Service Manager for Sales function
Responsible Support Roles	Servers & Network Telephony. BW. SAP & ORACLE. System Delivery. Desktop Support. Application Support
Key Customers	Telesales
Frequency/Trigger For Service	Daily - BW data load to update the data store.
Scope	Aid Telesales to sell and renew ads with high customer responsiveness
Impact Of Failure	Unable to provide timely information to Telesales, resulting in loss of revenue
Dependencies	Shared core infrastructure
Status & Plan	Changing telesales requirements need to be captured
Service Review Actions	Worklist response times poor.
Current Issues	User interface needs to provide targeted information in a faster manner
Metrics	
Availability	Good availability
Performance	Acceptable response times
Application Quality	No major bugs
Information Quality	Issues with customer master data
Responsiveness	Telesales looking for quicker response to current problems

This new format was welcomed both by the IT service managers and the key business stakeholders. It provided an insight into the core goals and issues, supporting a more agile approach to service changes.

IV. DISCUSSION AND CONCLUSIONS

In this section we reflect on the differences in the two periods and the insights to be gained from Yell's approach to driving service agility.

A. A ContextualApproach

This paper presents a rich longitudinal case which, in the authors' view, generates key insights into the importance of a contextual approach to service management. There is a tendency with current service management practices to assume that a greater focus on predictable processes is the direction that IT departments should aim target. This is exemplified by the Capability Maturity Model which has five levels related to the process maturity of the IT organization (14). Further the use of generic service level measurements for availability often implies that a very high service is the ultimate target of an organization.

This traditional view was supported in our evaluation of the first period under review. Yell scored exceptionally high on both process and availability dimensions. It evidenced best-in-class ITIL processes and consistently high service levels in winning the IT Department of the Year Award. In this type of environment, the need is for IT professionals whose core competence is to understand and follow processes. Yell's service team demonstrated this capability.

However, the challenging market environment in the second period necessitated a different approach. The need for the business customers of IS was to respond quickly by delivering new products and services. In this context, best-in-class but rather rigid processes and very high but expensive service levels were not the most appropriate way of meeting the business needs and building trust with internal and external customers. Recognizing this shift, Yell IS's focus moved to more flexible and cost-effective service management. This had significant implications for measurements, processes and people. In effect, Yell targeted both time and range ability (12) with the aim to respond quicker but also provide a wider range targeted at the specific service needs.

Furthermore, through the revised service level agreements, IS nominated service managers who were accountable both for the service and for client relationship management with the business owner. The revised SLA and the interviews with business stakeholders demonstrated that this approach was paying dividends. Our conclusion is that the dual perspective on improved service process and enhancing service relationships aligns with the service management theory on building trust, although we have applied it in a different context. The comparison of the two periods highlights the need for a contingent approach to service management, one that recognizes the different business environment.

B. *Practitioner Insights*

We propose three main practitioner insights from the research. The measurement of a service is a point of much debate. From a customer perspective, service is seen to comprise the following measures: Availability, Performance, Application Quality, Information Quality and Responsiveness to Customers. We conclude that all of these areas need to be right in order to achieve customer satisfaction. Furthermore, the service has to be targeted to the business benefits of the organization by demonstrating how these are being realized or impacted by failure.

The second insight is that historically, IS has typically looked inwards by addressing service problems that are linked to a technology component without considering the impact on related services and hence all of the customers of that service. In practice, accountability for delivery of customer facing and component services is shared between several IS departments. No single IS department has the end-to-end responsibility or the line of sight for these services. Hence it is important to define what components constitute a services and who is accountable for each of the services. Furthermore, this service needs to be defined and focused on the end customer experience.

Our third insight was that customer facing services could be categorized as Products, Public Facing Services, Business Processes and User Productivity. Each of these service groups has different characteristics hence they required tailored approaches to enhance services. Yell product services are for consumers such as Yell.com. Public-facing services are for customers and partners such as E-Channel. Business process services are for internal customer functions and rely on IT applications such as SAP R/3 and CRM. User productivity services are aimed at individuals and include email and personal devices. We found that it is generally easier to take an end-to-end view for user productivity services than it is for services that rely on a complex set of integrated applications. Therefore a tailored and contingent approach is needed.

C. *Limitations and Future Research*

Our goal was that the analysis should provide insights into service management approaches within a context of changing market pressures. The senior management role of the second author helps in gaining insights but increases the risk of a biased viewpoint. We believe, by using data from an award submission backed by documentary evidence, that this offsets the concern about potential bias in the first period covered by the research. In the second period, the combination of interviews and focus groups led to challenging debates and insights which provided diverse inputs to contrast with the views of the second author.

Working with a single organization provides a depth but not a breadth of study. The authors therefore recommend further longitudinal research in other organizations to expand the data and provide a contrast to the experiences of Yell. In particular, we would welcome research into public sector and not-for-profit organizations to assess if similar patterns of market stimulus and service management response exist in both types of organizations. This should be extended to international settings.

REFERENCES

[1] I.R. Bardhan, H Demirkan, P.K. Kannan, R.J. Kauffman and R. Sougstad, An Interdisciplinary Perspective on IT *Services Management* and *Service* Science Journal of Management Information Systems, Spring2010, vol. 26 issue 4, pp. 13-64

[2] M.M. Montoya, A. P. Massey and V. Khatri, Connecting IT Services Operations to Services Marketing Practices, Journal of Management Information Systems, Spring2010, Vol. 26 Issue 4, pp. 65-85

[3] T. Wui-Ge; A. Cater –Stel and M. Toleman, Implementing IT Service Management: A Case Study Focusing On Critical Success Factors, Journal of Computer Information Systems, Winter2010, Vol. 50 Issue 2, pp.1-12

[4] Di Romualdo, A. and Gurbaxani, V. (1998), "Strategic intent for IT outsourcing", Sloan Management Review, Vol. 39 pp.67-80

[5] Benn, I. and Pearcy J (2002). Strategic Outsourcing: Exploiting the Skills of Third Parties. London, MCA/Hodder & Stoughton:

[6] Ren, M. and Lyytinen, K (2008) Building Enterprise Architecture Agility and Sustenance with SOA. By:. Communications of AIS, Vol. 2008 Issue 22, p75-86

[7] Earl M J and Feeny D F (1996), 'Information Systems in global business: Evidence from European multinationals', Information Management: The Organisational Dimension (editor M J Earl), Oxford University Press.

[8] Yell website, http://www.yellgroup.com/english/aboutyell, October 2010

[9] Yell website, http://www.yellgroup.com/english/media-pressreleases-2004-yellinformationtechnologyteamwinstop, October 2004

[10] DeLuca, D; Gallivan, M.J. and Kock, N. (2008) Furthering Information Systems Action Research: A Post-Positivist Synthesis of Four Dialectics, Journal of the Association for Information Systems, Vol. 9 Issue 2, pp 48-71

[11] Yell internal document, Terms of Reference for Service Management Review

[12] Sengupta, K and Masini, A. (2008), IT agility: striking the right balance, Business Strategy Review, Summer2008, Vol. 19 Issue 2, p42-48

[13] Goo, J; Kishore, R; Rao, H. R.; Nam, K. (2009), The Role Of Service Level Agreements In Relational Management Of Information Technology Outsourcing: An Empirical Study MIS Quarterly, Mar2009, Vol. 33 Issue 1, p119-145,

[14] Carnegie Mellon, Software Engineering Institute. Capability Maturity Model® for Software (SW-CMM®). Retrieved on August 18, 2010 from: http://www.sei.cmu.edu/cmm/

An IT Service Reporting Framework for Effective Implementation of ITIL Continual Service Improvement Process Conforming to ISO/EC 20000

Mohammad Kajbaf, Negar Madani, Ali Suzanger
ITSM department
Infoamn IT Consulting Co.
Tehran, Iran
{m.kajbaf, n.madani, a.suzangar}@infoamn.com

Shirin Nasher, Mehrdad Kalantarian
Computer Engineering Dept.
Iran University of Science and Technology
Tehran, Iran
{nasher, kalantarian}@vc.iust.ac.ir

***Abstract*—In this paper, an IT service reporting framework has been presented to help organizations in implementing IT service improvement process in accordance with ISO/IEC 20000 PDCA-cycle and reporting requirements. It defines six types of reports and includes guidelines for automation of these different types. Afterward, a process for reporting by focusing on defining report templates based on the organization requirements is provided. Proposed report types, process flow, ARCI matrix, and, process integration points as a general IT service reporting framework, helps organizations to organize their communication using reports.**

Keywords-IT service management; continual service improvement; service reporting.

I. Introduction

A. Service Improvement in Different ITSM Frameworks

Different ITSM frameworks and standards have discussed service measurement and improvement. ISO 20000 requires continual improvement of IT services via the PDCA cycle. It also defines requirements for service reporting to measure, analyze and communicate observations to help improvement of organization activities [1].

The continual service improvement (CSI) is one of the main aspects of ITIL v3; it includes a 7 step process for IT service improvement [2].

The CSI process includes 3 main activities: (1) defining metrics for measurement of activities and service performance; (2) monitoring, measuring, reviewing and reporting defined performance metrics; and, (3) taking corrective actions to improve service performance [3].

In MOF v4, the improvement concept is inherent in the MOF lifecycle, but there is no explicit process or service management function (SMF) in charge. Two management reviews cover measuring performance of IT services and processes, and taking corrective actions: the Policy and Control review and the Operational Health review [4].

In COBIT, the Monitor and Evaluate domain covers reviewing, monitoring and continual improvement of IT services. It includes defining performance indicators and reporting them, acting upon deviations, third-party reviews, and, integrating IT reporting with business reporting [5].

B. Toward an ITSM Management Model

The ITIL® framework is a major source of good practice in IT service management (ITSM) used by organizations worldwide [6]. ITIL defines many policies and key performance indicators (KPI) for different IT services and processes. Although ITIL describes policies and rules for service reporting process in the CSI book [3], it does not clearly define how the process must work.

Organization that are implementing ITIL for the first time, face new challenges. One of them is taking the correct approach in defining performance indicators, and measuring KPIs and reporting them. For most organizations the implementation of true common cross-organizational management processes may be the most difficult aspect of the ITSM project [7]. They need to know how to find which processes are working and which are not.

Keel *at al.* [8] discuss some main challenges in adopting ITSM. It suggests that successful implementation of ITSM strategy relies on the quality of processes.

Keeping sight of return on investment (ROI) and balance, of cost, time and quality is of a very high importance for companies [9]. Lahtela *et al.* [10] suggest an ITSM measurement system to support improvement activities. Lima *et al.* [11] propose a model to quantify IT service quality, to enhance the “Check” phase of PDCA cycle.

In Section II, description of the problem is provided. Main aspects of the provided framework are defined in Section III, and the process flow is presented in Section IV. Section V lists terms and definitions, and Section VI describes future works.

II. Description of the Problem

Many problems would arise during implementation of IT security and service management frameworks for customers, especially if they are implementing a business process model for the first time. Employees do not know what they must do, and managers do not know what to expect from them. One of the main areas of ambiguity is the improvement process. By investigating problems in implementing ITSM for customers, we established a reporting structure to decrease ambiguities.

At the first sight, the structure seems like some extra work required from employees. At final steps of the project, all parties can relate their day-to-day activities to concepts they were familiarized with during their ITIL/ISMS classes. For successful implementation, reports must be such that all parties have a common understanding on them.

The proposed service reporting framework is a tool for organizations to clarify their internal communications and simplify defining and monitoring KPIs and metrics, and thus the CSI process. The reporting process helps managers to define clearly what they expect in each activity.

ISBN: 978-1-61208-116-8

III. IT Service Reporting Framework

Our proposed service reporting framework contains 2 parts. First, a hierarchy of reports for different activities of ITSM is defined. This hierarchy includes six types of reports for different activities of operational, tactical, and strategic types. Then, a Service Reporting process is proposed for defining report templates according to the reporting requirements of the organization. This process could be implemented as a sub-process of ITIL continual service improvement process.

A. Report Types

There are different types of reports in an organization. Some contain low level details of daily observations, while the others are based on high level analysis and contain conclusions required for decision makers. Considering all the activities in the ITIL framework, and by investigating different types of reports communicated internally in our customer companies, six generic types of reports have been identified. Then, they were refined based on the level of activities, the material included in each report, required level of analysis on each and required competency for it.

1) Reporting routine tasks

This category of reports is of operational level, about happenings, successes and failures in on-going daily activities of operation and monitoring type. Audiences include section managers or process owners, whom are accountable for the task. Low amount of analysis is required, and, the report must include patterns, frequencies, time intervals, inconsistencies and extermums in observations. They are often scheduled for small periods of time, from a few times a day up to once in a few weeks.

2) Reporting assigned tasks

This category is of tactical level, about the progress of projects (for teams) or assigned tasks (for persons). These reports are usually in design, development, deployment and implementation activities. Audiences are project or function managers, or the change advisory board (CAB) (in the case of new or change services). They contain some basic analysis on the progress of the project/task, achievements, pitfalls, remained works, estimations, lessons learned, etc. They are always pre-scheduled based on the project/task steps or small periods of times from once in a few days up to once in a few months.

Two types of reports discussed above were about measuring activities in IT processes. The next two types, in contrast, are mostly about IT service measurement.

3) Reporting on events

This type of report is of operational level and is about low level data on events. The data are usually gathered in operation, monitoring and support activities and contain summaries of different types of events in IT services, specially security events, or details of a major/critical event. Detailed data included in reports are defined in the Event Management process.

Audiences are operation and related service managers. In the case of major events, managers of affected services, owners of related monitoring processes, and, the incident manager, are included.

In the case of normal events, summaries and number of different events are included in the report. In the case of major events, descriptions of events, decisions made, activities done, and the results gained, and some calculations on impacts and costs may be required. Normal reports on events are scheduled mostly for small periods of time, from daily up to weekly. Reports on major events are not scheduled but reactive due to the nature of events.

4) Reporting on services

This category is of a tactical level and is about service status data. The included data are gathered and analyzed in monitoring and service management activities. The reports include current and predicted service levels, statistical analysis and trends, user complaints and customer satisfaction measures, measurement of metrics defined in SLAs, OLAs and contracts, and breaches in agreed levels, as well as prediction of required resources and capacities against target levels.

Audiences include service managers and the portfolio, service level and CSI managers, as well as business or external customers about status SLA-signed services. The scheduled periods are of a medium level, from weekly up to once in some months. For business and external customers, reporting schedules are defined in agreements. Out-of-schedule reports may be required due to proactive monitoring or where service level breaches.

5) Reports on review meetings

These reports are of a strategic or tactical level and are produced after review meetings in different phases of IT service management. They always contain summary of discussions, ideas presented by different parties, and details of decisions made, tasks assigned, and further activities/meetings scheduled. Decisions must often be reflected in other types of documents, such as policies, plans, contracts, etc. Therefore, the report must include all required details discussed in the meeting. Report schedules depend on meeting schedules.

Audiences include all attendant parties, as well as, CSI manager, and other relevant IT managers. If decisions include changes to the existing documents, IT services or the infrastructure, the report must be provided to the change manager as well.

The meetings include, but are not restricted to, CAB and emergency CAB (eCAB) meetings, portfolio management meetings, service and technology designers meetings, meetings between development teams, meetings with suppliers and business or external customers, post implementation reviews, non-conformances, etc.

6) Management reports

This category of reports is of strategic level and is about all high level aspects of IT services and ITSM framework. Audiences are top IT and business management officers, the CSI manager, and other parties as defined in the *report template*. Management reports contain analytical data on the performance of IT services, trends, suggestions and prospects, costs, risks and values, performance and achievements of IT processes and functions according to defined KPIs, demand and market analysis, new methods and technologies, etc. Analysis on business achievements against strategies and objectives or new suggestions may be required.

In summary, this category of reports reflect all the other types of reports to the top IT and business managers. All six types of reports are presented in the table below.

TABLE I. SUMMARY OF DIFFERENT REPORT TYPE SPECIFICATIONS

Different Report Types				
Report Type	***Types of Activities***	***Types of Audiences***	***Included Material***	***Analysis level***
Routine Tasks	operational, monitoring	section managers, service owners	happenings, summaries, inconsistencies, observations	low
Assigned Tasks	development, deployment	project/section managers	progress, achievements and pitfalls, learnings	medium
Events	operational, monitoring, support	section managers, service owners	summaries, descriptions, actions done	medium
Service Status	service management, monitoring	service managers, business customers (against SLAs)	service status, statistics, customer satisfaction	high
Review Meetings	management, planning, relationships, development	Top management, design and planning teams, customers	decisions on policies, plans, objectives, strategies, agreements, …	high
ITSM Framework	planning, strategy, service management	Top IT/business managers	measured KPIs, trends, costs and benefits, conclusions	low, medium, high

B. Automation Level

Another factor which can be defined for reports is possibility of using automation tools. Different automatic tools for reporting network and security events and measuring service performance exist. Using these tools for reporting, helps employees in measurement and reporting and decreases human errors. They also ease scheduling of reports. Therefore, we suggest customers to use automated reporting systems, as long as it helps IT personnel in doing their activities at best.

For routine tasks, automated tools are widely present which provide data forms for employees to fill. These tools help employees to generate reports on schedules and usually help to communicate them directly to audiences.

A similar approach for reports on assigned tasks and review meetings could be taken. However it could not cover all aspects of these reports, because they include many analytical data. Therefore, often templates are prepared and rules are defined to fill them.

High levels of automation could be used in event reporting. Different tools for network and security events reporting exist. They automatically gather data and generate reports on their summaries, using different methods like web-services or intelligent agents [12-14].

These tools are often used to provide required data to operators to compose event reports. Therefore, audience of automated event reports is the operator/administrator in charge. In the case of major events, automated systems could alert the incident resolver in charge, the service owner, or event trigger an automated troubleshooting system.

Automated performance measurement systems could be used to gather and analyze service level data and help the report owner in composing the service status reports. they can automatically measure service levels and identify SLA breaches [15-16]. These systems could also alert service manager and person in charge, when a critical service fails.

Reports on ITSM framework are hard to automate. Different systems exist to gather and analyze data, draw charts and even analyze different aspects of a decision and suggest a choice. However, decision making is one of a few activities which still require a human in charge and cannot be fully automated.

Table II demonstrates possibility of using automated tools for each type of report.

TABLE II. POSSIBILITY OF AUTOMATION IN EACH REPORTING TYPE

Report Type	**Possible automation level**	**Usual type of automation**
Routine tasks	Medium	Form-based automation programs
Assigned tasks	Low	Pre-defined templates
Events	High	Automated event notification systems
Service status	Medium	Automated service level measurement
Review meetings	Low	Pre-defined templates
ITSM framework	Low	Pre-defined templates

IV. IT SERVICE REPORTING PROCESS FLOW AND ACTIVITIES

A detailed discussion of the processes and activities in the Report management process are provided in this section.

A. Service Reporting Policies

Service reporting policies must be clearly defined, and communicated to all IT personnel.

- Each report and report template must have a unique identifier. Reports must contain reference to their report template.
- Report templates should define purpose, audience, person in charge, required metrics and data sources of reports.
- All report templates must be clearly defined, agreed upon by all parties, and recorded in the configuration management database (CMDB). Also, all changes to report templates are subject to change management policies.
- All reports must be generated in time and contain accurate data, according to their report template.

B. Process flow activities

1) Define goals and objectives:

Several reports are prepared and distributed throughout the organizations. However, only the reports prepared based on specific goals and objectives add value to the business and lead to better performance, therefore “what the reports are going to present” and “what they are going to be used for” should be clearly determined in the report template.

This is mainly achieved by reviewing the portfolio and the business policies, rules, goals and objective that refer to reporting requirements and alignment of business and IT. There are three major aspects in reporting:

a) Reactive reports: reports on what has occurred

b) Proactive reports: reports on near breaches

c) Forward reports: reports on scheduled activities

2) Determine scope:

The scope of a report is the section that the report refers to. A report could be prepared for a process, service, activity, section, a specific department or the complete IT organization.

3) Select report types:

Depending on the goal and scope of the report, one or multiple report types are selected. The following steps should be taken for each of the selected report types.

4) Review prior identified requirements:

The prior identified requirements are extracted from the reporting requirements identified in the business and the designing processes such as service level management, availability management, capacity management, information security and service continuity management. These identified requirements are categorized and documented, and will be addressed in the report.

5) Define the measures and metrics

Each identified requirement is studied in order to determine the measures and metrics which present that requirement. These measures and metrics are included in the report.

6) Specify data sources

The value of a report highly depends on the owner and credibility of the evidence that supports its statements and content. Therefore the personnel assigned to prepare the report, the sources, evidence and basics of calculation should be included thoroughly in each report.

7) Review related schedules:

Schedules that refer to major activities or activities which produce or alter a considerable amount of information have a significant impact on report schedules; therefore they should be carefully reviewed and taken into account. These schedules mainly include data gathering schedules, processing schedules, analyzing schedules, management meetings, review meetings, purchase schedules, etc.

8) Specify audience:

Depending on the target audience, the reports are classified in three main categories:

The business category: including the customers, the internal providers such as the business managers and the external providers such as the stakeholders.

The senior IT management category: such as the CSI manager and the business/IT analyst.

The internal IT category: including the middle and low-level management and IT staff such as the service owner, service manager, service level manager, process owner, process manager and etc.

9) Report scheduling:

In order to schedule a report factors determined in the previous steps should be taken into consideration, including the goals and objectives, the report type, the related schedules and the audience.

10) Select communication methods & tools:

Whenever the report is generated by the responsible person, the report should be communicated to the target audiences. The communication may be through the following methods:

- Paper-based hard copies,
- Online soft copies,
- Web-enabled dynamic HTML,
- Real-time portal/dashboard

11) Determine access levels

It is obvious that the reports prepared for the higher level staff should not be accessed by the lower level. However, depending on the organizations rules and policies, some personnel are authorized to access certain reports while their parallel colleagues should not be authorized to access those reports.

12) Review and approve:

Before finalizing the report template, all fields discussed above should be checked and approved. After refining the report template, if necessary, it should be approved by higher authorities such as business, senior or IT management. After approval, the report template must be communicated to the corresponding parties.

The CSI manager or his/her representatives are responsible for monitoring of reports to be generated based on the related report template and to be communicated according to the reporting policies.

Figure 1 demonstrates the complete service reporting process flow.

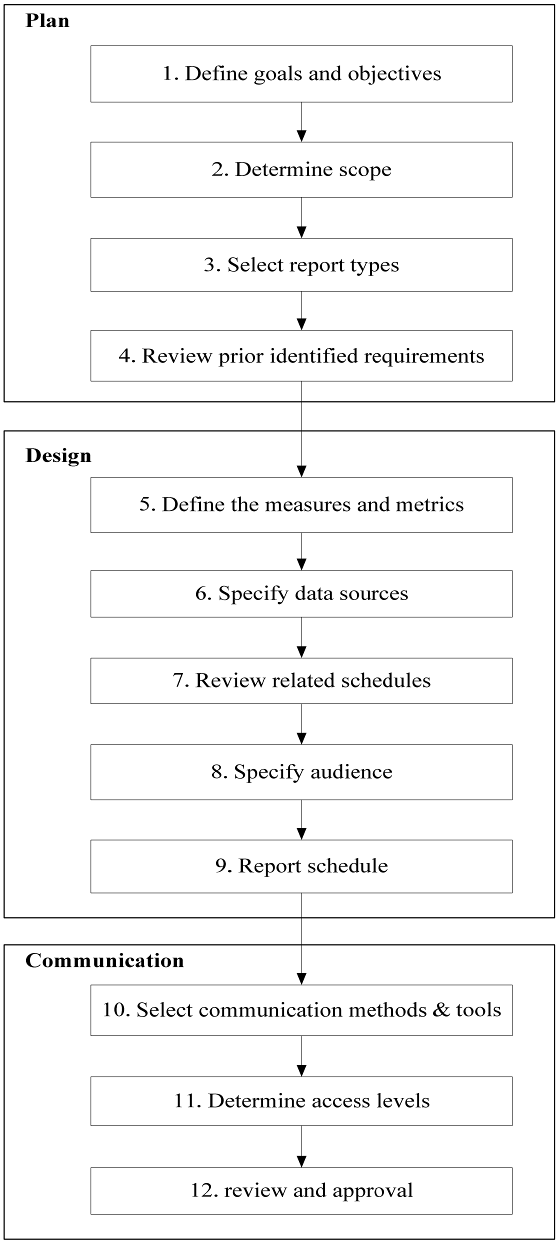

Plan
1. Define goals and objectives
2. Determine scope
3. Select report types
4. Review prior identified requirements
Design
5. Define the measures and metrics
6. Specify data sources
7. Review related schedules
8. Specify audience
9. Report schedule
Communication
10. Select communication methods & tools
11. Determine access levels
12. review and approval

Figure 1. Service Reporting Process Flow

C. Reporting process ARCI matrix

The CSI process owner is the person accountable for service reporting process. IT manager and report owner must cooperate to define right metrics according to business requirements. The information security manager should ensure data flow is in accordance with security policies.

Table III summarizes role of different stakeholders in service reporting process as an ARCI chart.

TABLE III. SUMMARY OF DIFFERENT REPORT TYPE SPECIFICATIONS

Reporting process activities		CSI manager	CIO	Report	Portfolio	Security	Report
A. Plan							
1	Define goals and objectives	A/R	C		C		C
2	Determine scope	A	C	R			C
3	Select report type	A	C	R			C
4	Review prior identified requirements	A	C	R	C		C
B. Design							
5	Define the measures and metrics	A	C	R			I
6	Specify data sources	A	C	R			C
7	Review related schedules	A	C	R			C
8	Specify audience	A		R			C
9	Report scheduling	A		R			I
C. Communication							
10	Select communication methods	A	C	R			C
11	Determine access levels	A	C	R		C	C
12	Review and approve	A/R		I	C		I
Legend	*A= Accountable,* *R= Responsible,* *C = Consulted,* *I = Informed.*						

D. Reporting Process Integration Points with other ITIL processes

Here, all documents which are communicated within the proposed service reporting process are listed.

Table IV presents all documents and information from different ITIL processes which are required in the service reporting process.

TABLE IV. SERVICE REPORTING PROCESS INPUTS

Inputs to reporting process	
Portfolio	-Business and IT rules, policies, goals and objectives -Business report requirements -Business and IT meetings and schedules
Demand	-Demand management requirements for a report -Demand management meetings and schedules
Financial	-Evaluation of reporting costs
SLM	-Service level agreements for the reporting process -SLM requirements for a report
Availability	-Level of availability of each report
Continuity	-The continuity required for providing a report
Security	-The security levels required for each report
Supplier	-Required reports for each supplier
Catalog	-Services available in the organization
Configuration	-CI's available in the organization
Release	-Progress of preparing the report template
Change	-Request for a new report template
Knowledge	-The accuracy of information received

Event	-Event schedules -Events related to inadequate reporting
Incident	-Incidents related to inadequate reporting
Problem	-Problems related to inadequate reporting
Request fulfillment	-The customer report requirements
Access	-The access levels for each report
CSI	-The collected, processed and analyzed data

Table V lists all information generated in the service reporting process.

TABLE V. SERVICE REPORTING PROCESS OUTPUTS

Outputs of reporting process	
Report Template for:	-performance against service level targets; -non-compliance and issues, e.g. against the SLA, security breech; -workload characteristics, e.g. volume, resource utilization; -Performance reporting following major events, e.g. major incidents and changes; -Trend information; -Satisfaction analysis. etc.

E. Service Reporting Process KPIs

- Reduction in the number of issues caused by inadequate reports.
- Improved decision making supported by efficient and effective reporting.
- Decrease gaps between the expected reports and the provided reports.
- Reduction in the number of unauthorized access to reports.
- Increase of relevant information reported to relevant audience.
- Reduced reporting costs.

V. TERMS AND DEFINITIONS:

- *KPI*: Key performance indicators are important metrics and measures used to report on the performance of process, IT service or activity.
- *PIR*: Post implementation review takes place after an activity has been completed and evaluates the success of the activity.
- *ARCI*: A Table that helps defining roles and responsibilities. ARCI stands for Accountable, Responsible, Consulted and Informed.
- *Report template*: A template is defined for each reporting purpose required in the organization. A report template is the result of the report management process.
- *Report designer*: Is a role assigned by the CSI manager which is mainly responsible for preparing the report template.
- *Report owner*: A role responsible for reporting based on the determined and assigned report template.

VI. CONCLUSION AND FUTURE WORK

Measurement and reporting is a key requirement of service improvement, which in turn is required for value creation and success in a competitive market. The proposed reporting framework helps organizations to define reports in a way that eases service improvements.

Many of the proposed concepts could be used for external reports, which are reports on service levels to clients. Further works can be done to classify and clarify external reports. More works are required to identify different types of reports required in different IT process and types of metrics and measures which can be defined for each of them.

REFERENCES

[1] International Standard Institute, "ISO/IEC 20000:2005 Information technology — Service management standard." 2005.

[2] The Office of Government Commerce (OGC), "ITIL Service Management Practices v3 - Core Books", The Stationary Office, UK, 2007

[3] The Office of Government Commerce (OGC), "Service Improvement Book", in ITIL v3 Service Management Practices, The Stationary Office, UK, 2007

[4] Microsoft® Operations Framework v4: Microsoft Publications, 2008.

[5] The IT Governance Institute®, ITGI, "COBIT 4.1 - IT Governance Framework." 2007.

[6] ITIL® V3 : Managing Across the Lifecycle Best Practices: The Art of Service Pty Ltd, 2007.

[7] M. Jantti and K. Kinnunen, "Improving the Software Problem Management Process: A Case Study," in the 13th European Conference on Software Process Improvement, Joensuu, Finland, 2006, pp. 40-49.

[8] A. J. Keel, M. A. Orr, R. R. Hernandez, E. A. Patrocinio, and J. Bouchard, "From a technology-oriented to a service-oriented approach to IT management," IBM Systems Journal, vol. 46, pp. 549-564, 2007.

[9] L. Haber. (2003). How do you measure ITSM success. ITSM Watch, http://www.itsmwatch.com/itil/article.php/3291421/How-Do-You-Measure-ITSM-Success.htm, (Dec 14, 2010).

[10] A. Lahtela, M. Jantti, and J. Kaukola, "Implementing an ITIL-based IT Service Management Measurement System," in Fourth International Conference on Digital Society, St. Maarten, 2010, pp. 249-254.

[11] A. S. Lima, J. N. de Sousa, J. A. Oliveira, J. Sauve, and A. Moura, "Towards business-driven continual service improvement," in Network Operations and Management Symposium Workshops (NOMS Wksps), 2010 IEEE/IFIP, 2010, pp. 95-98.

[12] C. A. Carver, J. M. Hill, J. R. Surdu, and U. W. Pooch, "A methodology for using intelligent agents to provide automated intrusion response," in Proceedings of the 2000 IEEE Workshop on Information Assurance and Security, 2000, pp. 110–116.

[13] K. Nakajima, Y. Kurata, and H. Takeda, "A web-based incident reporting system and multidisciplinary collaborative projects for patient safety in a Japanese hospital," Journal of Quality and Safety in Health Care, vol. 14, p. 123, 2005.

[14] S. Natarajan, A. Harvey, H. Lee, V. Rawat, and L. Pereira, "Technique for providing automatic event notification of changing network conditions to network elements in an adaptive, feedback-based data network," Google Patents, 2003.

[15] T. Fahringer, M. Gerndt, et al., "Knowledge specification for automatic performance analysis, APART Technical Report," vol. FZJ-ZAM-IB-2001-08, FORSCHUNGSZENTRUMJÜLICH GmbH, Zentralinstitut für Angewandte Mathematik, 2001.

[16] A. Sahai, V. Machiraju, M. Sayal, A. Van Moorsel, and F. Casati, "Automated SLA monitoring for web services," Management Technologies for E-Commerce and E-Business Applications, pp. 28-41, 2002.

Reverse Commerce

Adding Information System Support for Customer-centric Market Coordination

Robert Neumann, Konstantina Georgieva,
Reiner Dumke
Department of Distributed Systems
Otto-von-Guericke University Magdeburg
39106 Magdeburg, Germany
{robneuma, ina, dumke}@ivs.cs.ovgu.de

Andreas Schmietendorf
Berlin School of Economics and Law
Fachbereich II
10245 Berlin, Germany
andreas.schmietendorf@hwr-berlin.de

Abstract—**While it is sometimes hard for individuals to find all suppliers in a market in order to be able to determine the one with the lowest price for a product or service, they pay often more than necessary. There might have been another cheaper supplier, which –due to the opaqueness of the market- they have not found. Furthermore, if suppliers for a product or service are scarce, individuals tend to accept lower service quality, because they are happy to have found after all the item of their desire. If consumers were able to simply announce their demand on a global level where implicitly all suppliers were integrated, they could leave the process of investigating the market and negotiating prices completely to the suppliers. In this article, we are going to provide a formal model that describes a consumer-centric approach for market coordination and leverages the support of information systems.**

Keywords-Electronic Commerce; market coordination; efficiency; electronic markets; fixed pricing; dynamic pricing.

I. Introduction

In the Internet era, electronic commerce (e-commerce) plays an important role for both consumers as well as suppliers. E-commerce enables individuals to procure goods and services from their computers at home. Suppliers on the other hand benefit from a distribution channel that is cheap and very close to the customer. The number of internet users has increased considerably since the mid-1990s. At the same time, the revenue that is generated by internet-based e-commerce has increased steadily and keeps increasing [1].

While the development of e-commerce and its technological change are remarkable, the fundamental way of how e-commerce is conducted has not changed at all. Generally spoken, if consumers use the e-commerce channel for fulfilling their demand, they browse through a number of websites of suppliers that offer the product or service and decide for one. Though tools, such as product search engines, kept improving thus providing even richer experiences to the consumer, our contemporary understanding of e-commerce is still very supplier-biased.

Especially, the information phase of an e-commerce transaction is marked by a high level of pro-activity on the consumer side [2]. After consumers have realized that they have a demand for something, they need to investigate the market, browse through websites of potential suppliers, negotiate the price and finally decide on a supplier they want to bargain with. The suppliers on the other hand want to keep the market just transparent enough to still being visible to the consumer, but too opaque for the consumer to be able to find other, maybe even cheaper suppliers that offer better service.

The outcomes of such market constellation are manifold, though we tend to not often recall them to our mind. While it is sometimes difficult for us to find all suppliers in a market in order to be able to determine the best offer, we pay more than necessary as there might have been another cheaper supplier, which we have not found. Furthermore, if suppliers for a product or service are scarce, we tend to accept lower service quality as we are happy to have found the item of our desire at all.

If consumers were able to simply announce their demand on a global level where implicitly all suppliers were integrated, they could leave the process of investigating the market and negotiating prices completely to the suppliers. This way, they would avoid spending time on drilling through the opaqueness of markets and at the same time receive a number of bids from all suppliers that could potentially deliver the desired product or service. Their task would simply be to choose the supplier that offers the lowest bid. Thereby, consumers in such kind of markets would benefit from lower prices and eventually even better service.

Throughout this article, we will refer to the traditional way of e-commerce conduct as “Forward Commerce”. Contrarily, we will introduce a model that emphasizes customer-centric market coordination and refer to it as “Reverse Commerce”. We will provide a conceptual framework that underlines the essence of Reverse Commerce. Furthermore, we will emphasize the role of information systems and show how they build the technological backbone of Reverse Commerce. Finally, we will detail economical implications and discuss the potential of Reverse Commerce to positively affect both the overall service quality as well as the total welfare of an economy.

II. Related Work

The fundamental idea behind Reverse Commerce is not new. Dynamic price making as well as price making through bidding are mechanisms that are well known from virtual or real marketplaces that support auctioning. Reverse Pricing and Reverse Auctioning [3] are concepts that can be found on very specific and mostly B2B-only trading platforms, but that are not yet generally accepted among end-consumers.

While reversed market coordination allows for much higher dynamicity and fluctuation of prices and quantities traded, it also seems to be capable of much more accurately capturing and exhibiting short-term perceptions, moods and feelings of market participants. Fixed pricing [4], contrarily, due to its very nature aims at flattening out the fluctuations that exist in dynamic pricing, thereby making the market more reliable, stable and anticipatable.

As we will discuss later, however, reliability and stability in fixed pricing are traded for welfare. The reason for that can be comprehended by taking a look at the border cases that exist in fixed pricing. Though consumers and suppliers exist that would agree on a bargain below the fixed price, there is no way for both to communicate this to each other. Consequently, certain consumers cannot fulfill their demands while certain suppliers cannot make profit. The overall economy misses out on welfare.

Even though dynamic price making promises certain benefits for both consumers as well as suppliers (compared to fixed pricing it is “Pareto optimal” [5]), end-consumer markets are still dominated by fixed price systems. Sufficient examples exist, however, that dynamic price making works fine in end-consumer markets (e.g. eBay, stock markets) bringing up the question of why it has not yet gained wider acceptance.

Throughout the remainder of this article, we will combine the two concepts dynamic pricing with RFQ (request for quotation) [6]. We will emphasize how new technology can help with expanding the field of application of those concepts into end-consumer markets, thereby referring to the overall effort as Reverse Commerce.

III. Market Coordination

In this article, we propose Reverse Commerce as an alternative to the traditional way of business conduct and refer to the status quo as “Forward Commerce”. Based on [7], we define a market M as:

$$M = (\Sigma \cup \Gamma \cup \Delta, R_M) \tag{1}$$

Thereby, $\Sigma = \{\sigma_1, \ldots, \sigma_n\}$ is the set of involved suppliers, $\Gamma = \{\gamma_1, \ldots, \gamma_m\}$ is the set of consumers, $\Delta = \{\delta_1, \ldots, \delta_k\}$ is the set of available items in this context and R_M summarizes the relation between all participants in the market above.

Furthermore, an electronic market M^{eB} can be defined as:

$$M^{eB} = (\Sigma^{eB} \cup \Gamma^{eB} \cup \Delta^{eB}, R_M{}^{eB}), \tag{2}$$

whereby $\Sigma^{eB} \subseteq \Sigma$, $\Gamma^{eB} \subseteq \Gamma$, $\Delta^{eB} \subseteq \Delta$ and $R_M{}^{eB} \subseteq R_M$. Note that some σ_i are not part of the electronic market M^{eB}. Otherwise, this could also infer that some $\sigma_i{}^{eB}$ are not part of Σ. In this case, $\Sigma^{eB} \supset \Sigma$, which is a phenomenon that corresponds well with recent trends, but will not be considered throughout this article. Though we are aware of differences existing between M and M^{eB}, throughout the remainder of this article we are going to neglect them and will use M^{eB} as synonym for M, Σ^{eB} as synonym for Σ, Δ^{eB} as synonym for Δ and $\sigma_i{}^{eB}$ as synonym for σ_i.

In order to describe certain aspects of M, we need to refine our perspective on Σ, Γ and Δ and provide more detail. Therefore, we define σ = {ident$_\sigma$, presentation$_\sigma$, reputation$_\sigma$, {δ_σ}}, δ = {ident$_\delta$, price$_\delta$} and γ = {ident$_\gamma$, motivation$_\gamma$, {δ_γ}}.

IV. Forward Commerce

The contemporary understanding of how consumers find the product they are looking for and the supplier they want to bargain with has grown with the evolution of the human being as a very static and manifested pattern. If an individual is searching for a new item it wants to buy, this pattern is best described through the following process [2]:

1) Individual experiences need for an item δ_i
2) Individual starts investigating the price (price$_\delta$*)* of the item by visiting various suppliers (Σ)
3) Individual decides for the supplier σ_i that it personally likes best (influenced by price, reputation, etc.)
4) Individual enters bargain with supplier σ_i and obtains the item in exchange for a liability or funds

The above process is marked by a very strong proactivity of the consumer towards the bargain. What that means is that regarding the supply side, all the supplier usually needs to do is offering and advertizing the item that s/he wants to sell. The rest of the engagement is covered by actions of the consumer. The consumer investigates on products and suppliers (in-store, on the internet or in magazines). The consumer contacts the supplier and/or visits shops. Eventually, the consumer negotiates the price with the supplier, since it is in the immediate interest of the consumer to yield a lower price.

In essence, in Forward Commerce the consumer γ, when searching for a product δ_γ, is influenced by the supplier’s product presentation and his reputation. The three components δ_γ, presentation$_\sigma$ and price$_\delta$ get projected onto a supplier σ in the market who can deliver the product. Finally, the supplier σ -depending on the price of the product and the consumer’s motivation- decides, whether to sell δ to the consumer or not.

$$\begin{gathered} \gamma \Rightarrow (\delta_{\gamma,}\, \text{presentation}_{\sigma,}\, \text{reputation}_{\sigma}) \Rightarrow \sigma \\ \sigma \Rightarrow (\delta_{\sigma}, \text{ motivation}_{\gamma,}\ \text{price}_{\delta}) \Rightarrow \delta \end{gathered} \tag{3}$$

V. Reverse Commerce

The concept of Reverse Commerce proposes that in order to find the item δ_γ as well as the supplier σ the consumer wants to bargain with, all the individual needs to do is announce or "advertise" its interest in the item. Suppliers that are able to deliver the requested item receive a note about a new bargaining opportunity. They apply at the consumer for a bargain by providing a quote at which they are willing to sell the item. The consumer decides from the list of suppliers that have applied for a bargain for a favorite supplier. The payment is completed and the item delivered. The Reverse Commerce process is listed below [2]:

1) Individual experiences need for an item δ_i
2) Individual posts a request the item δ_i
3) Suppliers Σ apply for a bargain with the consumer γ by providing a quote
4) Individual enters bargain with chosen supplier σ and obtains the item in exchange for a liability or funds

From a consumer's point of view, the concept of Reverse Commerce would allow for a totally new consumer experience. Individuals who wanted to buy an item do not need to spend time on investigating items and suppliers anymore, but would rather simply announce that they have a demand that is to be fulfilled. The remaining actions would lie on the side of those suppliers that could potentially fulfill the demand.

In contrast to Forward Commerce, in Reverse Commerce the consumer solely decides for a product while in his decision making he leaves the supplier completely out. The decision for a product δ_γ is driven by the consumer's motivation to buy the product (motivation$_\gamma$) as well as the product's price (price$_\delta$). The three components δ_γ, motivation$_\gamma$ and price$_\delta$ get projected onto the product δ. Afterwards, the consumer gets automatically assigned the best feasible supplier influenced by his presentation and his reputation.

$$\gamma \Longrightarrow (\delta_\gamma, \text{motivation}_{\gamma,}\ \text{price}_\delta) \Longrightarrow \delta$$
$$\delta \Longrightarrow (\delta_\sigma, \text{presentation}_{\sigma,}\ \text{reputation}_\sigma) \Longrightarrow \sigma \qquad (4)$$

VI. Opaqueness Margins

The key benefit of Reverse Commerce is that consumers do not need to understand the market behind the items they are interested in anymore in order to be able to find the best offer. While suppliers are competing against each other, consumers benefit from this competition by picking the supplier that offers the best deal. They leave the price making and negotiation as it occurs in Forward Commerce to the competing suppliers.

Contrary to Forward Commerce, Reverse Commerce implicitly ensures that the consumer finds the supplier that offers the best price. This is because in Reverse Commerce only the supplier that offers the best deal out of all suppliers that exist in the market will enter the bargain with the consumer. This is not to be taken for granted as normally in Forward Commerce the cost for the consumer to fully explore the market and reveal all suppliers within is very high. As in Reverse Commerce the consumer more or less "pulls" the offers out of the market, all suppliers that could potentially fulfill a demand will implicitly provide the best offer as their bid competes with the bid of other competitors.

In Forward Commerce, the responsibility of finding the best deal lies with the consumer. Depending on how well s/he investigates the market, the consumer will find the supplier that offers the lowest bid. Thereby, it is very likely that the consumer -no matter how intensively s/he investigates the market- will not find the lowest bid. In many cases, the market opaqueness is simply too high as there might be a huge number of suppliers that are spread across a large geographic area and not easy to locate. The opaqueness of the market will finally compel the consumer to restrict his search and employ an individual search heuristic. The consumer will choose a supplier that in accordance with the applied heuristic offers the lowest bid. In the average case the consumer chooses the average bid as depicted in Figure 1.

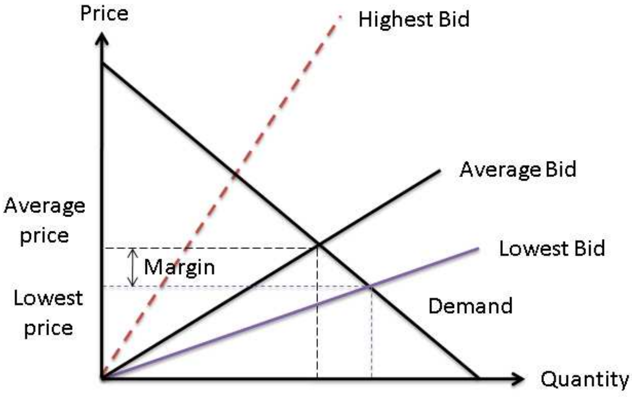


Figure 1. Best heuristic bid in Forward Commerce.

While the search heuristic allows the consumer to yield relatively good results (the average case) when dealing with the opaqueness of the market, the consumer will likely not find the best possible bid. In this case, the consumer will pay a price for the item that is by a "margin" higher than the lowest bid available in the overall market (see Figure 1). We will from now on refer to this margin as opaqueness margin. The opaqueness margin is the margin the supplier makes with the consumer not having found the lowest bid available in the market. Thereby, it is important to note that the supplier can only realize the opaqueness margin because the market is not completely transparent to the consumer. On the other side, the opaqueness of markets equips suppliers with a certain advantage in terms of bargaining power. Baileys and Bakos have referred in their fundamental work on electronic markets to this phenomenon as "bargaining asymmetry" [8]. While markets aim at aggregating buyer demand in order to achieve economies of scale to reduce supplier-biased bargaining asymmetry, markets that function in line with traditional Forward Commerce will hardly be able to completely abandon the opaqueness margin and thus remove the supplier-biased bargaining asymmetry.

VII. Conceptual Basics of Reverse Commerce

In the following, we are going to provide a theoretical framework that outlines the conceptual difference between Forward and Reverse Commerce and points out the significance of the concept that underlies the opaqueness margin.

Let δ be the item an individual is looking for in order to fulfill a demand. Let μ be the determinator function that an individual employs when choosing a supplier to bargain with from a set of multiple suppliers. Let furthermore φ be the search heuristic an individual follows when searching for suppliers in a market which is not completely transparent and σ be the supplier the individual finally decides to bargain with. Let Σ^{δ} be the wholeness of suppliers that exist in an opaque market and that could potentially deliver the desired item δ. Let Σ^{δ}_{NC} be the wholeness of suppliers that exist within an opaque market and that due to the heuristic φ that was employed by the consumer to search the market, were not found and could not be considered. Finally, let OM be the opaqueness margin.

Reverse Commerce variables:

δ: desired item
Σ^{δ}: suppliers who can deliver δ
$\Sigma^{\delta} = \bigcup_{i=1}^{n} \Sigma^{\delta}{}_{i}$
Σ^{δ}_{NC}: suppliers not considered by consumer
φ: individual search heuristic
μ: determinator function that chooses supplier to bargain with
σ:chosen supplier
OM: opaqueness margin

with:

$$\varphi\colon \gamma \Longrightarrow \left(\delta_{\gamma}, \text{motivation}_{\gamma,}\ \text{price}_{\delta}\right) \Longrightarrow \delta$$
$$\mu\colon \delta \Longrightarrow (\delta_{\sigma}, \text{presentation}_{\sigma,} \text{reputation}_{\sigma}) \ \square\ \sigma$$

In traditional Forward Commerce, the supplier σ, which the individual determines, can be expressed as a function μ of Σ^{δ}, the individual search heuristic φ and the desired item δ. Thereby, Σ^{δ} represents the suppliers which exist in the opaque market (M) and which can deliver δ.

$$\sigma = \mu(\Sigma^{\delta}, \varphi, \delta) = \mu\left(\varphi(\Sigma^{\delta},\delta)\right) = \mu(\varphi(\Sigma^{\delta})) \tag{5}$$

The number of suppliers the consumer found in the opaque market after having employed the search heuristic φ can be expressed as $|\varphi(\Sigma^{\delta})|$. Accordingly, the total number of suppliers that exist in the opaque market and that could potentially deliver the item δ can be expressed as $|\Sigma^{\delta}|$. If the quotient of both the number of suppliers the consumer found and the total number of suppliers in the market is smaller than 1, this implies that the consumer –due to the heuristic search- did not find all suppliers who could potentially deliver the desired item δ.

$$\frac{|\varphi(\Sigma^{\delta})|}{|\Sigma^{\delta}|} < 1 \rightarrow \text{not all suppliers were found} \tag{6}$$

The set of suppliers that –due to the heuristic search- were not considered by the consumer for a bargain is represented by Σ^{δ}_{NC} . Thereby, Σ^{δ}_{NC} is simply the quantitative difference between the total of suppliers in the market and the set of suppliers the consumer found after having applied φ to Σ^{δ}.

$$\Sigma^{\delta}_{NC} = \Sigma^{\delta} \setminus \varphi(\Sigma^{\delta}) \tag{7}$$

If there is a supplier in Σ^{δ}_{NC}, which the consumer does not know about, but which, if hypothetically chosen by the consumer, offered a better bargain than the best supplier within the set of suppliers the individual had to choose from after the heuristic was applied (φ(Σδ)), then the opaqueness margin OM is larger than zero. In this case, the chosen supplier σ made a margin with the market being not completely transparent.

$$\mu(\Sigma^{\delta}_{NC}) \succ \mu\left(\varphi(\Sigma^{\delta})\right) \rightarrow \text{OM} > 0 \tag{8}$$

In case of Reverse Commerce, above argumentation would look fundamentally different. One difference in Reverse Commerce lies in σ_{RC} being chosen from the total of suppliers in the Reverse Commerce market. The individual heuristic in Reverse Commerce (φ_{RC}) implicitly yields the wholeness of suppliers to the consumer. In Reverse Commerce, φ_{RC} makes all suppliers in the market visible to the consumer as all suppliers that would potentially want to bargain with the consumer themselves receive a notification from the market that a new bargaining chance exists. Suppliers automatically start providing the consumer with offers while readjusting their bids when competing with other suppliers' bids. The consumer merely needs to pick the lowest bid.

$$\varphi_{RC}(\Sigma^{\delta}) = \Sigma^{\delta} \tag{9}$$

Consequentially, the supplier which the consumer chooses from the set of suppliers it gained after having employed the heuristic $\varphi_{RC}(\Sigma^{\delta})$ is determined by $\mu\left(\varphi_{RC}(\Sigma^{\delta})\right)$ or in other words $\mu(\Sigma^{\delta})$. Thereby, the best supplier, which the consumer chooses from $\varphi_{RC}(\Sigma^{\delta})$, is also the best supplier which the consumer would hypothetically choose from Σ^{δ}, if the market was completely transparent to him. This results into the opaqueness margin OM being zero.

$$\sigma_{RC} = \mu\left(\varphi_{RC}(\Sigma^{\delta})\right) = \mu(\Sigma^{\delta}) \rightarrow OM = 0 \tag{10}$$

Finally, the quotient of $|\varphi_{RC}(\Sigma^{\delta})|$ and $|\Sigma^{\delta}|$ equals one, as no supplier exists that could not be found and taken into consideration by the consumer.

$$\frac{|\varphi_{RC}(\Sigma^{\delta})|}{|\Sigma^{\delta}|} = 1 \tag{11}$$

The consumer has found the supplier that offered the best deal without actively searching for it.

VIII. Information System Support

While the above conceptual framework has outlined that customer-centric market coordination as it occurs in Reverse Commerce eliminates opaqueness margins, we are now going to detail how information systems can support large scale Reverse Commerce scenarios.

One essential requirement for Reverse Commerce to function properly is that both consumers and suppliers are given a medium via which they can communicate. In case of Forward Commerce, consumers are utilizing search engines in order to find the supplier they want to bargain with. In many cases, however, the search results are fuzzy, as neither the consumer was able to exactly describe what s/he was looking for, nor did all the suppliers appear in the search results. Even though price search engines specialized on increasing the response quality of queries that were placed by consumers, the accuracy of the returned results is still not very high. While easy queries already yield good results, more complicated queries that contain several restrictions on the product the consumer is looking for still fail in returning precise offers. With respect to an ongoing automation of matching consumer demand with bids posted by suppliers in Reverse Commerce, it is essential to increase the accuracy of query results.

In Forward Commerce, we do not see that the customer is provided with a “simplified and pleasing interaction” [9] that allows for an instantaneous gathering of product information. We believe that this is because of the shortcomings of the general approach of matching queries posted on web search engines with offers posted by suppliers, maybe even on their private website. No matter how smart the algorithms that search engines employ, they will always try to match poorly structured queries with poorly structured offers placed on the worldwide web. In order for Reverse Commerce to function properly, this concept needs to be able to rely upon high query accuracy.

In order to underline our argumentation above, let us assume the following theoretical example. Given a query Q that consists of n components $c_1..c_n$:

$$Q = \{c_1, c_2, c_3, \ldots, c_n\} \tag{12}$$

This query is provided (most likely as a string) to a search engine E that returns a result R consisting of the components $c_1..c_m$.

$$R = \{c_1, c_2, c_4, \ldots, c_m\} \tag{13}$$

Thereby, the search engine E projects the query Q onto the result R.

$$E: Q \rightarrow R \tag{14}$$

Intentionally, the user of the search engine would expect Q to be equal to R, as he expects the search engine to return accurate results. In this case, subtracting R from Q would yield an empty set $\emptyset$. Furthermore, m would equal n.

$$Q \setminus R = \emptyset \tag{15}$$

In most cases, however, due to search engines operating upon both weakly structured queries as well as weakly structured product descriptions, Q and R will be different.

$$Q \setminus R \neq \emptyset \tag{16}$$

This could indicate that either R does not contain all components of Q or it contains more components than Q as m would be different from n. The reason for the distortion that exists between Q and R is the poor structure of the query as well as the product description. Our approach towards a solution to the above problem includes strongly structured queries on the consumer side as well as strongly structured product descriptions on the supply side. Our approach aims at ensuring that $Q \setminus R$ always yields $\emptyset$.

In an earlier work, we have talked about Organic Product Catalogs (OPCs) [10] that aim at providing suppliers with a centralized, cloud-based and low-cost platform on the internet to describe and maintain their products and expose them to the worldwide web. We have designed the architecture of our OPC as in Figure 2.

Figure 2 provides two perspectives on a cloud-hosted OPC. In the horizontal perspective, based on the cloud platform, the tenant layer ensures that each participant has access to an own customized content area. The actual EPC distinguishes and individually manages the product descriptions for each client separately and thus provides full support for multitenancy. Furthermore, the OPC provides a general purpose product description language (PDL) that can be used to describe the products and services of each tenant.

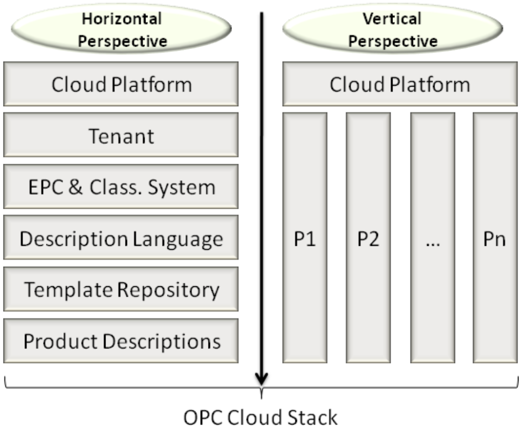


Figure 2. Cloud stack of an OPC.

In most cases, however, tenants will use templates that were created by other tenants earlier for describing their products [10] [11]. Finally, the bottom-most layer of our OPC centrally stores all product descriptions that a tenant is maintaining. The vertical perspective (or the per-tenant perspective) visualizes that on each layer in the horizontal perspective, tenants are provided with their own customized (multitenant) experience.

The general purpose product description language serves as basis for the definition of products and services of any kind. Thereby, it is important to note that the PDL aims at being universal. What that means is that the PDL is capable of modeling any product or service, no matter how trivial or complicated it is. One fundamental assumption we make in order to correspond with the universality requirement says that every product or service can be expressed as a list of hierarchically ordered attributes.

While our PDL can be considered as an XML dialect, it adds structure and with that machine readability to the product description. In accordance with our organic product catalog approach, we want to achieve the same structured machine readability regarding the queries that are posted by the consumer. The only way to achieve this is to equip the consumer with a possibility to express his demand in a standardized way. Regardless of how we will obtain the standardized query from the consumer (in the most trivial case s/he will write XML), once we have transformed the query into a machine-readable format, we will be able to run it on our organic product catalog.

The accuracy of the query result depends on the level of detail the consumer has provided in his query description. If a one hundred percent matching of the query with existing product descriptions is possible, the consumer will get immediate response from the system and with that an immediate quote. If the matching is only partially successful, because the consumer specified characteristics in his product that no product available in the product catalog has, suppliers would receive a notification from the system that they could potentially fulfill a demand, but need to customize their offers. In addition to that, suppliers could update their quotes, if they were outbid by competitors who posted lower bids. Of course, the outbid competitors could again update their bids, so that eventually the price of the final quote would be determined dynamically. In any case, the consumer would receive the lowest bid available in the market.

IX. TOWARDS A REVERSE COMMERCE PROCESS MODEL

As we have indicated earlier, in order to function properly, structured queries are necessary. Figure 3 features a process flow for how Reverse Commerce transactions could conceptually look.

The consumer begins with a web-based keyword search in order to determine the type of product s/he is looking for (1). S/he enters the search string similar to what is known from web-based search engines and submits it to the server. (2) If templates for the query are available, the consumer will proceed with customizing the query template while adding query components (1..n) to his query (Q). In addition to substituting template placeholders with concrete values, new components, which are not part of the original template, can also be added. Once the query was constructed, it is submitted in (3) and applied to the organic product catalog (OPC). As there can be multiple products matching the same query, the OPC returns a list of found products -the result set (R)- to the consumer (4). After the consumer has chosen his favored product (5), suppliers that could potentially deliver the product start placing bids on the price (6), either manually or automatically. Finally, once the consumer believes s/he has received sufficient quotes, s/he can decide for a supplier and the process concludes (7).

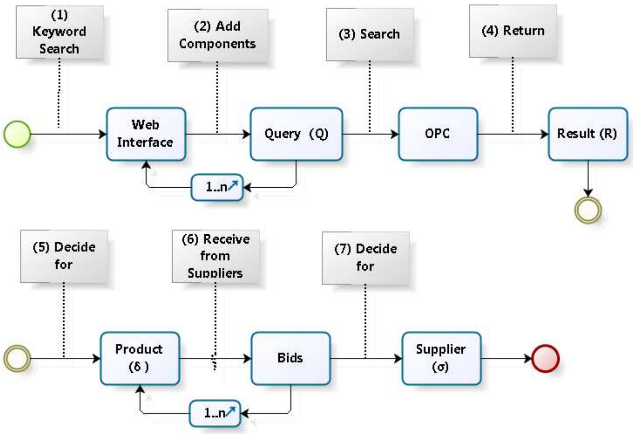


Figure 3. Process model for designing structured queries based on an OPC.

X. ECONOMIC IMPLICATIONS OF REVERSE COMMERCE

Since in Reverse Commerce, the opaqueness margin is implicitly abandoned while ensuring that the consumer receives the lowest bid without any meaningful action on his side, this could have significant economic consequences. We are now going to show how Reverse Commerce-driven markets could eventually even impact the welfare of a society.

In order to understand the concept of welfare, we will quickly review the match making process as it occurs in markets. According to Figure 4, the point where supply meets demand is represented by the tuple E = (market clearing quantity, market clearing price), whereby E is the market clearing equilibrium [12]. Figure 4 furthermore depicts that, if the supply tilts to the right, the supply hits the demand at a lower angle, thus resulting into a lower market clearing price and a larger market clearing quantity.

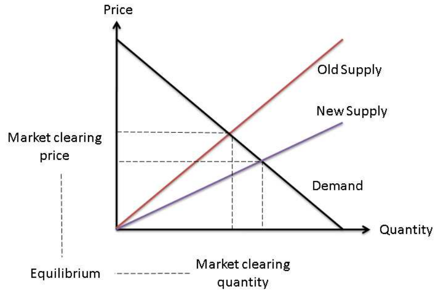


Figure 4. Market Coordination [12] [13].

According to [13], welfare is defined as the aggregation of the utility of single individuals or groups. Graphically seen, welfare can be expressed as the triangle between the demand and the supply function (see Figure 5).

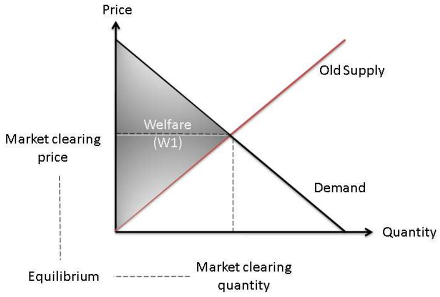


Figure 5. Graphical representation of welfare [13].

In accordance with Figure 4, by lowering the prices and thus tilting the supply function to the right, the area of the triangle between demand and supply increases. At the same time, a larger triangle means a higher welfare (see Figure 6).

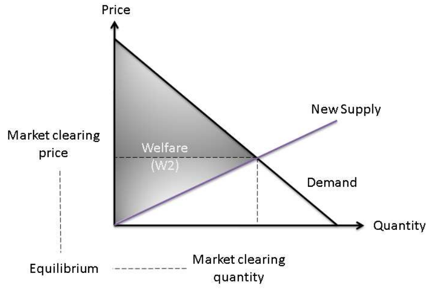


Figure 6. Increasing welfare by reducing prices [13].

Since we earlier stated that Reverse Commerce implicitly yields the lowest bid for the consumer, the Reverse Commerce demand function meets the supply function with the maximum possible right tilt. As the area of the triangle that determines the welfare exclusively depends on the tilt angle, the smaller the inclination of the supply, the higher is the welfare. While in Forward Commerce the consumer's demand function could as well meet steeper supply functions, the welfare is not implicitly maxed out, as it is the case in Reverse Commerce.

It is important to note that the welfare argumentation above is not exclusively applicable to Reverse Commerce. It is a phenomenon of electronic markets in general. While electronic markets typically show lower transaction costs of business transactions as compared to traditional markets [14], the consumers benefit from lower prices. This, however, does still not imply that consumers always find the bid with the minimum possible price. Solely the concepts and structures of Reverse Commerce ensure that consumers always find the lowest possible bid, which means that their demand meets the supply function with the flattest slope.

XI. INTRINSIC SERVICE QUALITY

At this point, we would like to discuss another economic implication of Reverse Commerce. With Intrinsic Service Quality (ISQ) we refer to the service quality which customers experience when interacting with vendors or suppliers. Even though there is no universal definition of quality, the general contemporary understanding is that quality means "meeting or exceeding customer expectations" [15]. The ISQ describes our attempt towards the definition of a concept that measures the quality of service which customers experience when interacting with vendors or suppliers.

In Forward Commerce, due to the opaqueness margins, which suppliers make with their customers, there is no immediate need for suppliers to guarantee a high level of service quality to their customers. This is because in Forward Commerce suppliers are often not chosen by consumers for a bargain because they offer the best service, but because they allegedly offer the lowest bid. In Reverse Commerce, however, the consumer implicitly assumes that every supplier who wants to bargain with the consumer already offers the lowest bid possible.

As we have shown earlier, in Reverse Commerce prices would eventually converge against the lowest bid available in the market as no consumer would voluntarily pay more than necessary. The only way for suppliers to distinguish their offers from competitors in Reverse Commerce is by excelling in the way they deal with their customers. By offering an intrinsic service quality to the consumer that is better than the competitors' ISQ, suppliers will convince consumers to enter a bargain with them. Figure 7 visualizes the conceptual difference of the ISQ in Forward Commerce and Reverse Commerce respectively.

As in fully supplier-biased markets consumers will encounter severe problems with finding alternative suppliers (e.g., monopoly), the bargaining power for products or services is more or less to a 100% with the supplier. Accordingly, the supplier can decide about prices and with this about the opaqueness margin OM s/he wants to make. Furthermore, there exists no incentive for the supplier to offer a high service quality (ISQ → 0%) as from the consumer's point of view he is the only supplier to deliver the item.

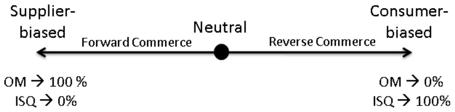


Figure 7. ISQ in Forward Commerce vs. Reverse Commerce.

In a fully consumer-biased market as it can appear in Reverse Commerce, the market would be totally transparent to the consumer. The consumer would implicitly know about all alternative suppliers in the market as those would "apply" at him for a bargain with their bid. As those suppliers that are not able to provide bids on the same price level as the cheapest bid in the long run would vanish from the market, there would be no opaqueness margin to make any longer (OM → 0%). On the contrary, if all suppliers offer the same low bid, the only way of distinguishing themselves from competitors is by offering a better service quality (ISQ → 100%).

XII. Conclusion and Future Work

In this article, we have introduced the concept of Reverse Commerce and have opposed it to the traditional way of e-commerce market coordination. We have described in a formal model how Reverse Commerce could support customer-centric market coordination by reducing prices on products and services while at the same time improving service quality. We have furthermore detailed that Reverse Commerce can only function properly, if it is supported by information systems.

For this reason, our team has already started developing a prototype of the first Reverse Commerce-enabled electronic marketplace. Due to our Reverse Commerce marketplace eventually becoming subject to a comparatively high load, we have decided to leverage the power of the Cloud Computing [16] paradigm and implement the marketplace as a cloud-based application. Concerning the cloud platform, we have decided for Microsoft's Windows Azure, since the available tools as well as the infrastructure seem to be highly effective.

Once we put our first functioning prototype of a Reverse Commerce marketplace online, we hope to be able to soon gather data on consumer behavior and the acceptance by consumers as well as suppliers. If Reverse Commerce turned out to be the new way of e-commerce conduct, consumers could soon relax and lay back watching how the prices on their requests keep dropping with every new bid that was posted by a supplier.

References

[1] Fritz, W. 2004. Internet-Marketing und Electronic Commerce. Gabler Verlag. Wiesbaden, Germany.

[2] Neumann, R. 2010. The EBF Application Foundation: An Approach towards the Design of an E-Commerce Framework for Small and Medium Enterprises. Lambert Academic Press. Saarbrücken, Germany.

[3] Beall, S., Carter, C., Carter, P., Germer, T., Hendrick, T., Jap, S., Kaufmann, L., Maciejewski, D., Monczka, R., and Peterson, K. 2003. The Role of Reverse Auctions in Strategic Sourcing. CAPS Research Report. CAPS Research. Tempe, USA.

[4] Clark, W. 2010. Fixed price contracts = more risks for contractors. Business Research Services, Inc. Hampton, USA.

[5] Brickley, J., Zimmerman, J. L., and Smith, C. W. 2008. Managerial Economics and Organizational Architecture. McGraw Hill. New York, USA.

[6] Phillips, J. 2008. Procurement Management in Project Management - Taking Out a Contract. PM Hut. [Online]. Available: http://www.pmhut.com/procurement-management-in-project-management-taking-out-a-contract [Accessed: Dec. 30, 2010].

[7] Skyttner, L. 2005. General Systems Theory – Problems, Perspectives, Practice. World Scientific Publ. New Jersey, USA.

[8] Bailey, J. P. and Bakos, Y.1997. An exploratory study of the emerging role of electronic intermediaries. International Journal of Electronic Commerce Volume 1 Number 3, Armonk, USA, pp. 7-20.

[9] Rajput, W. E. 2000. E-Commerce Systems Architecture and Applications. Artec House. Boston, USA.

[10] Neumann, R., Schmietendorf, A., and Dumke, R. 2010. Cloud-based Organic Product Catalogs – A Highly Pervasive E-Business Approach for Micro Enterprises. Proceedings of the 2010 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government, Las Vegas, USA, pp. 123-129.

[11] Neumann, R., Schmietendorf, A., and Dumke, R. 2010. Organic Product Catalogs - Towards an Architecture for Cloud-based Micro Enterprise E-Commerce. IEEE Cloud 2010 - The 3rd International Conference on Cloud Computing, Miami, USA, pp. 530-531.

[12] Samuelson, P. and Nordhaus, W. 2009. Economics. McGraw-Hill/Irwin. New York, USA.

[13] Varian, H. R. 2009. Intermediate Microeconomics: A Modern Approach. W. W. Norton & Company. New York City, USA.

[14] Bichler, M. 2001. The Future of e-Markets: Multidimensional Market Mechanisms. Cambridge University Press. Cambridge, UK.

[15] Sower, V. E. 2010. Essentials of Quality with Cases and Experiential Exercises. Wiley. Hoboken, USA.

[16] Armbrust, M., Fox, A., Griffith, R., Joseph, A., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I., and Zaharia, M. 2009. Above the clouds: A Berkley view of cloud computing. University of California. Berkley, USA. [Online]. Available: http://www.google.de/#hl=en&source=hp&q=Above+the+clouds%3A+A+Berkley+view+of+cloud+computing.+&rlz=1R2SKPT_deDE409&aq=f&aqi=&aql=&oq=&gs_rfai=&fp=22b1b139f61051cd [Accessed: Dec. 30, 2010].

Moving E-Commerce Towards E-Commodity

A Consequence of Cloud Computing

Robert Neumann, Konstantina Georgieva,
Reiner Dumke
Department of Distributed Systems
Otto-von-Guericke University Magdeburg
39106 Magdeburg, Germany
{robneuma, ina, dumke}@ivs.cs.ovgu.de

Andreas Schmietendorf
Berlin School of Economics and Law
Fachbereich II
10245 Berlin, Germany
andreas.schmietendorf@hwr-berlin.de

***Abstract*—Many people believe that nowadays everything can be purchased online, but the reality could not be further from the truth. A recent survey that we have conducted in two medium-sized German cities revealed that only about 12% of all businesses are able to accept online orders. While 100% of all large enterprises run their own online store, only 6.9% of all micro, small and medium enterprises are able to receive online orders. The obvious outcome of such an e-commerce disparity is that consumers can only very rarely place online orders at their favorite local stores. In this article, we will describe the idea of providing electronic commerce as a commodity service to micro, small and medium enterprises. We will identify business-side as well as technology-side problems that exist in current e-commerce and discuss strategies to mediate them.**

Keywords-E-commerce; e-commodity; micro small and medium enterprises; cloud computing; product definition; agile e-commerce.

I. INTRODUCTION

In the Internet era, electronic commerce (e-commerce) plays an important role for both consumers as well as suppliers. E-commerce enables individuals to purchase goods from their computers at home and get them delivered to their address. Suppliers on the other hand benefit from a distribution channel that is cheap and very close to the customer.

The number of internet users has increased considerably since the mid-1990s. At the same time, the revenue that is generated by internet-based e-commerce has increased steadily and keeps increasing [1]. While this development is remarkable, contemporary e-commerce is limited to goods that are easy to describe, standardized and that have low asset specificity [2]. Adler adds that products that qualify best for e-commerce are so called “search products”, which can be purchased based only on search qualities [3]. Products and services that do not meet above requirements or are no search products and are not frequently or not at all traded on the internet.

Especially, so called micro, small and medium enterprises (MSMEs) -that are enterprises with less than 250 employees [4]- do not very often draw benefit from the development that has taken place in e-commerce. Reasons for that can be found in MSMEs (especially micro enterprises) often dealing with goods and services that are considered as being difficult to trade on the internet as well as in the complexity and cost related to the maintenance of e-commerce systems.

With 99% of all European businesses being MSMEs [4] and the vast majority of them not being able to accept online orders, end-consumers are seldom able to purchase those goods online which they normally buy from local stores around the area they live in. Goods that consumers buy on a daily basis (e.g., groceries, toiletries, etc.) are not yet often to be found on the internet. Appointments with local craftsman, practitioners, barbers or other service providers can still not be booked online. We have referred to a kind of e-commerce where those products and services that individuals buy on a daily basis can be purchased online as day-by-day e-commerce (DBDE-Commerce) [5].

In accordance with our vision of DBDE-Commerce, throughout this article we will provide a new perspective on how to turn e-commerce into commodity (“e-commodity”). We will refer to the concept of making e-commerce available to each and every business as E-Com-E-Com (“e-commerce as e-commodity”). Inspired by the fundamental ideas behind cloud computing to make computational power become utility (similar to electricity out of a socket), we will introduce a prototype for an intelligent cloud platform that aims at providing virtual storefronts to businesses as commodity service.

II. OVERVIEW AND RELATED WORK

The development which has taken place in the field of electronic commerce goes back to the progress of the World Wide Web. With the upcoming of hypertext and hypertext-enabled web browsers it became possible to provide product information online and to conduct purchases as well as sales over the internet. In the period between 1994 and 1999, many enterprises have moved to the internet while they were expecting an increase in sales caused by their online shops [6]. Amongst these enterprises, only a few succeeded in establishing a global platform for international trade with a diverse variety of products and services that can be purchased online. The most prominent examples can be found in the US-based enterprises eBay or Amazon, which at the moment can be considered as leaders in their field of business.

Both the internet as well as internet-based commerce keep growing continuously and the number of people who have access to web-based virtual storefronts is increasing steadily. According to a report of the German federation of information technology, telecommunication and new media (BITKOM), currently two thirds of all German individuals have permanent access to the internet [7]. Furthermore, the E-Commerce Guide, published by a consortium of different members who are engaged in e-commerce, mentions that about 70% of all German individuals with broad band connection are frequently purchasing goods and services online [8]. The business volume of electronic commerce in 1999 summed up to 1.3 billion €. In 2009, it rose up to 17.3 billion € while the 2014 sales forecast for Germany predicts 27.2 billion € worth of online orders [9]. The Online Shopping Survey (OSS) of the GfK, the largest German market research institution, revealed that the most popular products of electronic trade in 2009 were books (14,9 million individuals), clothes (14,7 million individuals), event tickets (12,5 million individuals), music (8,5 million individuals) and hotel reservations (7,4 million individuals) [10].

Given the Online Shopping Survey of the GfK, it appears that today's e-commerce is dominated by easy-to-describe search products [2]. Though we believe that it is more difficult to also expand e-commerce into those product domains that deal with more complicated products and services, we are convinced that it is possible.

III. Features of Today's E-Commerce

Different e-commerce suites have different functionality, but in general all of them provide certain base level features. These features include:

- Product/service information
- Usability
- Customer support
- Order fulfillment and payment
- Logistics
- Marketing
- Trust with security and privacy

When consumers browse websites with the intent to conduct online purchases, it is important to provide them with rich product and service information. Today's online shops have a lot of features for product presentation. Among the most popular ones are high-resolution images, webcasts or video tours for product configuration engines and online showrooms. Pricing information is always up to date and can change on a daily basis.

Primary reasons for leaving a website are due to long load time, poor navigation, inconvenient coloring or bad page design. To improve the duration of stay of a visitor on a website, developers are given tools and guidelines to create their design in a robust and appealing manner. Many storefronts support multiple languages to reach a large number of customers in different countries. Options for searching, filtering and sorting of product information enhance the end-user experience and provide for an easier way of finding the desired item.

Reoccurring product-related questions are addressed in frequently asked questions (FAQs). In addition, hotlines, web forms or live chats are available to customers for getting in contact with the vendor. In case of asynchronous communication, replies are normally guaranteed to be sent within 24 - 48 hours. Weblogs aim at equipping customers with an instrument to obtain information on products while those are still under development or for trouble shooting after the product was purchased.

Advanced online shops have an integrated order and payment system or at least feature the possibility to connect an order and payment system via certain interfaces. Today's online shops track shopping baskets throughout the entire session of the customer and at the end of the online purchase gather logistics and payment information. Several online payment options exist (e.g., invoice, prepayment, debit entry, cash on delivery, credit card, PayPal or Click & Buy).

Many shops provide fast delivery services offered by partnering logistics enterprises. Most frequently mentioned are DHL, UPS, or GLS. This way, it is possible to dispatch orders within 24 hours and provide order tracking services. Furthermore, certain enterprises (e.g., Alternate) offer short message services to confirm the order or to inform the customer about the status of the delivery.

While several aspects of traditional commerce had to be adapted to new and different requirements in online commerce, marketing also had to come up with new selling strategies. Marketing techniques for online commerce include search engine optimization (SEO) to be ranked among the top entries of search engine results. Cross selling aims at recommending similar or complementary products to the customer. Google has developed AdWords to include semantic information in the placement of ad banners, so that users of a website get presented advertisement that corresponds with the semantic context and topic of the site. Several enterprises have started integrating social media, such as twitter or Facebook, in their online marketing. Others provide shopping club memberships (e.g., Limango or brands4friends), which suggest potentially interesting offers or hot deals to members. Word-of-mouth recommendation (e.g., TRND) is a new way of electronic product marketing, where members are given the opportunity to test products and recommend them to their friends. Members can test products, share them with friends and post the results on trnd.com. Another online marketing strategy is based upon enterprise coupon services, where every day vendors provide a new "Deal of the Day" (e.g., DailyDeal, CityDeal or Groupon).

Finally, most online shops have features that aim at establishing a feeling of trust, security and privacy to their customers. To build up trust, online shop providers use seals of quality, such as Trusted Shops or Safer Shopping. Another helpful instrument to establish trust are guest books where customers can leave comments and feedback regarding previous purchases. Additionally, online shop owners use digital security certificates to protect users against data theft or interception.

Together, all of the features above form an all-in-one package to earn user satisfaction and establish a long-lasting customer relationship. Though e-commerce in general and e-commerce technology in particular have made an impressive development over the last two decades, we still do not see e-commerce having become commodity. E-commerce is yet too much driven by industry and there are still too many factors within an organization that determine whether or not to run a virtual storefront. The end-user experiences supported by most online shops might seem acceptable to the majority of online customers, but at the same time this is where the caveat lies. If future e-commerce technology was capable of attracting a more diverse group of online customers while providing a larger number of products and services online, the functionality and features of today's e-commerce would expand even further into our daily shopping experience. Towards the end of this article, we are going to provide a list of novel features which we are going to support with our next generation e-commerce platform "Goliath".

IV. E-COMMERCE AS E-COMMODITY STUDY

In June 2010, we have reviewed 1859 (Table 1) enterprises in two German cities, namely Magdeburg and Freiburg. We have chosen Magdeburg (population 230.000) and Freiburg (220.000) since both have a similarly big population, a similar number of businesses and since we have researchers in both areas who conducted the analysis. We have split the enterprises in both cities into 1757 MSMEs as well as 102 large enterprises (LE). For the classification of MSMEs and LEs, we have employed the European Commission's definition, where MSMEs are considered as enterprises with <250 employees and LEs as enterprises with >=250 employees [4]. Table 1 features the percentages of MSMEs with no own website (38.4%), enterprises with own website but now online shop (53.6%), enterprises with online shops (6.9%) and enterprises that use their website to promote special website deals (1.1%).

TABLE I. NUMBER OF ANALYZED MSMES

Information	*Magdeburg MSMEs*	*Freiburg MSMEs*	*Total MSMEs*
No website	199	475	674
Only website, no online shop	629	312	941
Online shop	52	70	122
Special deals	20	0	20
Total	900	857	1757

Out of the 900 MSMEs reviewed in Magdeburg, 199 (22.1%) have no own website, 629 (69.9%) have an own website but are not able to accept online offers, 20 (2.2%) have a website to inform about special offers and deals (or limited features) and only 52 (5.8%) enterprises run their own electronic storefront (Figure 1).

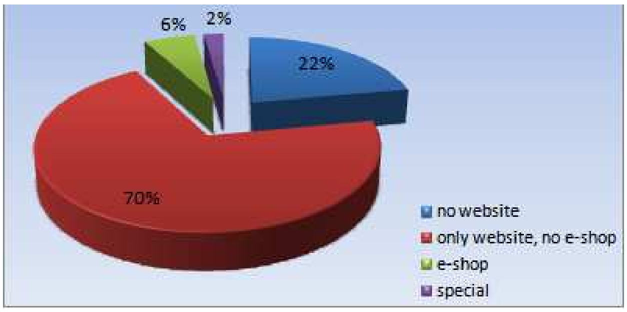


Figure 1. MSME e-commerce in Magdeburg.

In case of the city of Freiburg, out of 857 MSMEs 475 (55.4%) have no own website, 312 (36.4%) have an own website but do not accept online offers and only 70 (8.2%) have an online shop (Figure 2).

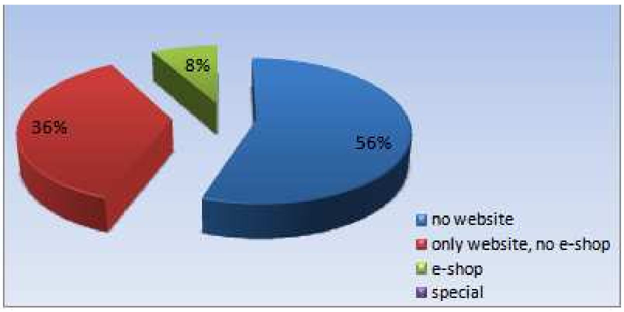


Figure 2. MSME e-commerce in Freiburg.

With merely 5.8% of all MSMEs in Magdeburg and 8.2% in Freiburg being able to accept online orders, the results are very similar. It is also important to note that out of the 1859 enterprises reviewed in both cities, only 102 (5.4%) can be considered as large enterprises that in one way or another act as retailers. Among those large enterprises, the websites of all 102 businesses (100%) feature an online ordering system.

While 93.1% of all MSMEs (together in Freiburg and Magdeburg) not being able to accept online orders seems to be quite a large percentage, we are next going to hypothesize in our KULI model reasons for MSMEs to yet reject the option of electronically exposing their products and services on the internet.

V. BUSINESS-SIDE E-COM-E-COM FACTORS

When enterprises consider the introduction of an electronic storefront for their business, there are many aspects that play into the decision making process. An online shop might, for example, only be feasible for enterprises which do not depend too much on drop-in customers, enterprises which offer products that are easy to describe or enterprises which have access to appropriate logistics facilities or logistics service providers. Another problematic aspect is that, as most enterprises do not have IT as their major field of business, a significant portion of MSMEs might lack the knowledge and technical understanding that is required for running e-commerce. Others might be too focused upon maintaining their existing business infrastructure to investigate the possibilities of e-commerce or might simply lack the time to take related initiatives. Again others might find e-commerce to be an appealing option for extending their market, but might have products that require certain agility from logistics service providers (e.g., perishable goods, such as food). Last but not least, running an online shop might simply be too expensive for certain MSMEs (cp. TCO of e-commerce in [5]) where this could mean that -due to the structure and products of the business- the e-commerce facility does not generate enough return on investment (ROI) to be an option to pursue or to even come up for the running cost of the online shop.

VI. KULI Model of E-Commerce Maturity

In our KULI model, we have picked up above thoughts, whereby KULI stands for Knowledge, Unawareness, Logistics and Investment. With KULI, we try to raise attention for the problems and difficulties MSMEs might face with respect to a possible adoption of e-commerce.

Knowledge: In order to setup, run and maintain an online shop, a variety of technical skills is required. Starting with the creation of a database and the setup of a web server towards the installation of the online shop software and the deployment to the web server, the majority of MSMEs would already get lost before they have specified even their first online product. Enterprises could compensate their lack of technical expertise by purchasing service hours of specialized companies and consultants. As we will see later, the cost for outsourcing the technical administration represents another major barrier for MSMEs towards the adoption of e-commerce.

Unawareness: As we have indicated earlier already, many MSMEs might simply be unaware of the chances e-commerce could bring to their business. We hypothesize that this is most often due to the size and the limited revenue MSMEs operate upon, so that no or only very limited resources exist within the organization that could deal with strategic questions and new technologies. To our opinion, it has always been and it will always remain a core problem of all e-business related research to communicate the many benefits of e-commerce to MSME decision makers in a way, so that they can access them and clearly identify where information system technology could add value to their organization.

Logistics: The third aspect the KULI model tries to point out is logistics. If enterprises depend on an own logistics facility in order to dispatch their online products, they will in most cases have to run their own fleet. Running an own fleet, however, is expensive and thus most enterprises would immediately withdraw this option, if no positive ROI was within sight. For certain businesses, however, it turned out to be an appealing extension of their business model to run e-commerce in addition to their traditional distribution channel. Pizza deliveries, restaurants or even pharmacies, for instance, offer to immediately dispatch their orders by using their own fleet. Businesses that wanted to online expose their products to a local market (as in the case of the pizza delivery) would also need to be able to rely upon an own fleet, or at least a logistics service provider that is capable of rapidly dispatching orders to local customers [11]. In this context, one should also evaluate the feasibility of a shared logistics provider or fleet that could be owned by a conglomerate of local businesses. This way, local MSMEs could share the cost for a fleet, which might at the same time be a solution for overcoming the cost barrier of an instantaneous dispatch handling. For many MSMEs, a shared logistics model could even represent the enabler for a future e-commerce go-live.

Investment: Last but not least, the KULI model of e-commerce adoption points out the role of the initial investment that is to be brought up for introducing e-commerce in an enterprise. Though several MSMEs might already have thought about adding e-commerce to their order handling, another significant number might find the investment for software, hardware, administration and consulting services to be out of proportion in the respective case of their business. In an earlier work, we have referred to this phenomenon where MSMEs might be attracted by the chances and possibilities of e-commerce, but on the other hand might not be able to bring up the initial investment as “MSME dilemma” [12]. Though critics might argue that nowadays there exists a variety of open source online shop software that can be procured for free, we would like to emphasize that MSMEs would still need to know how to set up the software (refer to “KULI/Knowledge”). Though so called rental shops, which represent online shops that are deployed and hosted by specialized companies, eliminate the procurement and installation of soft- and hardware on the MSME’s side, the majority of enterprises will still not know how to customize the software to reflect their individual requirements or how to model products in their online shop. Again, the services of administrators or companies specialized on content management would need to be purchased. One large-enterprise real-world example for this can be found in the German T-Systems that not only hosts large-scale e-commerce suites, but also handles the content management for their customers. Though similar service providers have appeared in the small and medium segment (e.g., Open IceCat), services offered might yet be too expensive, of a limited functional scope or simply assume MSMEs to at least have basic level of information system support available in their organization.

We have summarized all four KULI factors in Figure 3, where we propose one question for each KULI factor. Only if all four questions are answered with “Yes”, an MSME is mature enough to start running an online shop or engage in e-commerce.

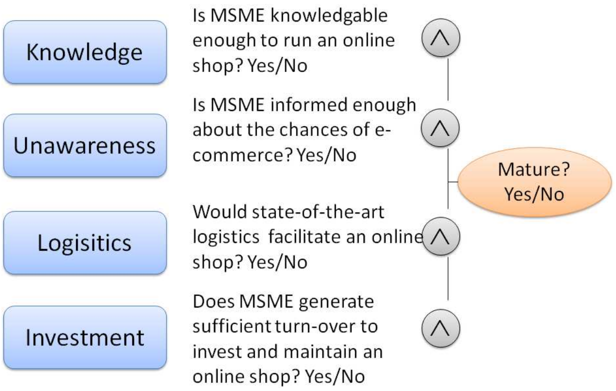

Knowledge
Is MSME knowledgable enough to run an online shop? Yes/No
Unawareness
Is MSME informed enough about the chances of e-commerce? Yes/No
Logisitics
Would state-of-the-art logistics facilitate an online shop? Yes/No
Investment
Does MSME generate sufficient turn-over to invest and maintain an online shop? Yes/No
Mature? Yes/No

Figure 3. KULI model of e-commerce maturity.

With the questionnaire above, MSMEs are given a tool to evaluate and review their KULI maturity. Enterprises with a positive KULI maturity (overall "Yes") should seriously start considering an introduction of e-commerce to their business. Enterprises with a negative KULI maturity (overall "No") should start analyzing how to increase their maturity by, for example, increasing their knowledge or awareness. Furthermore, MSMEs with a negative KULI maturity should evaluate how to deal with "logistics" or how to bring up the funds necessary for the initial investment.

VII. TOWARD AN E-COM-E-COM DESIGN MODEL

In the previous section, we have introduced a model for outlining the reasons of why traditional e-commerce technology is not yet accepted at large scale by micro businesses. In a related work, we have talked about Organic Product Catalogs (OPCs) [13] that aim at providing a smooth entry for MSMEs to electronic commerce. Throughout this section, we are going to detail the behinds of organic product catalogs and will demonstrate how a cloud platform serves as the ideal platform to develop upon.

VIII. ORGANIC PRODUCT CATALOGS

The concept of organic product catalogs (OPCs) tries to provide a platform that is very close to the end-user and operates at minimum cost. The closeness is expressed as OPCs being driven by their user community; their content is made by the community for the community. The general intention behind this closeness argument is that if person A has created something in the OPC that could be useful for person B, B can very easily reuse this for his purposes. We will later detail how this could look in practice.

One fundamental requirement to an OPC is universality: No matter how easy or complicated a product or service is, it must be describable in the OPC. Therefore, OPCs at their very basis expose a universal product definition language (PDL) to the user that allows for the description of any arbitrary product or service. The assumption that underlies the PDL is that each and every product can be expressed in hierarchical form as a composition of attributes. In accordance with the "divide-and-conquer" principle, complex products are expressed as a hierarchical composition of attributes. Figure 4 is featuring a simple PDL tree of a car that consists of the attributes "make", "cost", "color" and "engine", whereby some of the attributes have one or multiple child attributes.

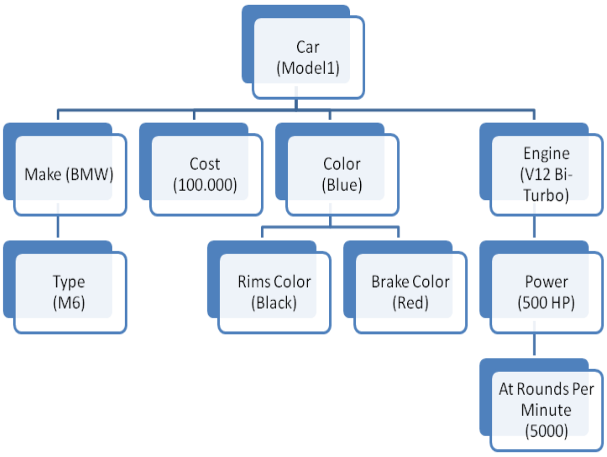


Figure 4. A sample PDL tree featuring a car.

Furthermore, the PDL allows for an on-attribute-level specification of dynamic behavior. When designing a product description for a product or service, users can bind so called attribute handlers to each and every attribute. The attribute handlers modify, update or interpret the attribute value before it gets returned to the calling instance. Attribute handler definitions are available in the OPC's attribute handler repository and can be freely bound to each attribute. In above example, the attribute "cost" could, for example, have an attribute handler "GetInCurrency" that returns the cost of the car in a user-specified currency by applying always up-to-date exchange rates.

While the PDL corresponds to the universality requirement, at the same time it violates another requirement: ease of use. Even though users are now given a tool to describe whatever product they have, most of them will most likely not "speak" the PDL. It is complicated, it is complex and it is something that needs to be learned. In order to design the OPC in a way so that it is usable for the majority of potential users, templates of commonly defined product descriptions are generated from the PDL. Hence, users, if they want to input their products into an OPC, first of all try to find a template that corresponds well with their product. In the next step, they customize the template (e.g., provide concrete values for placeholders) and save it to the OPC.

If a template for a product does not yet exist, the user can author a new template by using the PDL. He can submit the template to the OPC where it is going to be automatically reviewed and possibly added to the standard template base. This way, the own-authored template of one user becomes available to the whole community. If similar templates already exist, the OPC suggests using those or will try a merge resulting into a new template. As over time, users keep providing new or modifying existing product templates, the repository of product templates gains momentum. It is unpredictable how many templates the OPC will consist of and with that which products are maintained at a time t. The closeness to the community ensures that OPCs continuously reflect what is required by their users.

IX. Cloud Implementation Platform

So far, we have discussed how OPCs can add value to MSMEs. We have, however, not shown yet how OPCs can ensure a minimum cost for participation, so that they are accessible to every business, regardless of the size. Our solution to the problem of minimizing the cost for participation aims at minimizing the Total Cost of Ownership (TCO) of OPCs.

Ensuring easy-to-access maintenance and extension for future efforts [14] is one component of optimizing the TCO. Building OPCs as applications in the cloud, however, would probably grant the major portion of cost benefits. Figure 5 depictures that we designed our reference implementation as a cloud application as we are convinced by the many benefits of this paradigm with respect to the OPC scenario.

Figure 5 provides two perspectives on a cloud-hosted OPC. In the horizontal perspective, based on the cloud platform, the tenant layer ensures that each participant has access to an own customized content area.

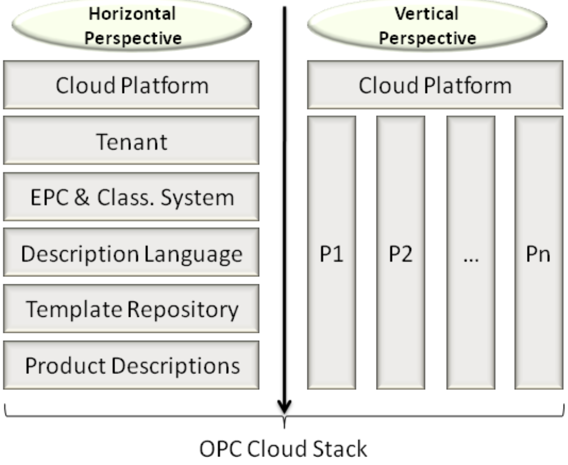


Figure 5. Cloud stack of an OPC [13].

The actual EPC distinguishes and individually manages the product descriptions for each client separately and thus provides full support for multitenancy. Multitenancy means enabling many "tenants" to use a central utility, in this case the OPC. The rationale behind multitenancy is scaling: In applications that are built to scale well, the operating cost for each client (the per-tenant TCO) will drop as more clients are added [15]. Furthermore, the OPC provides a general purpose product description language (PDL) that can be used to describe the products and services of each tenant. In most cases, however, tenants will use templates that were created by other tenants earlier for describing their products. Finally, the bottom-most layer of our OPC centrally stores all product descriptions that a tenant is maintaining. The vertical perspective (or the per-tenant perspective) visualizes that on each layer in the horizontal perspective, tenants are provided with their own customized (multitenant) experience.

X. Project Goliath – The Digital City

Inspired by the idea of providing MSMEs with an instrument to easily and cost-effectively model and expose their products and services on the internet, we are now planning the rollout of our OPC prototype to a delimited economic entity, such as the city of Magdeburg or Freiburg.

We refer to this effort as the "Goliath Project" [11] where we will eventually be trying to capture and "digitize" all products and services provided by the businesses in one city ("The Protocity") and make them available for online purchase.

For that, we are adding a website frontend to our OPC that will allow citizens of the Protocity to browse through all online offerings and place orders. If physical goods are ordered, another, not yet raised issue is complicating things: logistics. Up to now, we were only dealing with the software side and have completely neglected the fact that the orders would also need to get dispatched in a timely manner. In reality, however, we would need to be able to support a variety of logistics-related scenarios. Hot food, for example, would need to get dispatched within a very short time, medicine ordered from an online pharmacy maybe even faster. For that reason, we have developed a model that tries to reflect the different requirements of physical goods that could potentially be purchased online.

As the main focus of the Goliath project lies upon the technology that is necessary for "digitizing" a city, we were looking for an elegant way to deal with the logistics problem. Part of our solution was to leave the intelligence that is required for route planning and way optimization to the logistics service providers (LSPs) themselves. The LSPs on the other hand can offer their parcel services on the Goliath platform by modeling them with our PDL, just like every other service would be modeled. While Goliath provides the information system infrastructure for LSPs to offer their services and get booked for a delivery, it is our hope that from an LSPs' point of view Goliath will become an interesting market to advertize their services. In this sense, we would in particular encourage semi-LSPs (fleet owners that do not primarily act as LSPs (e.g., pizza deliveries, pharmacies)), as well as smaller local LSPs to take their chances and expand their markets into the Goliath business network. Though we would appreciate large LSPs, such as DHL, UPS or Fedex, to also participate in Goliath, we believe that as they would need to bring up significant investments to create an agile and locally operating fleet that can dispatch orders within minutes, in the beginning large LSPs will rather hesitate to join the Goliath effort. For this reasons, we anticipate the first LSPs to adopt Goliath being local LSPs or semi-LSPs that would like to expand into new markets.

XI. Conclusion and Future Work

In this article, we have introduced an approach to turn e-commerce into e-commodity and referred to this as E-Com-

E-Com. We have argued that the fundamental idea of E-Com-E-Com is to provide e-commerce as a commodity service to micro, small and medium enterprises with the intention to one day cover the vast majority of all products and services offered in a city.

With our KULI model of e-commerce maturity, we have tried to outline the business-side E-Com-E-Com factors that determine an MSME's e-commerce readiness. Though we believe that the only way for MSMEs to achieve KULI maturity lies in the businesses adjusting the factors they are short on by themselves, we have also stated that technology-side E-Com-E-Com factors can at least make it easier for MSMEs to achieve KULI maturity.

While having local businesses equipped with e-commerce technology would allow end-consumers to benefit from a variety of new and exciting digital shopping experiences (e.g., order delivery at a specified time by a local retailer), there are also aspects we need to carefully consider when continuing our work on E-Com-E-Com, OPCs and the Goliath project.

One of the biggest challenges we will be facing is providing local retailers with a way to dispatch their orders in an agile and timely manner to their customers. Though we have ideas of how to mitigate this problem (e.g., binding semi-LSPs and other local ISPs to our platform), we are not sure yet whether the business model behind will work out in practice. Another problem that might evolve from our approach of how to deal with logistics is trust. We need to find a way to ensure that those businesses that offer to dispatch orders for other businesses become a reliable and trustworthy part of the Goliath business network.

Since the development of our OPC and the preparations for Goliath are still under development, we plan on rolling out our platform to the first tier of MSMEs in the beginning of 2011.

REFERENCES

[1] Fritz, W. 2004. Internet-Marketing und Electronic Commerce. Gabler Verlag. Wiesbaden, Germany.

[2] Benjamin, I. and Wigand R. 1995. Electronic Markets and Virtual Value Chains on the Information. Sloan Management Review, Winter 1995, pp. 62-72.

[3] Adler, J. 1996. Informationsökonomische Fundierung von Austauschprozessen. Eine nachfrageorientierte Analyse. Gabler Verlag. Wiesbaden, Germany.

[4] European Commission. 2005. The new SME definition – User guide and model declaration. Enterprise and Industry Publications. [Online]. Available: http://ec.europa.eu/enterprise/policies/sme/files/sme_definition/sme_user_guide_en.pdf [Accessed: Dec. 30, 2010].

[5] Neumann, R., Schmietendorf, A., and Dumke, R. 2010. Cloud-based Organic Product Catalogs – A Highly Pervasive E-Business Approach for Micro Enterprises. Proceedings of the 2010 International Conference on e-Learning, e-Business, Enterprise Information Systems, and e-Government, Las Vegas, USA, pp. 123-129.

[6] Access eCommerce. Ecommerce Trends. 2010. University of Minnesota. [Online]. Available: http://www.access-ecom.info/article.cfm?id=85&xid=MN. [Accessed: Dec. 30, 2010].

[7] BITKOM - Bundesverband Informationswirtschaft, Telekommunikation und neue Medien e.V. Zwei Drittel aller Haushalte nutzen Ende 2010 Breitband. Berlin, Germany. [Online]. Available: http://www.bitkom.org/de/themen/54890_62900.aspx [Accessed: Dec. 30, 2010].

[8] Stahl, E., Breitschaft, M., Krabichler, T., Breitschaft, M., and Wittmann, G. 2009. E-Commerce-Leitfaden - Erfolgreicher im elektronischem Handel. ibi research. University of Regensburg. Regensburg, Germany. [Online]. Available: http://www.ecommerce-leitfaden.de/studien [Accessed: Dec. 30, 2010].

[9] E-Commerce-Center Handel (ECC Handel). 2010. Definitionen zu E-Business und E-Commerce. Cologne, Germany. [Online]. Available: http://www.ecc-handel.de/definitionen_zu_e-business_und__e-commerce.php. [Accessed: Dec. 30, 2010].

[10] Statista GmbH. 2009. Populärste Produktkategorie im Online-Handel nach der Anzahl der Käufer in Millionen (Hochrechnung). Berlin, Germany. [Online]. Available: http://de.statista.com/statistik/daten/studie/153035/umfrage/populaerste-produktkategorien-im-online-handel/ [Accessed: Dec. 30, 2010].

[11] Neumann, R. and Sokolova, K. 2009. Goliath – The Digital City. Painting the Picture of a New Era of E-Commerce (Whitepaper), Otto-von-Guericke University Magdeburg. Magdeburg, Germany.

[12] Neumann, R. 2010. The EBF Application Foundation: An Approach towards the Design of an E-Commerce Framework for Small and Medium Enterprises. Lambert Academic Press. Saarbrücken, Germany.

[13] Neumann, R., Schmietendorf, A., and Dumke, R. 2010. Organic Product Catalogs - Towards an Architecture for Cloud-based Micro Enterprise E-Commerce. IEEE Cloud 2010 - The 3rd International Conference on Cloud Computing, Miami, USA, pp. 530-531.

[14] Neumann, R., Günther, S., and Zenker, N. 2009. Reengineering Deprecated Component Frameworks: A Case Study of the Microsoft Foundation Classes, Proceedings of the WI2009. Vienna, Austria, pp. 737-746.

[15] Chong, F. and Carraro, G. 2006. Architecture Strategies for Catching the Long Tail. MSDN. [Online]. Available: http://msdn.microsoft.com/en-us/library/aa479069.aspx [Accessed: Dec. 30, 2010].

Understanding Mobile Payment Service in University Campus: A Context-awareness View

Hsiao-Chi Wu
Dept. of Information Management
National Taiwan University
Taipei, Taiwan
hciwu69@gmail.com

Jen Wel Chen
Dept. of Information Management
National Taiwan University
Dept. of Business Administration
Chinese Culture University
Taipei, Taiwan
d95725009@ntu.edu.tw

Ching-Cha Hsieh
Dept. of Information Management
National Taiwan University
Taipei, Taiwan
cchsieh@im.ntu.edu.tw

***Abstract*—Mobile payment services have gradually become popular and so drawn researchers' attention. Currently, most research has focused on issues of the technology itself and consumer's acceptance of this technology, but paid little attention to possible social and cultural ramifications. Mobile payment services can be used at various locations by different types of users, resulting in complex contexts and difficult design challenges. This study first examines the context of mobile payment services, and then comprehends how this context affects its implementation. A context-awareness framework is adopted to compare two cases in different Taiwanese university campuses. This framework investigates the significance of contextual factors during the adoption of mobile payment services, such as location, people, objects and their interaction. Our findings suggest that, for mobile payment services, (1) the socio-spatial dimensions of geographical regions affect consumer adoption; (2) the developer's social interpretation of technological artifacts facilitates their innovation; (3) and a deep understanding of contextual interactions enables the creation of a "must-have" mobile payment service, and helps identifies its niche.**

Keywords-mobile payment; context-awareness; location-based services; university campus

I. INTRODUCTION

The growing popularity of mobile payment services has increasingly drawn the attention of academic researchers. Some scholars argue that because mobile payments (MP) allow real-time, cashless, and wireless transactions for buying goods and services at any location or time, these payments will become a successful mobile service [1, 2].

This predicted success of mobile payment services is based upon not one single dimension, but rather considerations of many different aspects. Dahlberg et al. [3] point out that most MP research focuses on issues concerning the technology and consumers. Some researchers have explored the issue at the national level, such as MP environments [4-6], while others investigate MP standardization [7, 8] or provide simple overviews [9]. However, few researchers have adopted multiple perspectives of study [2, 10], and none have addressed the social and cultural influences of MP [3].

The purpose of this paper is to understand the context surrounding MP services. MP designers pay close attention to mobile environments when designed their applications and improving the functionality of MP services [23]. When using mobile technology, rapidly changing environmental factors are an important issue. Thus, development of MP services should not only consider consumers' technological acceptance but also the surrounding context, such as location, time, activity, physical objects, etc. Environmental changes will trigger various needs and activities, and affect the choices of payment services. Some consumers may be interested in MP for reasons beyond payments [11], and if a certain payment scheme is available for multiple purposes and at various locations, consumers will find it more useful [10]. Current research has paid little attention to the impact of situational changes on the implementation of MP services.

Due to this research gap, this paper asks two research questions. First, how does context affect MP services implementation? Secondly, what contextual factors should MP developers and service providers focus on when facing different contexts/environments? In examining these questions and mobile payment services, this paper investigates two case studies with different context environments: two Taiwanese university campuses undergoing MP services adoption.

The paper is organized as follows. Section II reviews the literature related to mobile payments. Section III describes data collection strategies and introduces a context-awareness framework. Section IV explains case background. Section V analyzes these cases. Conclusions and discussion are provided in Section VI.

II. MOBILE PAYMENTS

Customer's acceptance of MP mainly depends on issues of cost, security, and convenience [7]. The technology-based schemes of MP are divided into card-based and phone-based. The card-based scheme remains the preferred scheme for payments [12], and adopts the pre-paid solution [10]. Some successful MP systems match specific customer needs that enhance the convenience of micropayments for daily local expenditures, such as public transportation (e.g., Hong Kong's Octopus is one of the most successful smart cards in the world), toll booths (e.g., E-ZPass), and fast-food restaurants (e.g., McDonalds) [13]. MP systems are suitable for proximity and micropayments, and should not be limited to mobile commerce [7, 13].

Dahlberg et al. [3] constructed a framework for the literature reviews of two applied theories – the five forces model [14] and the generic contingency theory (73 papers, 1999 - 2006) on MP. The rate of technological study is about 40% of the related researches; these studies mainly focus on conceptual technical constructions, and the proposal of tools or mechanisms for MP transactions and security. The rate of consumer research is about 27%, and these researches explore consumer adoption/acceptance, and focus on ease of use, compatibility, cost, trust, and usefulness for adopted MP services. It is regrettable that there is no research that investigates the effects of social and cultural changes on the development of MP.

Specific social and cultural issues may influence MP adoption, such as demographics, lifestyle characteristics, and cultural differences [13]. Payment services refer to an individual's location. The location factor is one of the essential factors that separates mobile commerce from e-commerce [15]. As individuals remain in different contexts, the various needs of consumption affect their choices of payment services [3]. Under the mobile computing, the application and development of MP services should pay more attention to the context of social influence; however, the current researches neglect this issue.

University campuses promote card-based MP services, which are used as a payment tool as well as for student ID cards and access to buildings [16, 17], these functions that enhance administrative efficiency [16], reduce administration costs, and increase incomes [18]. Mirza and Alghathbar investigated smart card applications in universities around the world, including North America, West Europe, and Asia. They surveyed 20 universities with 34 applications, and generalized the most popular four applications, which are student ID cards, book borrowing, and access to libraries and photocopies [17]. Related applications can be generalized into three categories, namely, (1) student ID card as identification; (2) entrance cards for libraries, buildings, parking lots, etc.; and (3) e-purses for consumption (e-purses is a prepaid card which can be used for payment instead of coins [26]). This paper explores the impact of contexts on MP services and elaborates on the multiple-purpose applications of MP.

III. RESEARCH METHOD

A. Data Collection Strategies

This paper adopts a case study approach, allowing the researcher to connect the research phenomenon with actual situations [19], in order to recognize the dynamics of these phenomenon. Primary data collection methods include participant observation, in-depth interviews, secondary data, and questionnaires (see Table I). Adopting these multiple sources of data collection helps ensure data authenticity and repeatedly validity.

TABLE I. DATA COLLECTION STRAGEGIES

Data type	**Two Cases**	
	TU	**CU**
Participation observations	Observation period: 2008.07-2010.05	Observation period: 2006.03-2010.05 One author works in CU
In-depth interviews	11 person-interviews.	10 person-interviews.
	Each interview lasted 90 to 120 minutes. Participants included mobile payment team leaders, developers, merchants, etc.	
Secondary data	14 copies	32 copies
	Proposals, project reports, meeting minutes, etc.	
Questionnaires	1,852 users, such as students, staffs, and teachers	--

B. A Context-Awareness Framework

Mobile services focus on the "anywhere, anytime" method of development, as mobile computing does not occur at a single location in a single context, but rather spans a multitude of situations and locations [20]. The greatest challenge is to design successful mobile services that deal with multiple contexts. Individuals engaged in related activities may trigger individual payment situations and other MP services, such as photocopying, overdue fines on library books, etc. As MP services face dynamic environments, MP developers must be aware of the context of MP services, and be able to design an MP system that satisfies consumers' needs.

Schilit and Want [20] defined context as "*a constantly changing execution environment.*" Context can include computing, users, physical location [20], and even social situations [27]. Dey, Abowd and Salber [21] defined context as "*any information that can be used to characterize the situation of entities (i.e., whether a person, place, or object), that are considered relevant to the interaction between a user and an application, including user and application themselves. Context is typically the location, identity, and state of people, groups, and computational and physical objects.*" In this paper, three entities were identified—places, people, and things [21, 22].

- People include individuals or groups.
- Places apply to geographical spaces, such as rooms, offices, buildings, or streets.
- Things refer to physical objects, software components, or artifacts, etc.

In order to describe these entities, four characteristics of context are listed, as follows.

- Identity refers to assigning a unique identifier to an entity.
- Location includes position information and spatial relationships between entities.
- State (or activity) identifies the observable properties of an entity.

- Time characterizes a situation, either as a timestamp or as a time span.

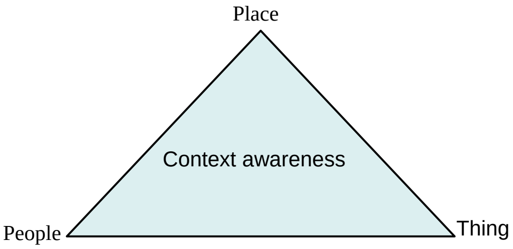


Figure 1. A context-awareness framework sourced from Dey, Abowd, and Salber [23]

In order to develop MP applications that satisfy consumers' needs, this study constructs a context-awareness framework based on the definition of context described by Dey, Abowd and Salber [21] (see Figure 1). The primary advantage of their approach is its inclusion of a complete but generalized scope of context, that can then be applied to different applications [22]. Application designers choose what context best fits an application upon an in-depth understanding of their contextual options. Similarly, application designers also determine what context-awareness behaviors are required to support their applications, through an in-depth understanding of how a given context is used [23].

IV. Case Background

This study examines two universities in Taiwan (anonymous: TU and CU) as case subjects. TU is a public university, while CU is private. After some time, both schools adopted the same card service as their campus MP system. This paper compares the consequences of MP implementation in distinct settings in order to understand the significance of contextual factors for the adoption of MP.

A. Case Study: TU TCard

TU is one of the top national universities in Taiwan. In 2007, TU replaced their student ID cards (TCard) with contact-less and store-value smart cards, namely, the EasyCard, in order to provide more convenient services for faculty and students. The new smart TCard is used not only for identification and entrance access to locations, but also as a payment card. Since September 2008, there have been 16 campus shops that provide MP services. In June 2009, TU disseminated an on-line questionnaire to investigate usage of the TCard among faculty and students, in response to decreasing MP consumption.

B. Case Study: CU UPass

In 1998, CU was the first university to issue MP cards (UPass) in Taiwan, and three well-regarded versions of MP cards have been used in the development of MP applications. Since 1998, CU has attempted to develop more services for the UPass. In 2005, CU again served as a pioneer and cooperated with the largest transportation card system (EasyCard) in northern Taiwan. CU changed the UPass from a debit card into an RFID smart card, which then provided various store-value services with contact-less Mifare standards. Since the School of Continuing Education (SCE) at CU operates independently of the university, the MP card application was first implemented by the academic centers of SCE, initiated under the planning of the IT Department of SCE.

In 2006, SCE introduced 12 campus services for UPass, including student identification, entrance access to specific locations, libraries, parking lots, venue reservations, photocopying, vending machine use, services for school administration, and intelligent management of classrooms in integrated academic buildings. As a result, numerous universities asked the SCE of CU to help promote MP systems in their own universities.

TU and CU eventually both selected the EasyCard as their campus MP system. EasyCard is utilized by northern Taiwan's largest transportation systems, and its development is similar to Hong Kong's Octopus card. Since April 2010, EasyCard has extended its service to transactions at over 10,000 locations, including four major convenience store chains, coffee shop chains, drug store chains, restaurants, fast food stores, and parking lots. Thus, the MP services of TU and CU expanded beyond their campuses to incorporate public transportation and other MP services.

On the whole, CU UPass adoption was more successful than the TU TCard. This is surprising, because UPass has fewer users and a more complex e-purse mechanism. Using a context-awareness framework (Figure 1), this study examines the differences between MP implementation in these two university campuses, and determines the significance of contextual factors during the adoption of MP services.

V. Analysis

This section analyzes the development and adoption of MP services in TU and CU, and employs a context-awareness analysis framework (Figure 1) as the analytic lens.

A. The Context of Places

TU is a comprehensive university with an open campus environment designed for leisure and public use. This campus includes academic and administrative buildings, libraries, dormitories, a post office, bank, museums, lakes, stadiums, arenas, etc. For consumption, there are over 50 retail stores on campus including cafeterias, coffee shops, restaurants, convenience stores, bookshops, a souvenir store, etc., but only 19 of these are equipped with MP services where patrons can use their EasyCard. TU is located near a famous shopping area and an MRT (subway) station. When interviewed, students offered several responses to this situation: *"It would be better if all stores on campus accepted the TCard, otherwise it is inconvenient;"* *"after school, I used to have dinner off-campus and not shop on* *campus.*" Furthermore, some complained, "*there is no obvious signage at stores to identify shopping with the TCard*."

In contrast, the SCE of CU is a closed environment, with eight academic centers located at the center of the city. However, the space for student activities is limited, as most space is utilized for classes, studying, discussions, and administration. Students can shop in campus stores with UPass, such as coffee shops, convenience stores, and photocopy centers.

B. The Context of People

After students at TU and CU register, they are issued a smart student card, which is embedded with an MP mechanism. According to our investigations and observations, the age and origin of TU and CU students are very similar. These students can quickly accept and easily use the card. Investigation of Lee, Cheng and Depickere [18] points out university's students generally accept using their university smart card.

The students of TU's main campus exceed 33,000. On-campus activities at TU are broad, and include studying, research, dining, living, sports, and entertainment. Additionally, many tourists enter campus for fun and shopping. According to our survey questionnaire, 70% of respondents agreed that the TCard is more convenient than paying by cash, and 62% of respondents were willing to use the TCard for payments. In fact, for 80% of respondents, average expenses per week are lower than 100NT dollars (approximately 3US dollars). Most students must pay administration service fees, such as photocopying (83%), vending machines (61%), overdue book fines (59%), and stadium rentals (55%). Respondents suggested that "*I want to use the TCard for all of photocopy machines and vending machine on campus.*" They want "*more discounts or reward mechanisms,*" and noted that the "*TCard just pays for bus and MRT.*"

The SCE of CU is actively devoted to the adult education market, and types of students include undergraduate students, part-time students with jobs, and senior learners. Thus, the campus activities of SCE's students focus on studying and discussions at the academic center.

C. The Context of Things

TU and CU adopted the same MP card system (Easy Card). This MP card is a RFID smart card, which then could provide store-value services with contact-less Mifare standards. Its usage is easy and convenient.

The TU TCard has only one e-purse: basic store value. The TU campus has only three convenience stores for adding value to their MP cards, and the minimum total amount for each value transaction is at least 500NT dollars. Students can also add value to their cards at MRT stations. Students responded that "*setting up more cash-to-card machines is good, but you cannot limit the amount of money I can add*"; and "*it would be helpful if the Internet could provide services for tracking purchases.*" Some complained that "*it is not that much faster if students shop with their TCard, and this decreases overall shopping*;" "*there is a receipt when paying with the TCard, which is a waste of paper and increases check-out time.*"

In contrast, the CU UPass has three e-purses: off-line dollars, on-line dollars, and store value. Off-line dollars and on-line dollars are only used on campus. Off-line dollars pay for photocopying, overdue book fines, and vending machines, while on-line dollars pay for net-printing and school services applications. Unlike TU, SCE's academic center has kiosks for adding value and checking balances, and students can also ask service center personnel to add value to their cards.

Here, the different contexts of MP services result in different implementation consequences. Table II shows the summary of the context-awareness dimensions for TU's and CU's MP cards.

TABLE II. CONTEXT-AWARENESS DIMENSIONS ON TU AND CU FOR MP CARDS

Context	TU	CU
Places	● Identity: open environment ● Location: research centers, academic buildings, shops, lakes, museums, arenas, etc.	● Identity: closed environment ● Location: SCE's academic buildings, convenient stores, coffee shops, photocopy centers, etc.
People	● Identity: students, faculty, staff, and tourists ● Activity: studying, reading, teaching, research, sports, shopping, living, traveling, etc. ● Time: all day	● Identity: students, faculty, staff, and tourists ● Activity: studying, reading, teaching, discussions, shopping, etc. ● Time: business hours
Things	● Identity: MP card ● State: MP card has just one e-purse ● Activity: payment, entrance access to buildings	● Identity: MP card and kiosks ● State: MP card has three e-purses; kiosk has value-added and balance functions ● Activity: MP card and kiosks are multi-purpose/ multi-function

VI. DISCUSSION

Students of TU and CU generally accept MP cards, and recognize their convenience. However, while the two cases have similar organizational contexts, the development and implementation of MP services in both are very different from each other. From our research findings, the additional dimensions of contextual factors influence the development direction of MP services and their implementation.

A. Socio-spatial dimensions of MP card adoption

The use of technological artifacts by individuals depends not only on their convenience, cost, and the user's technology acceptance, but also the context of their use [7]. Aoyama [24] suggested that socio-spatial dimensions, such as urban form, consumer preference, and cultural attributes all shape the patterns of technology adoption. Our observational results support Aoyama's findings.

The TU campus is an open environment surrounded by a famous shopping area, which means that there are many places for dining, shopping, and entertainment. Most sales of TCards are from the EasyCard issuer, and off-campus stores can also have this type of sale. Thus, it is difficult to attract students shopping with the TCard. Moreover, not all stores on campus accommodate TCards, which influences both the ability and willingness of students to use the card. Unsurprisingly, TCard purchases are decreasing.

In contrast, SCE's academic center of CU is a closed environment. All stores at CU provide shopping services with UPass. When shopping for vending machines, photocopying, and paying overdue book fines, UPass is an all-in-one card, and is convenient throughout the entire campus.

In other words, the interactions between places (e.g. socio-spatial dimensions) and consumer preferences affect the usage of MP. TU students have more opportunity for shopping in the famous shopping areas, not on their campus, and even though the TCard provides some payment convenience, students have alternative choices for purchases in this open environment, which results in uncertainty for MP usage. In a closed environment, the CU UPass serves as a payment tool in addition to other campus functions, which facilitates its continued use.

B. The service-oriented developments of MPcard

The MP card has physical limitations such as sending, receiving, or presenting information. However, the application developments of MP card are different between TU's and CU's. The information service for payment is taken as an example. The TU TCard provides an immediate paper receipt for each transaction, but students regard it as a waste of paper and prefer online inquiries. The SCE of CU applies multi-platforms to provide information services for students to check their balance, namely kiosks and the Internet. Jacob [12] points out that customers might be interested in MP for reasons beyond the payments themselves. In sum, UPass services broaden and have greater usefulness than the TCard.

The IT Department of SCE suggested that, other than the access control of buildings and parking lots, the entrance access of the UPass should also include venue reservations and spatial management. The IT manager of SCE stated: "*the purpose of UPass entrance access is not to control the people coming in and going out; it is about spatial management*." Teachers and students may directly reserve classrooms or discussion rooms online and access locations through the UPass. Security personnel monitor the states of classrooms and also manage campus safety by the PDA. The CU UPass is more than just a payment tool; it is also an all-in-one card, with multiple purposes and functions. These outcomes are the result of contextual interaction among people, places and things, and thus not simply due to any single contextual variable. Moreover, it is based on service-oriented approaches to apply multi-platforms, e.g., kiosks, internet, and PDA, to extend the scopes of UPass services.

The schemes of MP card can not actively capture the information of its involved environment in comparison with the phone-based MP schemes. MP card designers must pay more attention to the possible contextual interactions before they create new services. Furthermore, this study finds that a designer has more understanding on the contextual interactions, and then he/she would give more social interpretation of technological artifacts to develop specific innovations of MP services. The service-oriented development of MP card would more closely fit into the customer's specific needs; accordingly, it enhances the usability and functionality of the MP card..

C. The niche of MP services

In a mobile environment, the key successful point of MP services is that it is a "must-have" rather than "nice-to-have" service [25]. There are some failed cases [28], of which most fell into the "nice-to-have" strategy.

The TU TCard adoption has a "nice-to-have" strategy, as its MP services are just an alternative choice, but not necessary. For example, TU students expect to use the TCard for payment specific to the school's activities, such as photocopying and overdue fines for library books, and stadium space rental. .They also suggest adding more cash-to-card machines on campus. However, these expected shopping services have not been introduced, which also causes low usage of the TCard.

In contrast, the adopted strategy of the CU UPass is "must-have", such as photocopying, net-printing, venue reservation, payment listings, and access control to buildings, etc. These functions must use the MP card and fit with user's specific needs. CU provides two kiosks with cash-to-card machines in each building, which generates positive reactions from students, makes the CU UPass necessary on campus, and enables students to continually use the card. In summary, the innovative services of CU UPass consider all possible contexts of campus activities and integrate campus information, cash flows, and resources together. This all-in-one card is a result of strong linkage among contextual factors, i.e., people, place and things, which represents a niche for MP services.

VII. CONCLUSION

Consumers can use MP technology anywhere at anytime, which opens the technology up to a host of potentially complex contexts and renders the MP design process increasingly difficult. Complex contexts affect not only the implementation of MP services, but also determine the development of MP services. Here, two case studies adopt the same MP card in similar organizational settings, but with very different consequences of implementation, illustrating the interaction effects of contextual factors.

- The contexts of place include not only identity, location, state and time but also socio-spatial dimensions, which affect the adoption of consumer's MP services. In an open environment, the lack of boundaries of geographical regions raises more uncertainties for MP services, even though consumers may accept these services and agree to their convenience.

- The developers must have more awareness of involved contextual interactions for MP card during development. Designing MP applications must go beyond MP card's physical limitations, which is to base upon service-oriented contexts and apply multi-platforms to extend service scope. The developers' social interpretation of technological artifacts facilitates the creation of new MP services, and enables MP services beyond simple payment options.
- A strong linkage of contextual interactions among people, places and things, shapes a "must-have" strategy for MP development, and then identifies niches for MP services. MP services provide a positive impact on consumer's lives, which in turn helps ensure consumers continue using MP services.

This study selected university campuses as its research settings, which provided successful and failed experiences on MP service implementations. These findings can serve as reference for future studies, which may apply multiple in-depth cases to explore closer cooperation with users and merchants in a different natural setting to create more possibilities for MP services.

REFERENCES

[1] A. Herzberg, "Payments and banking with mobile personal devices," Communications of the ACM, vol. 46, no. 5, pp. 53-58, 2003.

[2] J. Ondrus, and Y. Pigneur, "Towards a holistic analysis of mobile payments: A multiple perspectives approach," Electronic Commerce Research and Applications, vol. 5, no. 3, pp. 246-257, 2006.

[3] T. Dahlberg, N. Mallat, J. Ondrus, and A. Zmijewska, "Past, present and future of mobile payments research: A literature review," Electronic Commerce Research and Applications, vol. 7, no. 2, pp. 165-181, 2008.

[4] K. Stroborn, A. Heitmann, K. Leibold, and G. Frank, "Internet payments in Germany: a classificatory framework and empirical evidence," Journal of Business Research, vol. 57, no. 12, pp. 1431-1437, 2004.

[5] J. Ondrus, and Y. Pigneur, "A Systematic Approach to Explain the Delayed Deployment of Mobile Payments in Switzerland," in Proceedings of the Fifth International Conference on Mobile Business(ICMB) Copenhagen, Denmark, 2006.

[6] P. Jaring, T. Matinmikko, and P. Abrahamsson, "Micropayment business in Finland-forming the basis for development of micropayment methods and business," in Proceedings of Helsinki Mobility Roundtable, Helsinki, Finland, 2006.

[7] N. Kreyer, K. Turowski, and K. Pousttchi, "Mobile payment procedures: scope and characteristics," e-Service Journal, vol. 2, no. 3, pp. 7-22, 2004.

[8] A. S. Lim, "Inter-consortia battles in mobile payments standardisation," Electronic Commerce Research and Applications, vol. 7, no. 2, pp. 202-213, 2008.

[9] A. Zmijewska, "Evaluating Wireless Technologies in Mobile Payments -- A Customer Centric Approach," in Proceedings of the Fourth International Conference on Mobile Business (ICMB), Sydney, Australia, 2005.

[10] Y. Chou, C. Lee, and J. Chung, "Understanding m-commerce payment systems through the analytic hierarchy process," Journal of Business Research, vol. 57, no. 12, pp. 1423-1430, 2004.

[11] K. Jacob, "Are mobile payments the smart cards of the aughts?," Chicago Fed Letter , Federal Reserve Bank of Chicago, No. 240, July, 2007.

[12] J. Ondrus, and Y. Pigneur, "A multi-stakeholder multi-criteria assessment framework of mobile payments: An illustration with the swiss public transportation industry," in Proceedings of the 39th Annual Hawaii International Conference on System Science, Hauai, HI, USA, 2006, pp. 42a.

[13] J. Ondrus, and Y. Pigneur, "A Disruption Analysis in the Mobile Payment Market," in Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Big Island, Hawaii, 2005, pp. 84c.

[14] M. Porter, Competitive strategy: techniques for analyzing industries and competitors: with a new introduction, New York: Free Pr, 1998.

[15] I. Junglas, and R. Watson, "Location-based services," Communications of the ACM, vol. 51, no. 3, pp. 65-69, 2008.

[16] C. H. M. Lee, Y. W. Cheng, and A. Depickere, "Comparing smart card adoption in Singapore and Australian universities," International Journal of Human-Computer Studies, vol. 58, no. 3, pp. 307-325, 2003.

[17] Abdulrahman A. Mirza, and K. Alghathbar, "Acceptance and Applications of Smart Cards Technology in University Settings," Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing, pp. 746-748, 2009.

[18] C. Clark, "Shopping Without Cash: The Emergence of the E-purse," Economic Perspectives, Federal Reserve Bank of Chicago, issue Q IV, pp. 34-51, 2005.

[19] R. Yin, Case study research: Sage publications Newbury Park, Calif, 1994.

[20] B. Schilit, N. Adams, and R. Want, "Context-aware computing applications." pp. 85-90.

[21] A. Dey, G. Abowd, and D. Salber, "A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications," Human–Computer Interaction, vol. 16, no. 2, pp. 97-166, 2001.

[22] M. Debes, A. Lewandowska, and J. Seitz, "Definition and Implementation of Context Information," in Proceedings of the 2nd workshop on positioning, Navigation and Communication (WPMC '05) & 1st Ultra-Widenband Expert Talk (UET '05), 2005, pp. 63–68.

[23] A. Dey, "Understanding and using context," Personal and ubiquitous computing, vol. 5, no. 1, pp. 4-7, 2001.

[24] Y. Aoyama, "Sociospatial dimensions of technology adoption: recent M-commerce and E-commerce developments," Environment and Planning A, vol. 35, no. 7, pp. 1201-1222, 2003.

[25] S. L. Jarvenpaa, K. Lang, Y. Takeda, and V. K. Tuunainen, "Mobile commerce at crossroads," Communications of the ACM, vol. 46, no. 12, pp. 41-44, 2003.

[26] L. Van Hove, "Electronic purses in Euroland: why do penetration and usage rates differ?," SUERF Studies, M. Balling, ed.: Vienna, Austria: Société Universitaire Européenne de Recherches Financières, 2004.

[27] O. Kwon, K. Yoo, and E. Suh, "ubiES: Applying ubiquitous computing technologies to an expert system for context-aware proactive services," Electronic Commerce Research and Applications, vol. 5, no. 3, pp. 209-219, 2006.

[28] C. Clark, "Shopping Without Cash: The Emergence of the E-purse," Economic Perspectives, Federal Reserve Bank of Chicago, issue Q IV, pp. 34-51, 2005

An Actor Network Theory Lens for Mobile Commerce: A Mobile Payment Case Study

Hsiao-Chi Wu
Dept. of Information Management
National Taiwan University
Taipei, Taiwan
hciwu69@gmail.com

Cheng-Chieh Huang
Dept. of Information Management
National Taiwan University
Taipei, Taiwan
d95725007@ntu.edu.tw

Ching-Cha Hsieh
Dept. of Information Management
National Taiwan University
Taipei, Taiwan
cchsieh@im.ntu.edu.tw

***Abstract*—Mobile device technology and mobile commerce are deeply embedded in many people's everyday lives. In this paper, we adopt Actor Network Theory to understand how consumers, merchants, service providers, mobile devices, applications, and services interact in aligned networks. Using a mobile payment system adoption case, we identify social-technical problems encountered by users. We examine how they redefine "convenience" while performing different activities under space-time limitations. The results imply that m-commerce technology and service designers should consider strategies that emphasize activity-oriented design, activity-based response, social-technical fit, and that lock users in a virtual closed network.**

Keywords-mobile payment system; actor network theory; system perspective.

I. INTRODUCTION

Devices and applications based on mobile technology are now commonplace in everyday life. In recent years these devices have moved beyond wireless phones, wireless-enabled handheld computers, and personal digital assistants (PDAs) to include global positioning systems (GPS), location-based services, and electronic payment systems. People expect these devices and applications to make their lives more convenient, while merchants aspire to earn money through mobile commerce transactions.

Traditionally, mobile devices and applications emphasize an "anywhere, anytime" development scheme, but m-commerce services are now at a crossroads [1]. Experts are increasingly discussing the reasons behind mobile technology adaptation and user intentions [2-6]. Despite these efforts, relatively little is known about how users, merchants, mobile technologies, applications and services interact in everyday activities or social life.

In this paper, we examine social actors and technologies in m-commerce environments from a system perspective. Actor Network Theory (ANT) is used to understand how technical (i.e. mobility applications, technologies) and different social actors (i.e. consumers, merchants) align and interact in a social-technical system, called an actor network. In testing this theory, we selected a complex social and technical mobile payment case study, a public Taiwanese university.

We address three questions. First, how do focal actors, such as merchants or issuers, enroll users to join the actor network? Secondly, how do these technical and social elements interact, and what are the consequences? Thirdly, under what conditions do users accept or reject the m-commerce actor networks these focal actors construct into their social life? Using the ANT lens to interpret our case study, we argue that m-commerce technology and service strategies should focus on activities-oriented designs rather than location-based or context-aware concepts.

The remainder of this paper is organized as follows. Section II illustrates Actor Network Theory. Section III outlines our research approach, while section IV describes our Taiwanese university mobile payment application case. Here we identify main actors, social environments, and relevant technical elements. We then analyze and discuss how the focal actors enroll users, and how the mobile payment system enables or constrains users (section V). Finally, we examine why the technical system is accepted or rejected by users (section VI). The concluding section (section VII) lists our contributions and the limitations of our research.

II. ACTOR NETWORK THEORY

Actor Network Theory (ANT) was developed within the sociology of science and technology [7]. A key feature of ANT is that it treats networks as involving both human and nonhuman actors (such as a technology) [7][8]. The theory examines the motivations and actions of human actors who align their interests around the requirements of non-human actors.

Within ANT, inscription and translation are pivotal concepts. Engineers who design, develop and diffuse a technical artifact inscribe into this artifact their instructions for use, their intentions, and their vision of the society and world in which the artifact best fits. In this sense, they become sociologists, or in Callon's word, engineer-sociologists. The technical aspects of the engineer's work are profoundly social. Therefore, it is difficult to distinguish between the technical and the social during the process of innovation. When we accept that the technical is social, the artifacts on which engineers inscribe their own social preferences become entities; in ANT terminology, an actor with the same nature and characteristics is regarded as a human actor. The distinctiveness of ANT is that it does not distinguish between human and nonhuman actors.

ANT helps describe how actors form alliances, enroll other actors, and use nonhuman actors (artifacts) to strengthen such alliances and to secure their interests. This process, called translation, is defined as “the methods by which an actor enrolls others” [8]. When an actor–network is created, translation consists of four stages [8].

- Problematization: The focal actors define interests that others may share, and establish themselves as indispensable resources in the solution of the problems they have defined. They define the problems and solutions and also establish roles and identities for other actors in the network. As a consequence, focal actors establish an “obligatory passage point” for problem solutions that all the actors in an actor-network must pass.
- Interessement: The focal actors convince other actors that the interests defined by the focal actors are in fact well in-line with their own interests. Through interessement the developing network creates sufficient incitement to both lock actors into networks.
- Enrollment: Enrollment involves a definition of the roles of each of the actors in the newly created actor-network. It also involves a set of strategies through which focal actors seek to convince other actors to embrace the underlying ideas of the growing actor-network and to be an active part of the whole project.
- Mobilization: The focal actors use a set of methods to ensure that the other actors act according to their agreement and remain loyal. With allies mobilized, an actor network achieves stability.

That is, the ANT lens is suitable for analyzing how social actors and technical elements negotiate and interact during the design and adoption phases of mobile technology.

III. Research Approach

A. Data Collection Strategies

Informed by this interaction process as a guiding framework, our subsequent research strategy sought in-depth case study data that would give further insight into the complex interaction processes in m-commerce environments. For this we contacted T University, a university in Taiwan that has struggled to introduce their mobile payment system on campus. This system involves many participating groups, including university mobile payment project teams, the mobile payment cards service provider, merchants and users, students, staff, and the teachers of T University.

We adopted multiple data collection strategies (see Table 1). First, we conducted in-depth interviews with the mobile payment project teams at T University and the mobile payment cards service provider, so as to understand their design intentions and issues. We also interviewed merchants about their experiences using mobile payment systems. Secondly, we distributed questionnaires to students, staff, and teachers to understand their use and interaction experiences. Third, we held two student focus groups to understand why they used or rejected the mobile payment system. Moreover, we gathered secondary data such as the mobile payment system’s proposal, function specifications, and meeting minutes. This secondary data helps supplement our survey and interview sources [9].

TABLE I. Data Collection Strategies

Strategies	Targets	Frequencies
In-depth interviews	Mobile payment project team	5 persons
	Merchants	6 persons
Questionnaires	Users (students, staffs, and teachers)	1,852 questionnaires
Focus groups	Students	2 * 5 persons
Secondary data	Proposals, meeting minutes, functional specifications	14 copies

B. Analysis Framework

Our analytical approach was to understand the different participating groups’ perspectives, use and interaction experiences relating to the T University mobile payment environment. Following our approach and ANT, we constructed our analytical framework (Figure 1).

Using this framework, we can understand how the mobile project team inscribed their interests on the mobile payment system during the design phase and how the users, merchants and mobile payment system interacted during the technology’s adoption.

IV. Mobile Payment Systems and Actors in T University

A. Mobile Payment Systems

Mobile payments are defined as the use of a mobile device to conduct a payment transaction in which money or funds are transferred from a payer to a receiver via an intermediary, or directly without an intermediary. Mobile devices include mobile phones or any wireless enabled device (e.g. PDA, laptop, card, watch).

Payment technology can be classified as card-based and phone-based systems [10]. Several successful mobile payment systems have already been launched in order to enhance the convenience of micro-payments for daily local expenditures. These solutions have been principally adopted by various fast service-oriented industries such as public transport (e.g. Octopus), toll booths (e.g. EZPay, FasTrak), gas stations (e.g. ExxonMobils Speedpass), fast food restaurants (e.g. McDonalds), retail vending machines (e.g. Sonera Mobilepay) and ski resort ticketing (e.g. Skidata) [11].

A payment market can be examined in terms of payment service providers, technology, and users. Payment service providers are typically financial institutions, network operators or independent issuers [12]. Users are divided into two different and demanding groups of adopters: consumers and merchants. In general, the mobile payment adoption environment is complex, and involves many social and technical elements.

B. Social Actors and Technical Elements in the Case of T University

Mobile payments systems in T University were introduced in August, 2008. T University administrators considered that "*mobile payments systems, combined with EasyCard and students' or staffs' identity cards will be convenient for payment needs on campus*," according to their meeting minutes on December, 2007. The intention of combining the university card with the public transportation EasyCard (issued by E Company) would increase its overall utility and provide simpler access to the largest public transportation system in northern Taiwan. That is, the new student/staff identity cards would conveniently pay for public transportation, parking, and other purchases on campus.

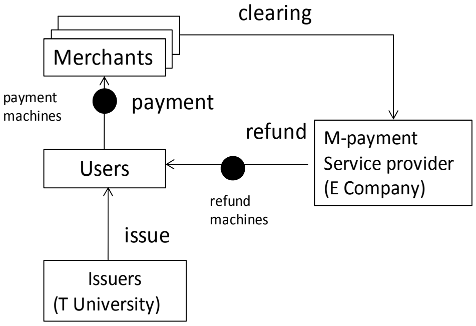


Figure 2. The mobile payment system business model in T University

The business model for this system is illustrated in Figure 2. Users, such as students, staff, and teachers of T University refill the payment cards (their identity cards) using refill machines, and pay merchants through payment machines, called EDC (electronic data collector). Merchants will send this transactions information to E company for every transaction. Merchants pay a 0.5 % transaction fee to E company, which is cheaper than credit card transaction fees.

In summary, the mobile payment system in T University is a card-based system. The service provider is E company, and the issuer is T University. The technical elements include mobile payment cards, payment machines, and refill machines. The social actors are T University authorities, E Company, and participating merchants, students, staff, and teachers.

V. A Mobile Payment Case in T University

Based on our analytical framework, the following are the interactions of social actors and technical elements in the design and adoption phases in the T University case.

A. Design Phase

The design phase began in January, 2008 after the T University combined EasyCard with student/staff identity cards in September, 2007 (see Figure 3). The T University project manager then invited the E company to design the mobile payment system on campus.

This was not the first time E company had introduced their EasyCard as a mobile payment system for institutions. The most successful case was D Hospital, which introduced EasyCard for patients to pay for their registration fees and medicine. Other education institutions have also incorporated these cards. Prior to implementation, T University surveyed more than forty merchants on campus, who initially welcomed the mobile payment system.

However, it was not easy to persuade these merchants to join the payment system. One E company representative said, "*When talking about transactions fees, only ten merchants wanted to join!*" The representative continued, "*This case is very different from D Hospital or other education institutions. T University administrators did not sufficiently help persuade merchants*."

The T University project manager responded, "*We are different from D Hospital or other private Universities!*" He said, "*It is state-run university, and our culture is free and respectful to everyone. We will not and cannot force every merchant and student or teacher to use the mobile payment system*."

A complication included differences between merchant points of sales (POS) systems. The card system required considerable integration costs. One of mechants said "*Why we promote more while costs/benefits are not equal?*" Nineteen merchants finally joined card payment services after T University agreed to pay the machine rental for two years.

B. Adoption and Consequences

To encourage user adoption, the T University project team conducted a promotional activity, where using a mobile payment card during the first month of the adoption phase earned the user a prize. They also made posters for participating merchants to attract user adoption. The project team and T University administrators viewed the cards as a valuable convenience.

TABLE II. Priority of Mobile Payment Usage Services

Services	Percentage of agreement	Introduced in the adoption phase
copy services	83%	No
vending machine	61%	No
book fines	59%	No
space rental	55%	Yes
other consumption	6%	Yes (partial merchants)

Despite this perception however, adoption has been surprisingly weak. According to our survey questionnaires, weekly average payment amounts per user are roughly 100 NT dollars, which is less than one-day average spending on campus. Although about 70% users agreed that the mobile payment card is more convenient than paying by cash, the cards are primarily used to pay for school administration activities, such as copy services, space rental, book fines, and vending machines (see Table II).

When asked, the first reason respondents did not use the payment card frequently was the risk of card loss. An interviewee said, "*I think that the payment card is small and easier to lose than cash. If I lose my payment card I lose my student ID card!*" Another interviewee described, "*I added only a small amount of money into the payment card to reduce risk*." A user explained his experiences, "*One day, I used my mobile payment card to pay for dinner at a coffee shop. I found the amount is not enough. But I really do not want to leave because my seat will probably be taken while I go to refill my payment card. It left me embarrassed!*" As a result, users will not make large purchases with their mobile payment cards. Students also complained that only a few merchants accept mobile payment cards; they still need to bring cash in their pocket to campus. Moreover, payment speed is not faster than paying by cash. "*Most of time, I still use my mobile payment card to pay for public transportation. I paid by mobile payment cards only because I was running out my coins*!" an interviewee explained.

Merchants also encounter problems with the mobile payment system. A merchant said, "*Sometimes users want to reorder, but the mobile payment system cannot withdraw*." A restaurant salesclerk also said, "*We have three salesclerks so that, for cash payers, they could choose any line for transaction during peak hours. However, for card users, they have no choice, but need to stand in the only line for the card payment machine, which sometimes takes longer time.*"

In summary, it seems convenient for users and merchants to use the mobile payment system, but in different social-technical interaction situations, they encounter different problems, and the system is not as convenient as focal actors expected. The payment cards are mostly used with the public transportation system beyond the T University campus.

VI. Analysis and Discussions

Based on our analytical framework, the following are the interactions of social actors and technical elements during the design and adoption phases in the T University case.

A. *Inscription and Translations*

As described within the design phase, the T University holds a "respect towards every merchant" policy, and did not push any merchants to join the mobile payment system (see Table III). Even after T University decided to pay for a two-year payment machines rental, only nineteen merchants joined. For E company, integration costs were too high, with little expected benefit. T University also did not want to invest much in installing refill machines. Finally, inscription of the mobile payment system on T University campus is like a "proof of concept" in a laboratory.

During the adoption phase, users encountered different social or technical problems in different situations, but the focal actors could not solve these problems. Merchants did not actively promote the mobile payment system. E company did not want to invest additional effort in resolving technical problems. Finally, users did not frequently use the card on campus. While the mobile payment card seemed convenient for students, these social-technical problems hampered overall convenience. Designers or merchants should thus consider such issues when solving these scenario problems, and not only technical functions or consumers' intentions.

TABLE III. Translation Strategies in the T University Case

Phase	Focal Actors	Translation
Design phase	T University, E company	1. problematization: T University defines "convenience for students" 2. interessement & enrollment: T University did promote system among merchants. E company also considered costs too high to form the network. 3. mobilization: merchants will remain loyal if T University pays for two years of payment machine rental.
Adoption phase	T University, E company, merchants	1. problematization: "convenient for students," but users encountered different problems in different situations. 2. interessement & enrollment: except for first month prize activity, little promotion. Merchants did not actively promote mobile payment system to users. 3. mobilization: did not have any agreement between users and merchants.

B. *From Physical Location to Actor Network*

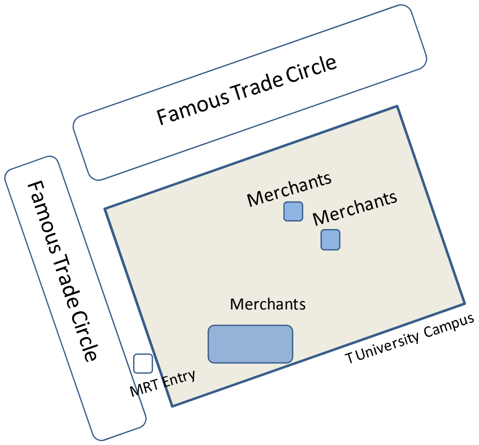


Figure 4. Location context of T University

Like other universities in Taiwan, T University is surrounded by many merchants beyond campus (see Figure 4). Moreover, the campus is near a famous trade circle, with many restaurants, coffee shops, and clothing stores. After school, most students will thus not stay on campus and instead explore the nearby shopping area.

Such an environment, where actors can easily enter or leave to conduct transactions, is called an open network. Compared with the closed MRT network environment, it is not easy to lock users into consuming on campus. Physical location still limits users' mobile commerce behaviors. Thus, when introducing mobile payment or mobile commerce services, it is worthy to consider the relationship between location context and users.

But from the ANT theory viewpoint, the focal actors can take strategies to form a 'virtual' closed actor network space to escape the physical location limitation. For example, actors may provide more discounts or other marketing activities for participating merchants on campus. Other strategies include persuading more merchants to join the actor network, installing more payment machines on campus or enrolling other non-school members to consume on campus by mobile payment cards.

Some students compared the mobile payment system on T University with a famous convenience store chain, with three stores located on campus. *"I do not know why, but when I enter the convenience store, I will always find myself paying by mobile payment card!" "The payment speed of the mobile payment card is also slow, but I would like to pay by the card when purchasing at the convenience store."* The convenience store installed refill machines, and introduced different promotion activities every two months. That is, the convenience store chain arranged suitable social and technological elements, and then constructed a virtual closed actor network while customers entered their many-location stores.

C. *Activity-Oriented Mobility*

Sometimes the location is not the only factor to consider; activities in specific space-time also matter. For example, students consider that they will use their cards most frequently for copy services, book fines, space rental, and vending machines (see Table II). Often they will remain a long time in a specific space, such as the library, while doing these activities. Users defined "convenience" as doing different activities within their space-time limitations. That is, the focal actors should consider different strategies according to different space and time-limited activities (see Figure 5). Recently, more and more mobile commerce literature focuses on location-based or context-aware services and their applications [3]. Balasubramanian et al. [13] suggest that mobile technologies change the flexibility of activities along the spatial and temporal dimensions, and also impact people's activity pattern. Therefore, a designer needs to take additional consideration in activity-oriented thinking when designing mobile devices, applications or services.

Although we expect mobile technology to achieve ubiquitous world dreams, most of time, mobile commerce is embedded within our living world, and activities within a specific space and time. Designers of mobile commerce applications or services should consider how to arrange the social-technical elements of different specific activities.

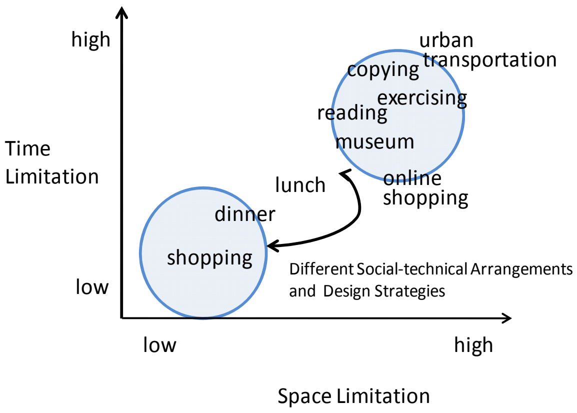


Figure 5. Activity-oriented mobility design

D. *Implications for Technology and Service Strategies*

As discussed above, both location and space-time matter for mobile commerce design. Based on the above analysis and discussion, we can indicate future directions for mobile commerce system technology and service design strategies (see Table IV).

Table IV indicates possible directions for design solutions, such as an activity-sensitive interface, activity-based responses, etc. Mobile systems can record activity, usage information and responses for different activities. Using T University as an example, mobile payment cards or service providers can record user-purchasing information, and merchants can respond with different promotion activities according to this purchasing data. Although some devices, such as the payment cards in our case, cannot store as much information as a smart phone, designers can still consider ways to use these cards for information storage, as social activities and other technical devices support these features. This is referred to as a social-technical fit for different activities (see Table IV).

Moreover, although ubiquitous computing is a dream for technology, focal actors should consider strategies for locking their users in a specific space-time environment: an actor network. This involves mobilizing more users to stay in an actor network and strengthening the stability of such a network.

TABLE IV. IMPLICATIONS FROM TECHNOLOGY AND SERVICE STRATEGIES

Strategies	Features
Technology	● activity-sensitive interface ● activity-based responses ● activity/event time configuration ● behavior information presentation ● usage summary information. ● location-sensitive responses
Service	● understanding target users' activities ● different strategies for various activities ● social-technical fit for activities ● event-based marketing ● lock users in a virtual closed network

ISBN: 978-1-61208-116-8

VII. CONCLUSIONS

In this paper, we take ANT theory as an analytic lens to understand the social-technical interactions of a mobile payment application case. We argue that mobile commerce is not only location-sensitive but also represents a social-technical arrangement of specific time-space activities. For designers, it is not sufficient to simply create a ubiquitous payment system, but an arrangement of applications, services, devices and activities that are embedded within the user's living world. These findings imply the following principles for designers:

- follow and understand user's everyday activity
- activity-based sensitivity and responses
- social-technical fits for different activities
- location-sensitive responses
- design a virtual closed actor network and strategies to lock users

In this study, we adopt multiple data collection strategies, and analyze both quantitative and qualitative data. Our primary limitation however is our single mobile payment case study. In future work, we will compare other institutions and different mobile commerce applications cases to understand the complex social-technical interactions within mobile commerce environments.

REFERENCES

[1] S. L. Jarvenpaa, K. R. Lang,, Y. Takeda, and V. K. Tununainen,: "Mobile Commerce At Crossroads," Communications of the ACM, vol 46, 2003, pp. 41-44.

[2] S. Sarker, and J. D. Wells, "Understanding Mobile Handheld Device Use and Adoption," Communications of the ACM, vol 46, 2003, pp. 36-40.

[3] I. A. Junglas, and R. T. Watson, "Location-Based Services," Communications of the ACM, vol 51, 2008, pp. 65-69.

[4] P. A. Pavlou, and M. Fygenson, "Understanding and Predicting Electronic Commerce Adoption: An Extension of the Theory of Planned Behavior," MIS Quarterly, vol 30, 2006, pp. 115-143.

[5] Y. Aoyama, "Sociospatial Dimensions of Technology Adoption: Recent M-commerce and E-commerce Developments," Environment and Planning, vol 35, 2003, pp. 1201-1221.

[6] E. W. T. Negi, ,and A. Gunasekaran, "A Review for Mobile Commerce Research and Applications," Decision Support Systems, vol 43, 2007, pp. 3-15.

[7] M. Callon, and B. Latour, "Unscrewing the big Leviathan," in Advances in Social Theory and Methodology, K. Knorr-Cetina and A.V. Cicourel, Eds, London: Routledge & Kegan, 1981, pp. 277-303.

[8] M. Callon, "Some Elements of a Sociology of Translation: Domestication of the Scallops and the Fishermen of St. Brieuc Bay," in Power, Action and Belief, J. Law, Ed, London: Routledge and Kegan, 1986, pp. 197-233.

[9] R. K. Yin, Case Study Research, Design and Methods, CA: Sage Publications, 1994.

[10] J. Ondrus, and Y. Pigneur, "A multi-stakeholder multi-criteria assessment framework of mobile payments: An illustration with the swiss public transportation industry," in Proceedings of the 39th Annual Hawaii International Conference on System Science, Hauai, HI, USA, 2006, pp. 42a.

[11] J. Ondrus, and Y. Pigneur, "A Disruption Analysis in the Mobile Payment Market," in Proceedings of the 38th Annual Hawaii International Conference on System Sciences, Big Island, Hawaii, 2005, pp. 84c

[12] Y. Au, and R. Kauffman, "The economics of mobile payments: Understanding stakeholder issues for an emerging financial technology application," *Electronic Commerce Research and Applications*, vol. 7, no. 2, pp. 141-164, 2008.

[13] S. Balasubramanian, R. Peterson, and S. Jarvenpaa, "Exploring the implications of m-commerce for markets and marketing," Journal of the academy of Marketing Science, vol. 30, no. 4, pp. 348-361, 2002.

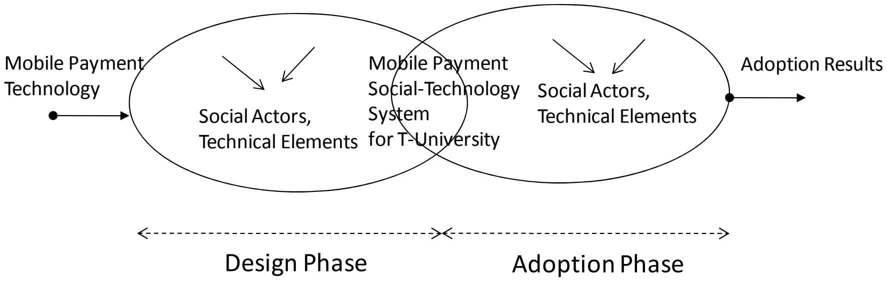

Mobile Payment Technology
Social Actors, Technical Elements
Mobile Payment Social-Technology System for T-University
Social Actors, Technical Elements
Adoption Results
Design Phase
Adoption Phase

Figure 1. Analysis framework based on ANT Theory

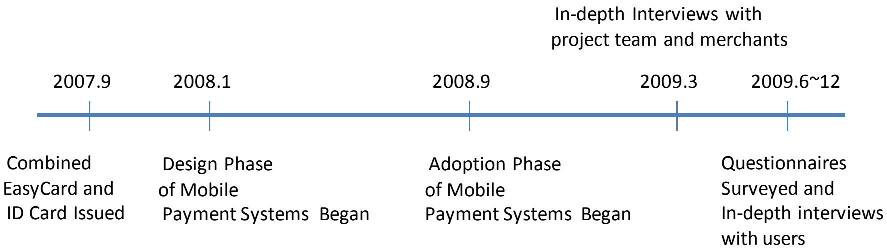

In-depth Interviews with project team and merchants
2007.9
2008.1
2008.9
2009.3
2009.6~12
Combined EasyCard and ID Card Issued
Design Phase of Mobile Payment Systems Began
Adoption Phase of Mobile Payment Systems Began
Questionnaires Surveyed and In-depth interviews with users

Figure 3. The mobile payment system implementation schedule of T University

ISBN: 978-1-61208-116-8

Generic Contract Descriptions for Web Services Implementations

Balazs Simon, Balazs Goldschmidt, Peter Budai, Istvan Hartung, Karoly Kondorosi, Zoltan Laszlo, Peter Risztics
Department of Control Engineering and Information Technology
Budapest University of Technology and Economics
Budapest, Hungary
Email: {sbalazs | balage | bucnak | hartungi | kondor | laszlo}@iit.bme.hu, risztics@ik.bme.hu

***Abstract*—The basic building blocks of SOA systems are web services. The domain specific language SOAL developed by the authors has a Java and C#-like syntax for describing web service interfaces and BPEL processes. The paper introduces an extended version of the language that supports Design by Contract. From the service contract specifications software artifacts are generated that check pre- and post-conditions on the server side at runtime, applying the delegation pattern. The proposed solution provides Design by Contract for both JAX-WS and WCF technologies used in most SOA products in the industry.**

***Keywords*-SOA; web services; Design-by-Contract; modelling; DSL.**

I. Introduction

Complex distributed systems are best built from components with well-defined interfaces and a framework that helps connecting them. Web services and BPEL (Business Process Execution Language) processes implementing WSDL (Web Services Description Language) interfaces represent nowadays the components that both enterprises and governments use to construct their complex distributed systems, thus implementing a Service Oriented Architecture (SOA) [1].

One of the advantages of building on web services technology is that one can get a vast set of standards from simple connections to middleware functionalities such as security or transaction-handling. [2]

The other advantage is that SOA applies classical principles of software engineering, like model-based development, loosely coupled components, separation of interface from implementation, etc., at a higher level of abstraction. Another classical principle is "Design by Contract" (DbC) introduced by Bertrand Meyer [3]. Contract is a key concept in SOA, thus it seems to be obvious to apply DbC for WebServices. Vendor support for it is usually minimal, and although there are third party solutions, the common drawback is their platform-dependency. Heterogeneous systems like those applied in e-government need a unified description of contracts that can be generally used in different SOA products. Having legacy systems the introduction of a new product or extending an existing product with native contract support is not an option. This paper shows how a high-level web service description language and code generating framework was extended with contract support in order to enable developers to automatically generate platform-specific contract-enforcing modules from the general contract specifications.

The activity of our research group aims at developing principles, recommendations, components and technologies that make the application of SOA in the e-government domain easier. Considering the diversity of legacy systems of different governmental organizations, all the solutions we elaborate should be platform-independent as far as it is possible.

The rest of the paper introduces a platform-independent solution of DbC in SOA systems, which can work in any SOA environment even it would be a heterogenous one. In the second section, related work is examined. In the third section, the contract metamodel, its representation in SOAL syntax, the architecture and details of code generation, and an overall evaluation are presented. In the fourth section, the results are summarized.

II. Related Work

Eiffel [3], [4] is an object-oriented programming language that supports Design by Contract. The code is either compiled to native code or to .NET CLR, where it could be applied web services written in the .NET framework. The JVM is not supported, therefore it cannot be used in Java environments.

D. Florescu et.al. [5] proposed a declarative domain specific language with pre- and post-condition support for web services. Their goal is to create web services that can be executed on any platform. However, instead of generating code for the different commercial products, they created a custom execution environment, the XL virtual machine. Their solution also lacks a metamodel behind the language.

WS-CoL [6] is primarily a monitoring language, however, it provides pre- and post-conditions on the interactions between services and it can also be used for BPEL processes, although it supports only the ActiveBPEL engine. WS-CoL has an extension with time constraints described in [7].

Another way of describing services is using semantic web technologies. The major goal of Semantic Web Services (SWS) is to create intelligent software agents to provide automated, interoperable and meaningful coordination of web services [8]. The three main directions of SWS are SAWSDL (Semantic Annotations for WSDL), OWL-S (Semantic Markup for Web Services) and WSMO (Web Services Modeling Ontology).

SAWSDL [9] does not introduce a new language. It is a WSDL extension for referencing ontological concepts outside WSDL documents. Beyond that it does not define any execution semantics for the implementation.

The OWL-S [10] profile ontology is used to describe what a service does, and is meant to be mainly used for the purpose of service discovery. The service description contains input and output parameters, pre- and post-conditions, and also non-functional aspects. The OWL-S process model describes service composition including the communication pattern. In order to connect OWL-S to existing web service standards, OWL-S uses grounding to map service descriptions to WSDL. The OWL-S environment provides an editor to develop semantic web services and a matcher to discover services. The OWL-S Virtual Machine is a general purpose web service client for the invocation. OWL-S therefore requires a custom execution environment and cannot be used in current commercial SOA products. Its underlying description logic OWL-DL has also a limited expressiveness in practice.

The WSMO [11] framework provides a conceptual model and a formal language WSML for semantic markup of web services. WSMO is used for modeling of ontologies, goals, web services and mediators. Ontologies provide formal logic-based grounding of information used by other components. Goals represent user desires, i.e., the objectives that a client might have when searching for services. Web services are computational entities, their semantic description includes functional and non-functional properties, as well as their capabilities through pre- and postconditions, assumptions and effects. Mediators provide interoperability between components at data, protocol and process level. The reference implementation of WSMO is the WSMX [12] framework, a custom execution environment. It is designed to allow dynamic discovery, invocation and composition of web services. It also provides interoperability with classical web services.

The main design goals of SWS standards are discovery, invocation and composition of web services. These standards are not primarily designed for modeling purposes; they are not supported by the major SOA software vendors and they require a custom execution environment.

Other efforts focus on modelling web services in UML. R. Heckel and M. Lohmann [13] introduce three levels of contract representations: implementation-level, XML-level and model-level. Their goal is to derive implementation- and model-level contracts from the model-level specification. Our aim is similar, but we think that a domain specific language and a metamodel for SOA are much more appropriate for this purpose. J. T. E. Timm [14] defines a UML profile that extends class and activity diagrams. This profile is used in transformations to automatically construct OWL-S specifications from diagrams and SWRL (Semantic Web Rule Language) expressions from OCL. The problem with this approach is the same as with SWS technologies: it cannot be applied in current major SOA products.

III. Design by Contract for SOA

This section presents the contract metamodel, its representation in SOAL syntax, the architecture and details of code generation, and finally, the overall evaluation of the proposed framework.

A. SOAL and SoaMM

In the SOA world, XML is used for interface- and process-description, and message-formatting. The aim is interoperability, but the drawback is that handling and transforming XML documents above a certain complexity is almost impossible. This is why most development environments have graphical tools for helping developers creating interface-descriptions, connections, process-flows and message-formats. The problem with the graphical approach is that it is neither efficient, nor reliably repeatable, nor sufficiently controllable, nor easily automatable.

On the other hand, the standards usually have a lot of redundant parts that result in poor readability and manageability. For example the `message` element in WSDL 1.x was omitted from WSDL 2.0 because of its redundancy. The development tools do not support the inclusion of special extensions in the interface specification (like pre- and post-conditions, authorization, etc.) Even some tools have special naming conventions that are to be accepted, otherwise the generated code is even less readable than necessary. We have examined a lot of products [15] and have found a lot of peculiarities, which have to be taken into account beyond the recommendations of the WS-I Basic Profile. In BPEL process-descriptions the `partnerLinkType-partnerLink` constructs for partners, or the `property-propertyAlias-correlationSet` constructs for correlations mean unnecessary redundancy. Unfortunately, the process designer tools usually map these constructs directly to the graphical interface instead of hiding them from the users.

To solve the above problems, an abstract SOA metamodel called SoaMM has been developed that can manage both BPEL and WSDL concepts, and, in order to describe the model, an extensible language called Service Oriented Architecture Language (SOAL) was specified [16], [17] that can be used for describing webservice interfaces and BPEL processes, and is more easily readable and manageable by humans. This model and language also enables automatization, vendor-specific WSDL and BPEL generation, and compile time type checking.

The following example illustrates a simple stack web service description in SOAL. The service has the URL http://localhost/Stack and can be accessed through SOAP 1.1 over HTTP:

```
namespace StackSample {
  interface IStack {
    void Push(int number);
    int Pop();
    int Top();
    bool IsEmpty();
  }
  binding Soap11HttpBinding {
    transport HTTP;
    encoding SOAP { Version = "1.1" }
  }
  endpoint Stack : IStack {
    binding Soap11HttpBinding;
    location "http://localhost/Stack";
  }
}
```

B. Design by Contract in SOAL

We extended SOAL with contract descriptions. A contract can be specified using the `contract` keyword and must implement exactly one interface. The implementation of an operation has to specify the pre-conditions with the `requires` keyword and the post-conditions with the `ensures` keyword. After these keywords a textual description has to be included about the condition being checked. This description is included in the error messages on the violation of the conditions. Invariants are not yet supported but they are subject to further investigation. The current instance of the service can be accessed through the `this` keyword and the `result` keyword represents the return value of the current operation. In the endpoint declaration the name of the contract must be specified with the `contract` keyword. Here is an example of the contract extension for SOAL applied to the `IStack` interface defined in the previous section:

```
contract StackContract : IStack {
  void Push(int number) {
    ensures "stack is not empty"
    { !this.IsEmpty(); }
    ensures "top equals to number"
    { this.Top() == number; }
  }
  int Pop() {
    requires "stack is not empty"
    { !this.IsEmpty(); }
    ensures "result is the old top element"
    { result == old(this).Top(); }
  }
  int Top() {
    requires "stack is not empty"
    { !this.IsEmpty(); }
  }
  bool IsEmpty() { }
}
endpoint GuardedStack : IStack {
  binding Soap11HttpBinding;
  contract StackContract;
  location "http://localhost/GuardedStack";
}
```

The operations may change the state of a stateful web service. In this case the method calls on `this` may return different values after the execution of the current operation than before. The `old` expression can be used to access the state prior to the execution of current operation. The input parameters of the operations are read-only, therefore there is no need to use `old` on them. The expressions in the `requires` and `ensures` clauses have the same syntax as the expressions in C#. In .NET 3.0 the API introduced the `System.Linq.Expressions` namespace with classes that can be used to construct expression trees in memory. Our extension to SOAL and the metamodel behind it is based on these expression tree nodes and supports the following node types (including lambda expressions) [18]: `Add`, `And`, `AndAlso`, `ArrayLength`, `ArrayIndex`, `Call`, `Coalesce`, `Conditional`, `Constant`, `Convert`, `Default`, `Divide`, `Equal`, `ExclusiveOr`, `GreaterThan`, `GreaterThanOrEqual`, `Lambda`, `LeftShift`, `LessThan`, `LessThanOrEqual`, `MemberAccess`, `MemberInit`, `Modulo`, `Multiply`, `Negate`, `UnaryPlus`, `New`, `NewArrayInit`, `NewArrayBounds`, `Not`, `NotEqual`, `Or`, `OrElse`, `OnesComplement`, `Parameter`, `RightShift`, `Subtract`, `TypeAs`, `TypeIs`, `Variable`.

The semantics of these expressions are defined by the C# language specification [19].

SOAL also supports array types. The .NET 3.0 framework allows arrays to be queried through the LINQ API, which provides a lot of useful query functions. Our extension to SOAL supports all of these [20]: `Aggregate`, `All<TSource>`, `Any`, `AsEnumerable<TSource>`, `Average`, `Cast<TResult>`, `Concat<TSource>`, `Contains`, `Count`, `DefaultIfEmpty`, `Distinct`, `ElementAt<TSource>`, `ElementAtOrDefault<TSource>`, `Empty<TResult>`, `Except`, `First`, `FirstOrDefault`, `GroupBy`, `GroupJoin`, `Intersect`, `Join`, `Last`, `LastOrDefault`, `LongCount`, `Max`, `Min`, `OfType<TResult>`, `OrderBy`, `OrderByDescending`, `Range`, `Repeat<TResult>`, `Reverse<TSource>`, `Select`, `SelectMany`, `SequenceEqual`, `Single`, `SingleOrDefault`, `Skip<TSource>`, `SkipWhile`, `Sum`, `Take<TSource>`, `TakeWhile`, `ThenBy`, `ThenByDescending`, `ToArray<TSource>`, `ToDictionary`, `ToList<TSource>`, `ToLookup`, `Union`, `Where`.

All of the OCL expressions are covered by the expressions above, except for the ones dealing with messages, object states and associations, which themselves are specific to UML. Since OCL has proved itself to be powerful enough in practice and its UML independent part is a subset of the LINQ expressions, it can be stated that our extensions to SOAL will suffice in most practical cases.

C. Architecture

The grammar for the contract extension in SOAL was implemented in the M language designed by Microsoft. The M language is a declarative language for working with data and building domain models. It is part of the SQL Server Modeling Services [21] (formerly Oslo) framework, which

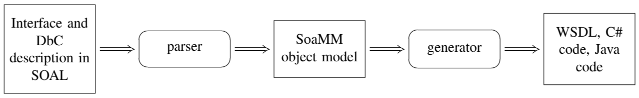


Figure 1. The architecture of the platform independent DbC for SOAL framework

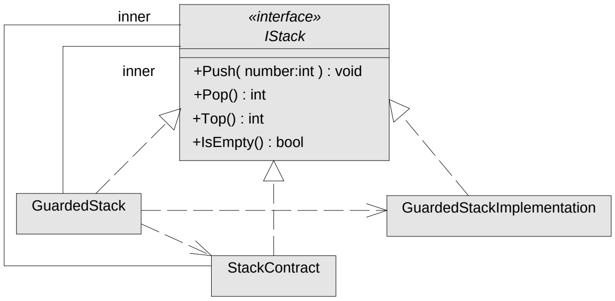


Figure 2. The design pattern of the generated C#/Java classes demonstrated on the stack example

also includes an editor called IntelliPad for domain specific languages and a repository for storing data models.

Based on the grammar the M language parser can process contract descriptions in the SOAL form and can transform the input into an object model described by the SoaMM metamodel. The object model is then easily processable from any .NET language. From this object model the framework generates directly importable projects for different SOA products. The projects include web services, which check the specified pre- and post-conditions automatically; the programmer only has to insert the implementation of the services. The code generator is written in the Text Template Transformation Toolkit for Visual Studio.

The architecture of the platform independent SOAL for WSDL framework can be seen in Figure 1.

D. Generated code

Our framework generates code for two popular web service technologies: Windows Communication Foundation (WCF [22]) and Java API for XML-Based Web Services 2.0 (JAX-WS [23]). The generated classes contain only standard elements, hence they can be used in any SOA product implementing these standards (e.g., Microsoft .NET, Oracle SOA Suite, IBM WebSphere, Apache CXF).

Figure 2 shows the design pattern of the target code regardless of the target platform (i.e., C# for WCF or Java for JAX-WS). From the `IStack` interface in SOAL an interface with the same name is generated in the target language. The contract `StackContract` is mapped to a class with the same name. The endpoint `GuardedStack` is transformed into two classes. The first one is the class `GuardedStack`, which is the web sevice endpoint published by the SOA products. The second one is the class `GuardedStackImplementation`, which contains the implementation of the service. Its operations must be filled by the programmer. The `GuardedStack` class uses the delegate design pattern to check the implementation by the specified contract using the following chain of calls: `GuardedStack` → `StackContract` → `GuardedStackImplementation`.

The following C# interface annotated with WCF attributes is generated from the `IStack` interface:

```
[ServiceContract(...)]
public interface IStack {
  [OperationContract(...)]
  void Push(int number);
  [OperationContract(...)]
  int Pop();
  [OperationContract(...)]
  int Top();
  [OperationContract(...)]
  bool IsEmpty();
}
```

The generated Java interface is similar, but of course it uses JAX-WS annotations, i.e., `@WebService` instead of `[ServiceContract]` and `@WebMethod` instead of `[OperationContract]`.

From the `StackContract` contract specification the following C# class is produced (the attribute named `inner` will contain the implementation of the service and every call is delegated to this object):

```
public class StackContract : IStack {
  private IStack inner;
  public StackContract(IStack inner) {
    this.inner = inner;
  }
  public void Push(int number) {
    this.inner.Push(number);
    if (!(!this.IsEmpty())) {
      throw new PostConditionViolationException(
                 "stack is not empty");
    }
    if (!(this.Top() == number)) {
      throw new PostConditionViolationException(
                 "top equals to number");
    }
  }
  public int Pop() {
    if (!(!this.IsEmpty())) {
      throw new PreConditionViolationException(
                 "stack is not empty");
    }
    int temp1 = this.Top();
    int result = this.inner.Pop();
    if (!(result == temp1)) {
      throw new PostConditionViolationException(
                 "result is the old top element");
    }
    return result;
  }
  public int Top() {
    if (!(!this.IsEmpty())) {
      throw new PreConditionViolationException(
                 "stack is not empty");
    }
    int result = this.inner.Top();
    return result;
  }
  public bool IsEmpty() {
    bool result = this.inner.IsEmpty();
    return result;
  }
}
```

As it can be seen, the `old` expressions are evaluated into temporary variables before the call is delegated to the implementation. After the execution of the implementation the post-conditions are checked correctly. The Java version of this class is similar. The expressions specified in Section III-B are translated to Java as well: the LINQ Standard Query Operators are backed up by a custom utility class, the lambda expressions are transformed into anonymous classes.

The `GuardedStackImplementation` class:

```
public class GuardedStackImplementation : IStack {
  public void Push(int number) {
    throw new NotImplementedException();
  }
  public int Pop() {
    throw new NotImplementedException();
  }
  public int Top() {
    throw new NotImplementedException();
  }
  public bool IsEmpty() {
    throw new NotImplementedException();
  }
}
```

This code is very clean, since the pre- and post-conditions are generated into the `StackContract` class. The programmer has to fill in the missing implementations in order to have a functioning web service.

The service endpoint class `GuardedStack` has to be published. It is also very simple; it builds up the delegation chain and delegates the calls to the other classes:

```
public class GuardedStack : IStack {
  private IStack inner;
  public GuardedStack() {
    this.inner =
      new StackContract(
        new GuardedStackImplementation());
  }
  public void Push(int number) {
    this.inner.Push(number);
  }
  public int Pop() {
    return this.inner.Pop();
  }
  public int Top() {
    return this.inner.Top();
  }
  public bool IsEmpty() {
    return this.inner.IsEmpty();
  }
}
```

The Java versions of the generated classes are similar to the C# examples above.

E. Evaluation

SOAL is a human readable domain specific language for SOA with a metamodel called SoaMM behind it. It provides much cleaner syntax than an XML based WSDL document. From a SOAL description a SoaMM object model can be constructed and from this object model WSDL files and program code can be automatically generated. This is a powerful tool in the top-down development process where the task is to create interoperable web services based on WSDLs, while it is as simple as the less interoperable bottom-up development process primarily supported by SOA products. The C#-Java-like textual syntax of SOAL is easier to maintain than the XML or graphical WSDL representations provided by the products.

Although Design by Contract is a key concept in SOA, the major software vendors do not provide any tools to make this task easier. In this paper we proposed a contract extension to SOAL. This extension offers the same maintainability and readability as the original version of the language. Pre- and post-conditions can be specified for each operation. These conditions are then woven as aspects into the generated code, while the programmer has only a single task: provide the implementation of the web service by filling in the methods of the implementation class. Every other configuration and source files for the SOA products are automatically produced by our framework. Currently two SOA products are supported: Microsoft Visual Studio for WCF and GlassFish ESB for JAX-WS. However, with the appropriate configuration files the Java code generated by our framework can be directly used in other products (e.g., Oracle SOA Suite and IBM WebSphere) as well.

Our proposition offers an extensive set of operators that can be utilized in pre- and post-condition expressions. These operators do not introduce any new concepts, therefore they are easy to learn. They also have a well defined semantics based on the C# language specification. The expressions can also contain lambda functions and thus the benefits of the LINQ standard query operators can also be harvested. These query operators build a superset of OCL (without the UML specific parts), which itself is also a powerful constraint description language. Although the expression syntax in SOAL is the same as in C#, the expressions are translated by our framework to Java as well.

The design pattern generated by the framework separates the interface, the publication part, the contract validation part and the implementation part of a service. This results in a clean, easily maintainable code. Although the delegation of method calls in this design pattern introduces a minor overhead, this is negligible compared to the time consumed by transforming the incoming and outgoing SOAP XML messages.

There is one weak point of the framework: the expressions are too powerful to be translated into BPEL, therefore BPEL code cannot be produced from them. However, a proxy web service can be generated that delegates calls towards the BPEL process instead of towards a local implementation class. This solution introduces a major overhead though, since the SOAP XML messages have to be serialized and deserialized twice, instead of once. Nevertheless, this proxy web service concept is useful for other purposes as well, e.g., testing other web services.

IV. Conclusion

We proposed a Design by Contract solution for SOA. Our solution is based on a simple, human readable domain specific language called SOAL. Our framework generates C# or Java code from this description, that can be directly imported into the SOA products of the major software vendors. The generated artifacts follow the delegation design pattern, which results in a clean and easily maintainable code. The syntax and semantics of the condition expressions are based on C# and the LINQ standard query operators are also fully supported. Design by Contract can also be applied to BPEL though proxy web services at the cost of some performance loss.

In our future work, we will extend the framework with other concepts, e.g., invariant conditions and versioning.

Acknowledgment

This work is connected to the scientific program of the ”Development of quality-oriented and harmonized R+D+I strategy and functional model at BME” project. This project is supported by the New Hungary Development Plan (Project ID: TMOP-4.2.1/B-09/1/KMR-2010-0002).

References

[1] OASIS, *SOA Reference Model*, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=soa-rm, accessed: 8.12.2010.

[2] ——, *WS-* Standards*, http://www.oasis-open.org/specs/, accessed: 8.12.2010.

[3] B. Meyer, “Applying ”design by contract”,” *Computer*, vol. 25, no. 10, pp. 40–51, 1992.

[4] ——, *Touch of Class: Learning to Program Well with Objects and Contracts*. Springer, 2009.

[5] A. G. D. Florescu and D. Kossmann, “Xl: an xml programming language for web service specification and composition,” in *Computer Networks: The International Journal of Computer and Telecommunications Networking*, 2003, pp. 1–25.

[6] L. Baresi, S. Guinea, M. Pistore, and M. Trainotti, “Dynamo + astro: An integrated approach for bpel monitoring,” *Web Services, IEEE International Conference on*, vol. 0, pp. 230–237, 2009.

[7] L. Baresi, D. Bianculli, C. Ghezzi, S. Guinea, and P. Spoletini, “A timed extension of wscol,” *Web Services, IEEE International Conference on*, vol. 0, pp. 663–670, 2007.

[8] M. Klusch, *CASCOM - Intelligent Service Coordination in the Semantic Web*. Birkhuser Verlag, Springer, 2008, ch. 3.

[9] W3C, *Semantic Annotations for WSDL and XML Schema (SAWSDL)*, http://www.w3.org/TR/sawsdl/, accessed: 8.12.2010.

[10] ——, *OWL-S: Semantic Markup for Web Services*, http://www.w3.org/Submission/OWL-S/, accessed: 8.12.2010.

[11] E. W. working group, *Web Service Modeling Ontology (WSMO)*, http://www.wsmo.org/, accessed: 8.12.2010.

[12] A. Haller, E. Cimpian, A. Mocan, E. Oren, and C. Bussler, “Wsmx - a semantic service-oriented architecture,” in *In Proceedings of the International Conference on Web Service (ICWS 2005*, 2005, pp. 321–328.

[13] R. Heckel and M. Lohmann, “Towards contract-based testing of web services,” *Electronic Notes in Theoretical Computer Science*, vol. 82, p. 2003, 2004.

[14] J. T. E. Timm, “Specifying semantic web service compositions using uml and ocl,” in *In 5th International Conference on Web Services*. IEEE press, 2007.

[15] B. Simon, Z. Laszlo, B. Goldschmidt, K. Kondorosi, and P. Risztics, “Evaluation of ws-* standards based interoperability of soa products for the hungarian e-government infrastructure,” in *International Conference on the Digital Society*. Los Alamitos, CA, USA: IEEE Computer Society, 2010, pp. 118–123.

[16] B. Simon and B. Goldschmidt, “A human readable platform independent domain specific language for wsdl,” in *Networked Digital Technologies*, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2010, vol. 87, pp. 529–536.

[17] B. Simon, B. Goldschmidt, and K. Kondorosi, “A human readable platform independent domain specific language for bpel,” in *Networked Digital Technologies*, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2010, vol. 87, pp. 537–544.

[18] Microsoft, *.NET LINQ ExpressionType Enumeration*, http://msdn.microsoft.com/en-us/library/bb361179.aspx, accessed: 8.12.2010.

[19] ——, *C# Language Specification Version 4.0*, http://www.microsoft.com/downloads/en/details.aspx?familyid=DFBF523C-F98C-4804-AFBD-459E846B268E&displaylang=en, accessed: 8.12.2010.

[20] ——, *.NET LINQ Standard Query Operators*, http://msdn.microsoft.com/en-us/library/bb882641.aspx, accessed: 8.12.2010.

[21] ——, *SQL Server Modeling Services*, http://msdn.microsoft.com/en-us/data/ff394760.aspx, accessed: 8.12.2010.

[22] ——, *Windows Communication Foundation*, http://msdn.microsoft.com/en-us/netframework/aa663324.aspx, accessed: 8.12.2010.

[23] Oracle, *JSR 224: Java API for XML-Based Web Services (JAX-WS) 2.0*, http://jcp.org/en/jsr/detail?id=224, accessed: 8.12.2010.

A Metamodel of the WS-Policy Standard Family

Balazs Simon, Balazs Goldschmidt, Peter Budai, Istvan Hartung, Karoly Kondorosi, Zoltan Laszlo, Peter Risztics
Department of Control Engineering and Information Technology
Budapest University of Technology and Economics
Budapest, Hungary
Email: {sbalazs | balage | bucnak | hartungi | kondor | laszlo}@iit.bme.hu, risztics@ik.bme.hu

***Abstract*—The basic building blocks of SOA systems are web services. The domain specific language SOAL developed by the authors has a Java and C#-like syntax for describing web service interfaces. Beside the syntax a metamodel (SoaMM) is also defined. The paper introduces an extended version of both SOAL and SoaMM that supports WS-Policy specifications. The original WS-Policy standards specify huge XML descriptions that are too complex and low level for efficient service design. The metamodel presented provides a high abstraction level that is still strong enough for generating vendor-specific service configurations for WCF and major JAX-WS implementations.**

***Keywords*-SOA; web services; WS-policy; modelling; DSL.**

I. Introduction

Complex distributed systems are best built from components with well-defined interfaces and a framework that helps connecting them. Web services and BPEL (Business Process Execution Language) processes implementing WSDL (Web Services Description Language) interfaces represent today the components that both enterprises and governments use to construct their complex distributed systems, implementing a Service Oriented Architecture (SOA) [1].

One of the advantages of building on web services technology is that one can get a vast set of standards from simple connections to middleware functionalities such as security or transaction-handling [2].

The other advantage is that SOA applies classical principles of software engineering, like model-based development, loosely coupled components, separation of interface from implementation, etc., at a higher level of abstraction. When implementing the abstract model, vendor specific issues start to dominate. In many situations, like systems in e-government or those of interoperating companies, such vendor-dependent issues undermine successfull systems integration. Without clearly defined, common configuration settings interoperability might be compromized. Therefore it is of utmost importance to have a common model of configuration descriptions that enables vendor independent specification of policies. This was the essential aim of the WS-Policy standards. Unfortunately these standards not only require huge, unreadable, and thus unmaintainable XML configuration descriptions, but in the actual products the descriptions usually still contain vendor specific details.

This paper shows how a high-level web service and BPEL metamodel, description language, and code generating framework was extended with WS-Policy standards support in order to enable developers to automatically generate platform-specific policy configurations from the general, standard-compliant policy specifications. The paper emphasizes the use of a clear and easily extensible metamodel framework that is the *sine-qua-non* of a really useable and effective solution.

The activity of our research group aims at developing principles, recommendations, components and technologies that make the application of SOA in the e-government domain easier. Considering the diversity of legacy systems of governmental organizations, all the solutions we elaborate should be platform-independent as far as it is possible.

The rest of the paper is organized as follows. In the second section, related work is examined. In the third section, the metamodel is introduced, the WS-policy representation is detailed, the code generating framework architecture is shown, and finally, an evaluation is provided. The fourth section summarizes the results.

II. Related Work

One way of describing services is using semantic web technologies. The major goal of Semantic Web Services (SWS) is to create intelligent software agents to provide automated, interoperable and meaningful coordination of web services [3]. The three main directions of SWS are SAWSDL (Semantic Annotations for WSDL), OWL-S (Semantic Markup for Web Services) and WSMO (Web Services Modeling Ontology).

SAWSDL [4] does not introduce a new language. It is a WSDL extension for referencing ontological concepts outside WSDL documents. Beyond that it does not define any execution semantics for the implementation.

The OWL-S [5] profile ontology is used to describe what a service does, and is meant to be mainly used for the purpose of service discovery. The service description contains input and output parameters, pre- and post-conditions, and also non-functional aspects. The OWL-S process model describes service composition including the communication pattern. In order to connect OWL-S to existing web service standards, OWL-S uses grounding to map service descriptions to WSDL. The OWL-S environment provides an editor to develop semantic web services and a matcher to discover services. The OWL-S Virtual Machine is a general purpose web service client for the invocation. OWL-S therefore requires a custom execution environment and cannot be used in current commercial SOA products. Its underlying description logic OWL-DL has also a limited expressiveness in practice.

The WSMO [6] framework provides a conceptual model and a formal language WSML for semantic markup of web services. WSMO is used for modeling of ontologies, goals, web services and mediators. Ontologies provide formal logic-based grounding of information used by other components. Goals represent user desires, i.e., the objectives that a client might have when searching for services. Web services are computational entities, their semantic description includes functional and non-functional properties, as well as their capabilities through pre- and postconditions, assumptions and effects. Mediators provide interoperability between components at data, protocol and process level. The reference implementation of WSMO is the WSMX [7] framework, a custom execution environment. It is designed to allow dynamic discovery, invocation and composition of web services. It also provides interoperability with classical web services.

The main design goals of SWS standards are discovery, invocation and composition of web services. These standards are not primarily designed for modelling purposes. They are weak in terms of security, transactional, reliability and other non-functional aspects even at the conceptual level [8]. Because of their custom execution environment, their implementations do not rely on existing SOA products of major software vendors, which can result in interoperability problems with classical web services published by these products.

Although there are directions to extend SWS standards with WS-Policy concepts [9] [10] [11], these solutions focus on service discovery and policy matching, and do not resolve the issues related to modelling and implementation.

Most BPEL workflow engines in the industry also lack support of the specification and enforcement of non-functional requirements. A. Charfi et.al. [12] proposed and implemented a container framework to include the most important WS-* standards regarding security, reliability and transaction handling in BPEL processes. However, their solution can only be used in the ActiveBPEL engine, and cannot be applied to other industry engines.

Another major drawback of the WS-Policy standard family is that the policy assertions of the WS-* standards can be very large XML structures, which makes them nearly impossible to be handwritten by humans. Luckily, most SOA products provide policy repositories (e.g., Oracle SOA Suite, IBM WebSphere) containing complete assertions that can be used for configuration. However, these assertions may differ between products and matching them can be a difficult task. There are also products, which offer GUI designers (e.g., GlassFish ESB) or transform WS-Policy assertions into their own configuration representation (e.g., Apache CXF, Microsoft Windows Communication Foundation (WCF)) making interoperability problems even harder.

Others focus their research on creating platform independent languages for describing WS-Policy assertions. These efforts come together under the umbrella of XACML [13] (eXtensible Access Control Markup Language). XACML is a declarative access control policy language implemented in XML and a processing model, describing how to interpret the policies. WSPL [14] [15] (Web Services Policy Language) is a subset of XACML and is designed for matching policy descriptions. WS-PolicyConstraints [16] is an even smaller subset of WSPL with the parts of WSPL that overlapped and conflicted with WS-Policy and WS-PolicyAttachment removed. Although these XACML-based solutions are platform independent, they are too complex and have unfriendly XML syntax. Their current tool support is also very poor and they cannot be used for modelling.

WS-Policy and XACML provide expressions to specify different configurations of policy assertions, however, most SOA products support only a single configuration option per endpoint. Different configurations are published on different endpoints. Hence, there is usually no need to dynamically choose between options; to be able to specify the exact same configuration on the client and server side between different products is a more important task in practice.

III. Meta-model for the WS-Policy standards

This section introduces the high level description language and the metamodel for WS-Policy standards.

A. SOAL and SoaMM

In the SOA world, XML is used for interface and process-description, and message-formatting. The aim is interoperability, but the drawback is that handling and transforming XML documents above a certain complexity is almost impossible. This is why most development environments have graphical tools for helping developers creating interface-descriptions, connections, process-flows and message-formats. The problem with the graphical approach is that it is neither efficient, nor reliably repeatable, nor sufficiently controllable, nor easily automatable.

On the other hand, the standards usually have a lot of redundant parts that result in poor readability and manageability. For example the `message` element in WSDL 1.x was omitted from WSDL 2.0 because of its redundancy. The development tools do not support the inclusion of special extensions in the interface specification (like pre- and post-conditions, authorization, etc.) Even some tools have special naming conventions that are to be accepted, otherwise the generated code is even less readable than necessary. We have examined a lot of products [17] and have found a lot of peculiarities, which have to be taken into account beyond the recommendations of the WS-I Basic Profile. In BPEL process-descriptions the `partnerLinkType-partnerLink` constructs for partners, or the `property-propertyAlias-correlationSet` constructs for correlations mean unnecessary redundancy in BPEL. Unfortunately, the process designer tools usually map these constructs directly to the graphical interface instead of hiding them from the users. The development tools do not support the inclusion of special extensions in the process description (like pre- and post-conditions, authorization, etc.)

To solve the above problems, an abstract model (SoaMM) has been developed that can manage both BPEL and WSDL concepts, and, in order to describe the model, an extendable language called Service Oriented Architecture Language (SOAL) was specified [18], [19] that can be used for describing webservice interfaces and BPEL processes, and is more easily readable and manageable by humans. This model and language also enables automatization, vendor-specific WSDL and BPEL generation, and compile time type checking.

The following example illustrates a simple stack web service description in SOAL. The service has the URL http://localhost/Calculator and can be accessed through SOAP 1.1 over HTTP:

```
namespace CalculatorSample {
  interface ICalculator {
    double Add(double left, double right);
    double Subtract(double left, double right);
    double Multiply(double left, double right);
    double Divide(double left, double right);
  }
  binding Soap11HttpBinding {
    transport HTTP;
    encoding SOAP { Version = SoapVersion.Soap11 }
  }
  endpoint Calculator : ICalculator {
    binding Soap11HttpBinding;
    location "http://localhost/Calculator";
  }
}
```

B. WS-Policy in SoaMM and SOAL

The most widely supported WS-* protocols by the industry are WS-Addressing, WS-ReliableMessaging, WS-SecureConversation and WS-AtomicTransaction. Each of them has a respective WS-Policy standard, which defines assertions that can be used to describe a set of parameters in a platform independent way to configure these protocols.

SOA products implement these protocols as web services stacks. These stacks consist of various layers, and the products offer different kinds of configuration mechanisms to set the parameters of these layers. We have reviewed all the WS-Policy standards corresponding to the protocols enumerated at the beginning of this section, and we have also examined the configuration mechanisms of the most popular SOA products. Oracle SOA Suite and IBM WebSphere directly use WS-Policy assertions, which can be selected from a repository. GlassFish ESB, using the Metro web services stack, also consumes WS-Policy assertions, however, it offers a graphical user interface built into Netbeans in order to make the settings easier. Apache CXF and Microsoft WCF transform WS-Policy assertions into their own configuration representation. These differencies between products make it very hard to achieve interoperability when the number of protocols in the stack increases, since finding the exact same options between the various representations can be error prone. It is not a coincidence, that GlassFish ESB being a Java implementation offers configurations labeled as .NET interoperable in its GUI settings.

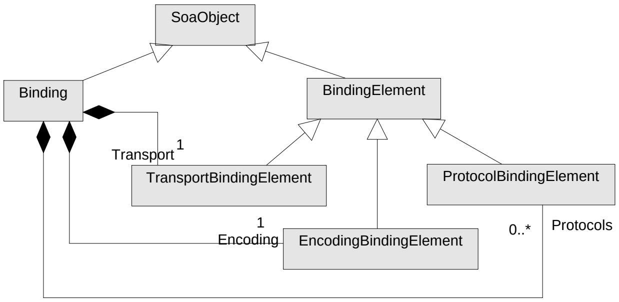


Figure 1. Bindings and binding elements in SoaMM

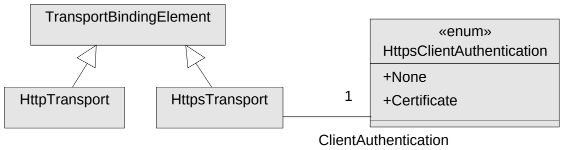


Figure 2. Transport binding elements in SoaMM

After reviewing all of these configuration representations we have created a platform independent metamodel as an extension of SoaMM in order to be able to describe all the parameters of the various WS-* protocols and to be able to generate directly interoperable configurations for the individual SOA products.

Figure 1 shows the basic building blocks of the policy metamodel. *SoaObject* is the root of the class hierarchy in SoaMM. *Binding* contains a set of protocols represented by *BindingElement*s. *TransportBindingElement*, *EncodingBindingElement* and *ProtocolBindingElement* denote transport protocols (e.g., HTTP, UDP, JMS, etc.), encoding protocols (e.g., SOAP, binary, etc.) and higher level protocols (e.g., WS-SecureConversation, etc.) respectively.

Figure 2 contains the two transport protocols currently supported by our metamodel: *HttpTransport* for HTTP and *HttpsTransport* for HTTPS. Through the *ClientAuthentication* property of type *HttpsClientAuthentication* the latter

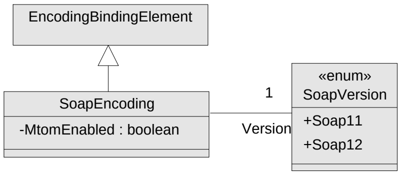


Figure 3. Encoding binding elements in SoaMM

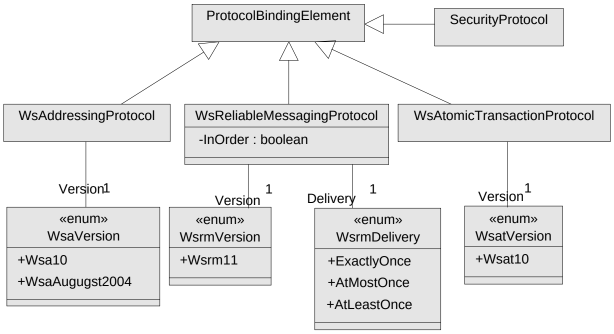


Figure 4. Protocol binding elements in SoaMM

can prescribe that the client must authenticate itself with an X.509. certificate.

Web services use the SOAP encoding protocol for communication. Our metamodel offers *SoapEncoding* (see Figure 3) to represent this protocol and also allows the MTOM option to be enabled for increased efficiency in the communication. The SOAP version can be specified through the *Version* property.

Figure 4 introduces the higher level protocols. *WsAddressingProtocol*, *WsReliableMessagingProtocol* and *WsAtomicTransactionProtocol* denote the settings of WS-Addressing, WS-ReliableMessaging and WS-AtomicTransaction respectively. Each of these has a *Version* property for the version number. The *Delivery* property of *WsReliableMessagingProtocol* defines what kind of delivery should be used in the WS-ReliableMessaging protocol while the *InOrder* property indicates whether the order of the messages should be preserved. The security protocols are omitted from this paper because of space limitations, however, in the figure they are marked by the *SecurityProtocol* class.

We have also extended the SOAL language so that the various protocol settings can be represented in the language. The main extension point is the binding declaration. For example the following binding can be used to configure WS-Addressing 1.0 and WS-ReliableMessaging 1.1 with messages delivered exactly once over HTTPS:

```
binding ReliableBinding {
  transport HTTPS {
   ClientAuthentication =
     HttpsClientAuthentication.None
  }
  encoding SOAP {
    Version = SoapVersion.Soap11
  }
  protocol WsAddressing {
    Version = WsaVersion.Wsa10
  }
  protocol WsReliableMessaging {
    Version = WsrmVersion.Wsrm11,
    Delivery = WsrmDelivery.ExactlyOnce,
    InOrder = true
  }
}
```

The corresponding WS-Policy assertions used by most SOA products would be:

```
<wsp:Policy wsu:Id="ReliableBinding_Policy">
  <sp:TransportBinding>
    <wsp:Policy>
      <sp:TransportToken>
        <wsp:Policy>
          <sp:HttpsToken
             RequireClientCertificate="false"/>
        </wsp:Policy>
      </sp:TransportToken>
      <sp:AlgorithmSuite>
        <wsp:Policy>
          <Basic256/>
        </wsp:Policy>
      </sp:AlgorithmSuite>
      <sp:Layout>
        <wsp:Policy>
          <sp:Strict/>
        </wsp:Policy>
      </sp:Layout>
      <sp:IncludeTimestamp/>
    </wsp:Policy>
  </sp:TransportBinding>
  <wsam:Addressing/>
  <wsrmp:RMAssertion>
    <wsp:Policy>
      <wsrmp:DeliveryAssurance>
        <wsp:Policy>
          <wsrmp:ExactlyOnce/>
          <wsrmp:InOrder/>
        </wsp:Policy>
      </wsrmp:DeliveryAssurance>
    </wsp:Policy>
  </wsrmp:RMAssertion>
</wsp:Policy>
```

The WCF configuration would be the following:

```
<binding name="ReliableBinding">
  <reliableSession reliableMessagingVersion=
                   "WSReliableMessaging11"
     ordered="true" />
  <textMessageEncoding
     messageVersion="Soap11WSAddressing10" />
  <httpsTransport
     requireClientCertificate="false" />
</binding>
```

As it can be seen, the SOAL description is much cleaner than the bloated XML-based WS-Policy assertions and is as compact as the custom configuration representation of WCF.

C. Architecture

The grammar for the WS-Policy extension in SOAL was implemented in the M language designed by Microsoft. The M language is a declarative language for working with data and building domain models. It is part of the SQL Server

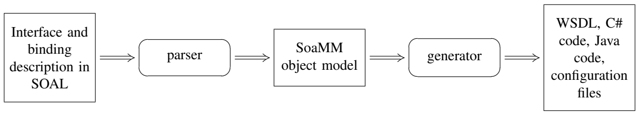

Interface and binding description in SOAL
parser
SoaMM object model
generator
WSDL, C# code, Java code, configuration files

Figure 5. The architecture of the platform independent SOAL framework

Modeling Services [20] (formerly Oslo) framework, which also includes an editor called IntelliPad for domain specific languages and a repository for storing data models.

Based on the grammar the M language parser can process binding descriptions in the SOAL form and can transform the input into an object model described by the SoaMM metamodel. The object model is then easily processable from any .NET language. From this object model the framework generates directly importable projects for different SOA products. The projects include web services with directly interoperable configurations even between different products; the programmer only has to specify the implementation of the services. The code generator is written in the Text Template Transformation Toolkit for Visual Studio.

The architecture of the platform independent SOAL for WSDL framework can be seen in Figure 5.

D. Generated code

Our framework generates code for two popular web service technologies: Windows Communication Foundation (WCF [21]) and Java API for XML-Based Web Services 2.0 (JAX-WS [22]). The generated C# and Java classes contain only standard elements, hence they can be used in any SOA product implementing these standards (e.g., Microsoft .NET, Oracle SOA Suite, IBM WebSphere, Apache CXF).

However, JAX-WS (unlike WCF) does not cover WS-Policy standards for non-functional requirements, therefore, configuration files for SOA products may differ between JAX-WS implementations. Our framework currently supports WCF and the following JAX-WS implementation stacks: Metro (used in GlassFish ESB), IBM WebSphere and Oracle SOA Suite. The open-source Apache CXF (also used in the JBossWS stack) is planned to be supported.

Table I shows the number of lines required in configuration files of the various web services stacks regarding the different protocols. It can be seen that SOAL is as compact as WCF, while the other configuration representations are more verbose. Therefore, generating these configuration files, so that they are even directly interoperable, results in great increase of productivity.

E. Evaluation

SOAL is a human readable domain specific language for SOA with a metamodel called SoaMM behind it. It provides much cleaner syntax than an XML based WSDL document. From a SOAL description a SoaMM object model can be constructed and from this object model WSDL files, program code and configuration files can be automatically generated.

Table I
NUMBER OF LINES OF THE VARIOUS CONFIGURATIONS IN SOA PRODUCTS

Legend: WS-A=WS-Addressing, WS-RM=WS-ReliableMessaging, WS-AT=WS-AtomicTransaction, WS-SC=WS-SecureConversation

Protocol	SOAL	WCF	Metro/IBM/Oracle	CXF/JBoss
HTTP	1	1	1	1
HTTPS	2	1	20	29
SOAP	2	1	1	1
MTOM	2	1	4	14
WS-A	2	1	3	5
WS-RM	4	1	10	11
WS-AT	2	1	10	-
WS-SC	4	5	40	40

This is a powerful tool in the top-down development process where the task is to create interoperable web services based on WSDLs, while it is as simple as the less interoperable bottom-up development process primarily supported by SOA products. The C#-Java-like textual syntax of SOAL is easier to maintain than XML or graphical WSDL representations provided by the products.

Furthermore, the authors are not aware of any other framework or metamodel that has such a high abstraction level policy representation and strong descriptive power at the same time as SoaMM and SOAL. The metamodel and the language also has advantage over the web services standard family: the descriptions are unified in a single language on all three (transport, encoding, protocol) configuration layers.

The flexibility and extensibility of the metamodel are demonstrated by the integration of standards specifying the most important non-functional aspects, like WS-Addressing, WS-ReliableMessaging, WS-Security, WS-SecureConversation, WS-AtomicTransaction. The metamodel can be easily extended at any time by further solutions for transport, encoding, or protocol.

The most important aspect of the metamodel extension is that configuration profiles can be specified on a high conceptual level. The actual implementation-specific configuration files are automatically generated from these high level descriptions. Interoperability, which is of utmost importance in systems integration, is thus guaranteed without further vendor-specific, low-level product configuration.

There is one weak point of the framework: current BPEL engines usually do not support advanced non-functional requirements (e.g., security, transactions), hence, configuration cannot be generated for them. However, a proxy web service can be produced that delegates calls towards the BPEL process. This solution introduces a major overhead though, since the SOAP XML messages have to be serialized and deserialized twice, instead of once. Nevertheless, this proxy web service concept is useful for other purposes as well, e.g., testing or adapting other web services.

IV. Conclusion

The original WS-Policy standards specify huge XML descriptions that are too complex and too low level for efficient service design. The paper presented an extension to the metamodel SoaMM and description language SOAL that was originally designed as a simple, human readable domain specific language specifically for webservices and BPEL. This extension provides a high abstraction level that is still strong enough for generating vendor-specific policy configurations. Furthermore, this description guarantees interoperability of SOA products of different vendors. For BPEL engines that do not support advanced non-functional requirements, a proxy web service can be produced that delegates calls towards the BPEL process.

In our future work, we will extend the framework with other concepts, e.g., versioning and SAML authentication.

Acknowledgment

This work is connected to the scientific program of the "Development of quality-oriented and harmonized R+D+I strategy and functional model at BME" project. This project is supported by the New Hungary Development Plan (Project ID: TMOP-4.2.1/B-09/1/KMR-2010-0002).

References

[1] OASIS, *SOA Reference Model*, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=soa-rm, accessed: 8.12.2010.

[2] ——, *WS-* Standards*, http://www.oasis-open.org/specs/, accessed: 8.12.2010.

[3] M. Klusch, *CASCOM - Intelligent Service Coordination in the Semantic Web*. Birkhuser Verlag, Springer, 2008, ch. 3.

[4] W3C, *Semantic Annotations for WSDL and XML Schema (SAWSDL)*, http://www.w3.org/TR/sawsdl/, accessed: 8.12.2010.

[5] ——, *OWL-S: Semantic Markup for Web Services*, http://www.w3.org/Submission/OWL-S/, accessed: 8.12.2010.

[6] E. W. working group, *Web Service Modeling Ontology (WSMO)*, http://www.wsmo.org/, accessed: 8.12.2010.

[7] A. Haller, E. Cimpian, A. Mocan, E. Oren, and C. Bussler, "Wsmx - a semantic service-oriented architecture," in *In Proceedings of the International Conference on Web Service (ICWS 2005*, 2005, pp. 321–328.

[8] O. Shafiq, M. Moran, E. Cimpian, A. Mocan, M. Zaremba, and D. Fensel, "Investigating semantic web service execution environments: A comparison between wsmx and owl-s tools," *Internet and Web Applications and Services, International Conference on*, vol. 0, p. 31, 2007.

[9] V. Kolovski, B. Parsia, Y. Katz, and J. Hendler, "Representing web service policies in owl-dl," in *In International Semantic Web Conference (ISWC*, 2005, pp. 6–10.

[10] K. Verma, R. Akkiraju, and R. Goodwin, "R.: Semantic matching of web service policies," in *Proceedings of the Second Workshop on SDWP, 2005*, 2005, pp. 79–90.

[11] N. Sriharee, T. Senivongse, K. Verma, and A. Sheth, "On using ws-policy, ontology, and rule reasoning to discover web services," in *Intelligence in Communication Systems*, ser. Lecture Notes in Computer Science, F. A. Aagesen, C. Anutariya, and V. Wuwongse, Eds. Springer Berlin / Heidelberg, 2004, vol. 3283, pp. 246–255, 10.1007/978-3-540-30179-0_22.

[12] A. Charfi, B. Schmeling, A. Heizenreder, and M. Mezini, "Reliable, secure, and transacted web service compositions with ao4bpel," in *ECOWS '06: Proceedings of the European Conference on Web Services*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 23–34.

[13] T. Moses, *eXtensible Access Control Markup Language (XACML)*, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml, accessed: 8.12.2010.

[14] A. Anderson, *An Introduction to the Web Services Policy Language, Fifth IEEE International Workshop on Policies for Distributed Systems and Networks (POLICY'04)*, http://labs.oracle.com/projects/xacml/Policy2004.pdf, accessed: 8.12.2010.

[15] T. Nadalin, *Web Services Security Policy Language (WS-SecurityPolicy)*, http://www-128.ibm.com/developerworks/library/specification/ws-secpol/, accessed: 8.12.2010.

[16] A. Anderson, *XACML-based Web Services Policy Constraint Language (WSPolicyConstraints)*, http://labs.oracle.com/projects/xacml/ws-policy-constraints-current.pdf, accessed: 8.12.2010.

[17] B. Simon, Z. Laszlo, B. Goldschmidt, K. Kondorosi, and P. Risztics, "Evaluation of ws-* standards based interoperability of soa products for the hungarian e-government infrastructure," in *International Conference on the Digital Society*. Los Alamitos, CA, USA: IEEE Computer Society, 2010, pp. 118–123.

[18] B. Simon and B. Goldschmidt, "A human readable platform independent domain specific language for wsdl," in *Networked Digital Technologies*, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2010, vol. 87, pp. 529–536.

[19] B. Simon, B. Goldschmidt, and K. Kondorosi, "A human readable platform independent domain specific language for bpel," in *Networked Digital Technologies*, ser. Communications in Computer and Information Science. Springer Berlin Heidelberg, 2010, vol. 87, pp. 537–544.

[20] Microsoft, *SQL Server Modeling Services*, http://msdn.microsoft.com/en-us/data/ff394760.aspx, accessed: 8.12.2010.

[21] ——, *Windows Communication Foundation*, http://msdn.microsoft.com/en-us/netframework/aa663324.aspx, accessed: 8.12.2010.

[22] Oracle, *JSR 224: Java API for XML-Based Web Services (JAX-WS) 2.0*, http://jcp.org/en/jsr/detail?id=224, accessed: 8.12.2010.

A Comprehensive Strategy Framework for e-Textbook in the Coming Digital Society for Learning

Jen Wel Chen[1,2, a], Dershing Luo[1,3, b],Ching-Cha Hsieh[1, c]
[1]Department of Information Management
National Taiwan University
[2]Department of Business Administration
Chinese Culture University
[3]Department of Information Management,
China University of Science and Technology
Taipei, Taiwan (R.O.C.)
[a] jwchen@sce.pccu.edu.tw
[b] luoder2008@gmail.com
[c] cchsieh@im.ntu.edu.tw

Chia-Ching Lu
Institute for Information Industry
Taipei, Taiwan (R.O.C.)
gaty@iii.org.tw

***Abstract*—The Princeton University held a pilot experiment on the current e-reader using in the classroom. The classroom experience was worsened by using e-reader, unsurprisingly. Meanwhile her final report in 2010 shows some valuable suggestion for future e-book manufacturers. However, the success of a product locates not only on its advanced technique but also on its stakeholders and business model. Take the e-book reader Kindle for example. Its success is largely owing to a specific position of "Reading" and online book sales. While during its marching to campus, it is obviously not valid any more. In this paper, we go through the research and practice of e-books recently as well as reflect the e-reader pilot at Princeton. We then propose a comprehensive strategy framework for e-Textbook that is composed of three dimensions: information goods, technology, and stakeholders. Tracing on the path of e-book reader, we march to e-Textbook. During this journey, we not only highlight solutions to the problems e-Textbooks will encounter, but also propose valuable business models for future e-Textbook.**

Keywords- e-book; e-Textbook; business mode; e-Reader

I. INTRODUCTION

There has been already 42 years for the development of e-books since Alan Kay originally proposed the concept of Dynabook in 1968. Meanwhile, the researchers, enterprises, and governments have developed and promoted the applications of e-books in various concepts, ways, and techniques. Such situation presented the e-books fruitful definitions and facets [4, 8, 10].

There are several different facets for the e-books. For the content side, the e-books may be digitalized books, online database, CD-ROM. For the device side, the e-books may be portable device for storage contents or browsing information. Even simpler, e-books maybe the digital contents as usually recognized. In this way, the development approaches are various as follows: content, composer, container, channel, and consultant [1]. Under different supports of projects, various technique development approaches come across their versatile requirements and user groups.

Same as the above situations, scholars have different definitions in the context and popularity for the e-books. That is the reason why the research issues presented versatile facets and results. In this paper, we adopt a broader definition: the e-books are not only digital contents, but also the portable devices/ carriers, access interfaces, and system for information browsing and integrating.

In 2007, the e-book reader, Kindle promoted by Amazon popularized the global e-books development and digital reading trend. The success is owing to the mechanism of online sale, digital reading techniques and hardware. The trend forced the other big electronic company, SONY, resumed its e-book development project. SONY promoted a new style e-book reader, and transformed its proprietary specification of e-book content into a general specification, e-PUB (short for electronic publication). In 2009, Barnes and Noble promoted e-book reader of its own brand, even she has already given up the sales of digital books long before in 2003. Insta, a market research institute, predicts the global e-book device shipments will raise from nearly one million in 2008 to thirty million, that will be a market scale of eleven billion. The next target for digitalized will be the great population of textbooks. Forrester predicts the textbooks will be the greatest e-book reader market five years later. A recent published report "E-Textbooks in Higher Education" [16], it examines the lessons learned from the various implementations of e-textbooks on college campuses in the 2009-2010 academic year and portends for the coming years. It also notes that the e-textbook is now in its important stage of learning curve.

In 2009, the Princeton University held a pilot experiment on the current e-reader using in the classroom. The classroom experience was worsened by using e-reader, unsurprisingly. Meanwhile her final report in 2010 shows some valuable suggestion for future e-book manufacturers such as annotation tools, pagination, content organization, and in achieving a more natural "paper-like" user experience

ISBN: 978-1-61208-116-8

[18, 19]. We emphasize here that all the required techniques will be achieved by the manufactures soon. What left unsolved are the habits of stakeholders, especially the teachers and students.

Reading is an important part of learning; however, it is not vice versa. Reading is not equal to learning. The textbook has its specific usage for learning. The textbook has its own characteristic use habits and application environments. Thus the development of e-textbook definitely needs a different position of e-book. Under current basis of e-book, we should develop learning related technique functions and usage process, provide suitable online platform as well. In this way, the learners and instructors may enjoy reading the e-book. Moreover, under the enabling of e-textbook, they may construct an environment for digital learning interaction, and reach the goal of e-learning.

Contrary to the prosperity of e-book, let us take a look at the textbook market. Currently, in the United States, the publishers find great obstacles selling the online components of the textbook, due to the maturity of environment as well as the technology resistance of instructors. In addition, there should be a better pricing mechanism in balancing the huge cost of developing online components and the great burden of students. The vibrant second-hand textbook market may portray part of facets of the ecology.

The remainder of this paper is organized as follows. Section 2 details the major problems in marching from e-book to e-Textbook, nature of e-Textbook, the benefits of users and markets, and the requirements for its innovations. Section 3 proposes our comprehensive strategy framework from the perspective of information goods, technology, and stakeholders. Section 4 states the possible business models under this framework. Finally, conclusions and future applications are given in Section 5.

II. MARCHING FROM E-BOOK TO E-TEXTBOOK

The success of e-book cannot guarantee the success of e-Textbook according to the characteristics of e-books stated above. In the developing of e-Textbook, we will face several major problems. We analysis with a systematic approach composed of three dimensions: information goods, technology, and students' habits. Followed this, we detail the nature of e-Textbook and the requirements for its innovations.

A. The Problems Faced When Moving from E-book toward e-Textbook

From the development process of e-book and the background of digitalized text textbook, we figure out six major problems for e-textbook from current studies [9, 12, 15, 17]. (1) learning vs. Reading, (2) E-ink or i-Pad, (3) service vs. product, (4) proprietary vs. Open, (5) interface function, and (6) cash flow mechanism. We detail them with a systematic approach that is composed of three dimensions: information goods, technology, and students' habits. They are depicted in Table 1.

TABLE 1 THE PROBLEMS FACED WHEN MOVING FROM E-BOOK TOWARD E-TEXTBOOK

Perspective	Option A	Option B	Considerations
Information Goods			
	product	service	Single product or bundled services
	pricing	costs	Reasonable, acceptable, comprises, sales, costs, services, for teachers. cash flow, stakeholders, needs
Technology			
	E-ink	i-Pad	usability, reading habits, interaction
	proprietary	open	Market, communication, benefit to all
	simple	flexibility	Interface, Behaviour, learning processes, flows, flexibility
			Software, hardware
Students' Habits			
	reading	learning	Interaction
			Chapter or whole book, e-Publisher industry

1) Problems Faced in the e-Textbook

a) The Perspective of Information Goods

(1) Service or Product?

We have to keep in mind that the big success of Kindle is not just because it is only a product. Besides the portable E-ink display interface, it is also composed of Amazon's online book ordering services. Through the use of network, it is empowered with the marketing, digitalized book, delivery of magazine and books, browser interface, connectivity to information, the online shelf of the platform enables every user have his own central storage spaces, the extensions of application and function enable every Kindle have its own e-mail account to communicate the digital contents. Thus, when we face the promotion from e-book reader to e-Textbook. We strongly posit that e-Textbook is just equivalent to the simple thinking of combining e-book and textbook digitalization [12].

We have to plan the function specification of e-Textbook properly. Moreover, we get to have an integration of product, service, and platform. Then we will have a way to success in promoting e-Textbook.

(2) Pricing

The iPod and iTunes of Apple have changed not only the pricing thinking but the market of digital music. The related stakeholders of digital music have now acknowledged that USD$ 0.99 each song is a reasonable price for consumers to get. Amazon has gradually leading the price strategy, too. She sets the price level of USD$ 9.9 each digital book because of her leading place in online book sales and success of Kindle. However, there are still many comprises among teaching and learning, limitation of sales, high costs, and the services for teachers.

Therefore, in the e-Textbook industry, the suitable cash flow mechanism must be well established. The stakeholders including authors, publisher, teachers, and students. They have to deal with the common pricing strategy to satisfy their needs. Through the clearing and cash flow mechanism, the reasonable price to get the digital textbook must be recognized by the stakeholders in getting not only the digital contents but also the paper version.

b) The Perspective of Technology

(1) E-ink or i-Pad

We have to face the problem of use after the digitalization of books. Many researches show that it is the usability, which hinder reader to use the e-book [3, 6, 7, 12]. For the reading habits, [1] stated the readers prefer paper books because it is difficult in reading on the screen. The e-book readers which use the reflective-like technique of E-ink electronic paper make themselves to be near to paper books. While the technique characteristic of low response of E-ink electronic paper also limit, e.g. it cannot scroll the digital contents or paging the contents. Besides, the better ability of interaction and information retrieval are what digitalized books should specially possess [14]. However, the iPad sales hit two million in April 2010 shows the opposite strength.

(2) Proprietary or open system

The development of e-book shows a trend from proprietary to open system both on the software and hardware. In early development, the system specification goes proprietarily for reaching each own objective. However, it limited the development of whole e-book industry. On the contrary, facing the trend of open system, Amazon develops Kindle DX for popular file specifications as pdf of Adobe and Office of Microsoft. The same situation occurred in the pioneer electronic company, SONY. She abandoned her own e-book specification and engaged the common standard of e-book content specification named e-PUB. Therefore, we should remember the lesson [9]. Open minds create more better developments for the industry.

(3) Interface Functions

Current interface functions of e-book reader are positioned to reading. These may not be completely matched all the necessary functions in promoting e-book reader to textbooks. Instead of directly fitting the learning behaviour of paper textbook to the e-Textbook [9, 12], it is necessary to research and confirm the learning processes needed in the e-Textbook. Such processes consist of interface, functions, and the flows.

c) The Perspective of Students' Habits

There are distinct human factors between the process of learning and reading. The study of campus e-book by [13] noted few people read the entire book. Most of the readers read only several chapters but not the whole book. The average time they spend on e-book is less than 20 minutes. From the teaching practice in the university, most of the teachers will choose some chapters as specified materials, and add some supplementary materials. Thus, the mechanism of e-Textbook should provide chapter by chapter instead of whole book.

Reading is much about personal activity. However, besides the above activity, learning effects are much owing to the interactions among teacher, student, and peers. Hereby, the current e-book readers have to make great progress on the interaction functions in order to apply to domain of learning.

In learning or textbook, the needs of chapter or whole book are diverse. This situation is quite different from in reading. The e-Publisher industry has begun to think a new pricing mechanism for chapter by chapter and a whole book.

2) Requirements for Consumers and Market

Many studies reveal the requirements for consumers and market for current e-books. The largest survey in the use of e-book so far is the survey of MyiLibrary of JISC in England. After a survey of 22,437 users from more than 120 universities, it showed that nearly 62 percent of students have already used e-books in their researches or instructions [13]. Reference [11] surveyed on Denver University showed that 51 percent of students do use the e-books, 56.6 percent of them read more than one chapter online but not printing out. More and more people use e-books because of saving printing fare and convenience [17]. The students use e-books three times of paper books [7]. The trend of e-book entering into campus for learning is obvious. As stated above, Forrester predicts the textbooks will be the greatest e-book reader market five years later.

Although the trend is emerging, there is still great inconvenience in using e-books, which hinders the popularity and promotion. The major inconveniences encountered are subsequently ”not used to reading on the screen for a long time”, ”special devices needed or technical problems”, ”inferior quality”, ”hard to read and browse”, and ”lack of ability in searching local contents” [5, 6, 11, 15].

The above inconveniences show the requirements for users in the development of e-books. In moving to e-Textbook, despite the digital contents, there must be convenient terminal reading devices, convenient operation, humanistic action flows, speed for information searching and interaction. To meet the needs of textbooks, there are some important requirements as follows: the total mechanism must be convenient and complete to support instruction, strategy of sales and Profit sharing to support whole book and chapter-by-chapter, function of printing on demand (POD), sales and service chain for click and mortar, an integrated platform for learning and logistics.

B. Nature of e-Textbook

The e-Textbook is kind of technology–mediated learning by nature. The technology here is as mediation and communication for learning. For the mediation, as stated above, we have to fit the e-Textbook to the characteristics of learning. For the communication, learning is not independent activity but a community of practice.

Learning also happens in the interaction among the teachers, students, and textbooks or auxiliary materials. Such situation is both un-substitute and un- duplicate.

C. Requirements for Innovations

Currently the e-Textbook is on its way to innovations. We begin with the lessons learned from some company with vision. We then draw the future requirements for e-Textbooks and finally explore the innovations for e-Textbooks. We depict there is a spectrum moving from e-book, e-book for learning, to e-Textbook.

1) Lesson Learned from e-books

O‘Reilly Media Group put her feet in the e-book market early in 1987. At that time, she boldly held the opinion of throwaway DRM (digital rights management) and multi channels simultaneously. The benefit today shows the value of such strategy. In O‘Reilly's opinion, publication being invisible is far terrible than being pirated. It can do more harm to the authors using the DRM to avoid pirates.

Owing to the strategy of open to digital contents and value-added services, O‘Reilly Media enables her e-book users to own various versions for different kinds of terminal readers, to do multiple downloads, to enjoy the latest version all their life. Instead of limiting, she shows the strategy of open, multiple channels, and sharing. This strategy makes the publication ubiquitously.

The lesson learned from O‘Reilly Media Group is valuable to the development of e-Textbooks. In the age of information explosion, publishers have to change their roles from information providers to services providers. The consumers do not pay for the contents but for the value. The value for the publishers is the knowledge rectification, arrangements, and mediation. Moreover, the information communication technologies (ICTs) provide the new value of ubiquitous and integrated services. This is the key to success for e-Textbook mechanism.

Owing to the quick changing of time and knowledge, traditional time-wasting publishing cannot meet the quick updating for today. On the other side, this is just the niche for electronic publishing (e-Publishing) companies. They have to be a suitable mediator of knowledge communication. The textbooks are oriented for knowledge creation and learning. The need for knowledge communication is specially suitable and urgent for e-Publishing. The system for e-Textbook has to provide open shelf system, combination of current sales model, click and mortar in campus, continuous updating and supplements, service mechanism of Wiki and Web 2.0, and corresponding payment system.

2) Future Requirements for e-Textbooks

We can derive future requirements for e-Textbooks from the imagination of future benefits for users. They are stated as follows.

a) The implementation of the concept of electronic bags

Past experience of the electronic bags shows insufficient value-added application software, poverty of contents, unclear position of users, lack of team coordination, late for content rights, neglect of user habits, and lack of market education. Without regarding the electronic bags as a whole system can account for the failure. In this case, it is hard to integrate each component as well as comprehensive planning.

The technique of e-Textbook plays an important part in the future implementation of the concept of electronic bags.

b) e-learning community of practice

A well-planned system of e-Textbooks may fully utilize the connectivity of ICTs. Hence, such e-Textbook may connect people and information as well as integrate into the e-learning community of practice

c) New styles of e-Publishing industry

The publishing industry may employ ICTs for transformation. Such transformation may add the value needed in the age of information explosion. The e-Publishing system enables a new style of flexibility of publishing, quick mediation ability, and instant communication speed.

III. OUR PROPOSED FRAMEWORK

Following the backgrounds and requirements for innovations described in former two sections, we propose a comprehensive strategy framework for e-Textbook based on the concept of information goods, and stakeholders. The stages we concerned covers pre-class, in-class, and after class. In this way, we make the learning happen ubiquitously. They are depicted in Figure 1 and elaborated as follows.

Levels	Items			
Information Goods	Non open (e.g. DRM)/ open materials			
Technology	OSPH (e.g. MID) /	OSOH (anyone) /	PSPH (eg. Kindle) /	PSOH (e.g. NoteBook)
Stakeholders	Authors, Publishers, Teachers, Students			

Figure 1. The comprehensive strategy framework for e-Textbook

A. The Level of Information Goods (open/ non-open)

As depicted in 1), in this level, we should consider not only the service or product, but the pricing mechanism. Through the lesson at Princeton, we may plan the function specification of e-Textbook properly. Moreover, we get to have an integration of product, service, and platform. Then we will have a way to success in promoting e-Textbook. Amazon has gradually leading the price strategy. She sets the price level of USD$ 9.9 each digital book because of her leading place in online book sales and success of Kindle. How can the related stakeholders of e-Textbook acknowledge a reasonable price for producers to take and for consumers to get. There are still many comprises among teaching and learning, limitation of sales, high costs, and the services for teachers. Through the clearing and cash flow mechanism, the reasonable price to get the digital textbook must be recognized by the stakeholders in getting not only the digital contents but also the paper version

B. The System for e-Textbook: Proprietary or Open System for Software (PS/OS)

Except running on a proprietary system, the e-Textbook reader may be developed in the open operating system, e. g. Android platform. Reference [2] argued that e-learning should promote from application level to function level. The application level contains sharing, presentation and media of materials and quizzes with ICTs. However, the function level consists of recording of learning process, statistics, reservation, and evaluation in passive function; analysing learning status, detecting learning myth, guiding learning, implementing planned instructions, and test/ evaluation in active functions.

Current e-book Readers do not emphasize the application and function level for e-learning. Even the Amazon Kindle has already provided a similar and basic Whispersync function for synchronization among multiple Kindles, it still lacks the learning records function, neither records learning status in order to analysis, no to mention the function of quiz or evaluation.

On the other hand, the Living Lab is a user centered innovation in real life scenarios. William J. Mitchell from MIT developed it[20]. It is user-centric research methods in real life environments to identify and build prototypes, and to evaluate multiple solutions. By means of the concept of Living Lab, we will put key stakeholders to co-develop in this experiment; we may evaluate and verify whether the functions of e-Textbook readers in Android platform can satisfy the learning needs.

For the platform and business mechanism, we also design an experiment of two classes in different schools. The consumer information and behaviour will be accumulated in the experiment. The results will validate the pricing strategy and institution.

C. The stakeholders in the e-Textbook industry

The stakeholders including authors, publisher, teachers, and students. For the authors, a well developed environment encourages their fruitful contributions. Since e-Publishing is unavoidable, the publishers have to embrace and try to create even more opportunities than ever. The most hardest part will be the habits of teachers and students. The habits of learning and teaching are not easily changed in a short period of time. In this case, they may suffer before they befit from the e-Textbook. The learning community of practice may be one of the possible solutions. As stated above, learning is not only personal activity but also a group one. The e-Textbook we design should meet the needs for learning community of practice. There are group installing and monitoring for the group learning, sharing and collaborating supports of various kinds of instruction approaches and learning style.

As the e-Textbook network platform, [14] stated that e-book should possess the following functions: interaction, hyperlink, browsing, searching, and link to online services, moreover, continuous updating. Several functions are expected in the e-Textbook network platform. There are Sharing, Web 2.0 knowledge spread, accumulate, and creation as stated above. Just think about it, even there is a great success at Princeton this time, the e-Textbook may not be available in other campuses on account of the different situation. That means only when we can grasp the characteristics of the students, the e-Textbook can be successful.

IV. THE BUSINESS MODEL

Based on the above framework for e-Textbook, there are several business models. The business model stated here focuses the trend of e-books, considers most of the backgrounds and requirements depicted in Section 2. It aims to create an open and integrated e-Textbook system. We stated the technical components and management issues in the business model respectively in this section.

As depicted in Figure 2, all the relations of the products and services are noted, the bones between them are the management issues to be handled. The technical components in the business model is four-fold: A. open, and sharing; B. personal services for authors, teachers, and learners; C. multiple channels of sales and cash flow; and D. the technologically ubiquitous learning environment.

A. Open and sharing

For the open and sharing part, there are (1) format transformation for e-book, (2) services across book sales platform, (3) shelving system, (4) service of sales chain with click and mortar, (5) supplementary materials of digitalized textbooks, and (6) mechanism of Web 2.0 like Wiki.

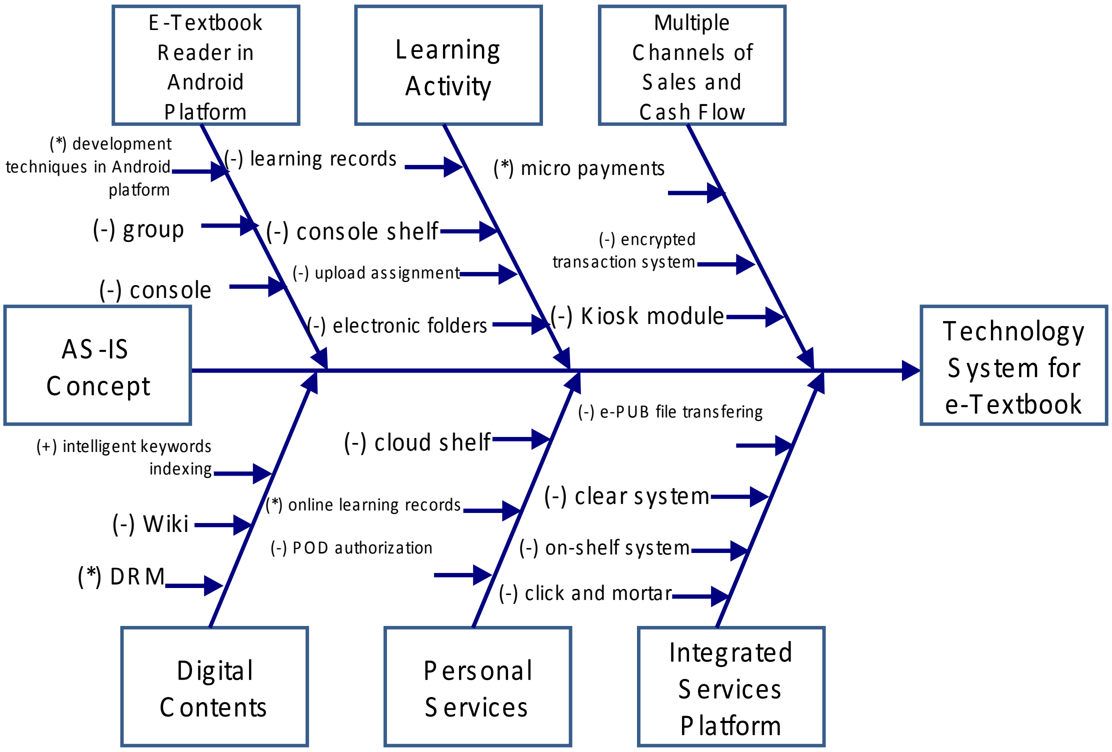


Figure 2. The relations of the products and services for the e-Textbooks (the signs denote the developments of product, technology, or service are * for developed, + for under developing, and - for will be developed in this project)

B. *Personal services for authors, teachers, and learners*

In this part of personal services for authors, teachers, and learners, there are (1) service of personal book store, (2) clearing and authorization chapter by chapter, (3) service of POD authorization, (4) push service, and (5) personal service.

C. *Multiple channels of sales and cash flow*

The multiple channels of sales and cash flow include: (1) sales of deposited encryption letter, (2) sales interface of Easy card depositing, and (3) module of cash flow.

D. *The technologically ubiquitous learning environment*

As a technologically ubiquitous learning environment, there are (1) records for learning activities, (2) console shelf, (3) assignment uploading and quizzes, (4) electronic folders, (5) group, and (6) console.

V. CONCLUSION AND FUTURE WORK

We have stated the efforts need to be elaborated when moving from e-books to e-Textbook. We highlight open problems regarding the technology as well as the way of distribution. Again, we emphasize that the lesson learned from the pilot at Princeton is not enough. Simplified speaking, advanced in technology only is not enough. The comprehensive strategy framework may cover the requisite needs. The stages covered in our proposed architecture covers pre-class, in-class, and after class. In this way, the learning will happen ubiquitously. The contributions of this paper to the approach and practice in e-Textbook industry are three-fold: (1) We go through the research and practice of e-books through a systematic survey framework, (2) We not only highlight solutions to the problems e-Textbooks will encounter, but also propose a framework for e-Textbook. (3) We also illustrate with possible business models for e-Textbook industry. At the end of this Section, we also give suggestions for the future study in e-Textbook.

A. *Conclusions*

(1) We go over the research and practice of e-books through a systematic survey framework of three dimensions: technology, stakeholders, and information goods. The results are noted in Table 1. Thus, we may have a complete comprehension starting from the e-books.

(2) Tracing on the path of e-book, we march to e-Textbook. With a project of the real case of Publisher Lisa, the journey is now on its way. We highlight solutions to the problems e-Textbooks will encounter in Table 1. Through the holistic point of view, we regard e-Textbook not just a product but a system. Such system aims the learning activities and relates together all stakeholders in social, economics, institution.

(3) Besides the comprehensive strategy framework for e-textbook depicted in Figure 1, we also illustrate with possible business models for e-Textbook as depicted in Figure 2. Figure 2 demonstrates all the related products and services with management issues through the proposed approach.

B. Future work

There are still some further issues worthy to study. The framework we proposed may try to apply to diverse ways of technology-mediated learning. Other salient issues are the evaluation criteria for the technical system, the possible sources of technology/ services and patents, the searching and management of intellectual property.

Acknowledgment

This study is conducted under the Project Digital Convergence Service Open Platform of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China

References

[1] Abdullah, N. and Gibb, F. (2008). Students' attitudes towards e-books in a Scottish higher education institute- part 1. Library Review, 57: 593-605.

[2] Alessi, S.M. and Trollip, S.R. (1991). Computer-based instruction: Methods and development (2nd ed.). Englewood Cliffs, NJ: Prentice Hall

[3] Anuradha, K. and Usha, H. (2006). E-books access models: an analytical comparative study. The Electronic Library, 24: 662-679.

[4] Borchers, J.O. 1999. Electronic books: Definition, genres, interaction design patterns. In ACM CHI 1999, Electronic Book Workshop. . Pittsburgh, PA.: Citeseer.

[5] Carlock, D. and Perry, A. (2008). Exploring faculty experiences with e-books- a focus group. Library Hi Tech, 26: 244-254.

[6] Chu, H. (2003). Electronic books: viewpoints from users and potential users. Library Hi Tech, 21: 340-346.

[7] Croft, R. and Bedi, S. 2005. e-books for a distributed learning university: The Royal Roads University case. Journal of library administration, 41: 113-137.

[8] Henke, H. (2001). Electronic books and ePublishing: a practical guide for authors: Springer Verlag.

[9] Herther, N. (2005). The e-book industry today: a bumpy road becomes an evolutionary path to market maturity. The Electronic Library, 23: 45-53.

[10] Lemken, B. (1999). Ebook: the missing link between paper and screen. ACM CHI 1999, Electronic Book Workshop. Pittsburgh, PA.: Citeseer

[11] Levine-Clark, M. (2006). Electronic Book Usage. portal: Libraries and the Academy, 6: 285-299.

[12] McFall, R. (2005). Electronic textbooks that transform how textbooks are used. The Electronic Library, 23: 72-81.

[13] Nicholas, D., Rowlands, I., Clark, D., Huntington, P., Jamali, H., and Olle, C. (2008). UK scholarly e-book usage: a landmark survey. London: Aslib., pp. 311-334.

[14] Nielsen, J. (1998). Electronic books–a bad idea. Jakob Nielsen’s alertbox, 26. Retrieved

[15] Shelburne, W. (2009). E-book usage in an academic library: User attitudes and behaviors. Library Collections, Acquisitions and Technical Services, 33: 59-72.

[16] Simba Information (2010). E-Textbooks in Higher Education 2010-2011. Simba Information.

[17] Storch, J. (2009, February 6). Needed: A single electronic source for textbooks. The Chronicle of Higher Education, Retrieved from: http://chronicle.com/weekly/v55/i22/22a03201.htm on 2010/12/30

[18] Princeton University (2010a). The E-reader pilot at Princeton. Final report, (long version). The Trustees of Princeton University.

[19] Princeton University (2010b). The E-reader pilot at Princeton. Final report, (executive summary). The Trustees of Princeton University.

[20] Schuurman, D., T. Evens and L. De Marez (2009). A living lab research approach for mobile TV, ACM.

QoS Driven Semantic Based Grid Service Composition Using AND/OR Trees

Xhemal Zenuni
Faculty of Contemporary Sciences and Technologies
South East European University
Tetovo, FYR of Macedonia
e-mail: xh.zenuni@seeu.edu.mk

***Abstract*—Composition oriented service discovery is an important requirement and research challenge in Service Oriented Architectures (SOA), such as in the third generation of Grid. It provides added value services, more rapid application development and improved reusability of existing services. This paper proposes and formulates a composition approach where semantic information is used to determine Grid services dependencies in form of AND/OR tree associated with semantic QoS information, thus transforming the issue of service composition discovery to constrained AND/OR tree search problem. Different constraint forms and QoS aggregation patterns in such trees are analyzed, and some major constrained searching techniques that can be applied to such trees are discussed. Our findings show that AND/OR trees are expressive in addressing QoS – aware semantic Grid service composition and able to employ different discovery searching techniques for that purpose.**

Keywords-Grid service composition; QoS; AND/OR tree; Semantic Web; Ontology;

I. Introduction

The Web Service Resource Framework (WSRF) [1] set documents of specifications defines a new formal framework for building current and future Grid applications based on Service Oriented Architecture (SOA) principles. In such Grid, services are becoming the fundamental building blocks and the basic collaboration element which can be used to build grid applications and resolve complex scientific problems.

In this conceptually new approach, composition oriented service discovery brings multiple benefits. In many situations, individual services in isolation are limited to respond to more complex user demands. However with the combination of several ones together, new solutions not anticipated in individual services can be implemented and more complex problems can be solved effectively. In addition, the same service can be combined in many composite ones, thus enabling better service re – usability.

On the other hand, when the composition process is automated, new services can be constructed faster and with less effort, thus accelerating a rapid application development in Grid. Moreover, a good service composition middleware can hide the composition details, by making visible to users only the available interfaces. In effect, this black – box encapsulation can simplify their usage.

Service composition system is part of a larger lifecycle development in Grids and imposes a list of requirements that can not be definitive. However, in order to achieve enhanced service composition process, some fundamental requirements and challenges need to be addressed timely, especially faced with the service proliferation. The composition middleware should effectively and efficiently discover service dependencies and coordination rules of different services in repositories, and conducting this in an automatic manner. Secondly, Grid systems are dynamic, with services created and destroyed on the fly. Service composition system should be adaptable and must detect those changes, and make quality decisions at run – time. Moreover, in large repositories and for a given problem, more than one solution may exist. The system should allow the users to define extra non – functional properties and preferences as a discriminating and/or ranking factor. Furthermore, QoS becomes a common model for narrowing the list to best discovered solutions.

Considering all these factors, this paper presents a holistic approach to service composition based on three fundamental elements. Semantic Web technology has been used to model and describe Grid Services functionalities and QoS features and as enhanced background to determine service interdependencies. Then, from functional point of view, these dependencies are expressed in form of AND/OR tree. Finally, composition services are discovered by QoS constrained search in such AND/OR trees.

The rest of the paper is organized as follows. Section 2 briefly presents the related work. Section 3 explains the semantic model developed to describe Grid service features, and explains how this model enhances the discovery of services. Section 4 introduces AND/OR trees and how service dependencies can be expressed through them. Section 5 explains the different forms of QoS preferences that a user may express, the aggregation patterns that occur in AND/OR tree structure and how to calculate end – to –end QoS dimension of composition services in such situations. Section 6 investigates some major constrained searching algorithms that are applicable in AND/OR trees. Finally, Section 7 concludes the work and gives the future directions for improvements.

II. Related Work

Many works on service discovery and automated service composition have been reported, especially as Artificial Intelligence planning problem [2]. However, the focus here is more on graph approaches based on I/O data and semantic information of services, which closely relates to our work. Liang [3] proposes a semi – automated method for service composition based on AND/OR graphs and applies the

REV* searching algorithm to find composite services. However, in this approach, the role and usage of semantic information in such model is not comprehensible. In addition, it didn't consider the scale of services. On the other hand, Gu [4] advances the work by presenting a faster service composition method, where indexing of services plays critical role for acceleration of composition algorithm. It proposes also a method how to handle semantic relationship between I/O data, but yet the method doesn't fully explore advantages and conversion of semantic information. Yan [5] goes further, especially by improving the searching algorithm to support the recognition, conversion and usage of semantic information described in OWL format. However, in this approach, as with all the other above, service composition is mainly seen from functional aspect of services. They have been used to determine the service dependency graph and then a searching algorithm is applied to find composite services. Our work is distinguished at least in two aspects, which we consider as contributions. First, we developed QoS ontology for describing the non – functional aspects of services, allowing users to define QoS preferences as well. This becomes necessity with service proliferation, and when many solutions may be anticipated. Secondly, we extended some searching algorithms for AND/OR trees to find composite services that fits to user constraints.

III. Semantic Grid Services

Industry standards for Grid Services, such as UDDI (Universal Description, Discovery and Integration) [6], WSDL (Web Service Description Language) [7] and SOAP (Simple Object Access Protocol) [8] focus on operational and syntactical details, which in turn make service publication, discovery and composition process very restricted. To overcome the limitations of keyword oriented searching with such standards, Semantic Web technologies have gained momentum as an approach that can provide better background and enhanced service discovery and composition mechanisms based on semantic information.

To this point, the ontologies for service discovery and composition can be defined at two main levels. The first level consists of domain – specific ontology, which describes specific domain concepts, in form of class and sub – class hierarchy, individuals of such classes and other relationships in them, such as synonyms, etc. The second level consists of an upper ontology, which provides uniform description of provided services. Several upper ontologies [9][10] for service modeling have been reported, mainly to describe them in terms of IOPE (Inputs, Outputs, Preconditions, Effects) which drive the composition process, but as discussed in [11], such ontologies have two main drawbacks when applied to Grid Services. Grid services usually act upon some resource, and the semantic information of Grid "resource" is absent. And secondly, the QoS features of services are not represented. To overcome these limitations, a new ontology model that includes these two aspects was presented in [11].

This ontology model allows semantic description of different aspects of Grid Services. The robustness of this model can be seen in many directions. First, a search algorithm can recognize different relationships defined in domain specific ontology, such as instance and concept relationships, other relationships such as synonyms, and use the semantic information to better recognize the relationships of different services and their coordination rules.

Secondly, service providers can describe multiple dimensions of any arbitrary QoS parameter for the services they provide, such as its overall impact, if it is measurable, the metric, the unit used and so on. This features become important ranking and/or discriminating factor in service provisioning.

Finally, service requestors and providers can express their QoS preferences in different forms, especially using different measuring units, and a search algorithm can support their equivalence if the relationships of different units if they are priory defined.

IV. AND/OR Background

An AND/OR graph (and AND/OR tree as special case) [13] is a structure commonly used in automatic problem solving where the solution involves decomposing the problem into smaller problems, and then the solution is found by solving this small tasks. It is a generalization of directed graphs, consisting of two types of nodes (connectors), namely AND connectors if there is a logical AND relationships in such nodes, and OR nodes if there is such logical relationship. In such graphs, the terminal nodes are solved nodes. If non – terminal node has OR successors, then this node is considered as solved only if at least one of its successors is solved. Contrary, if non – terminal node is AND node, then it is solved only all of its successors are solved.

These characteristics place AND/OR graphs as powerful formalism to express service dependency graph of services, because it can handle *n – to – m* relationships of I/O. Such dependency graphs are created by analyzing inputs and outputs of available services.

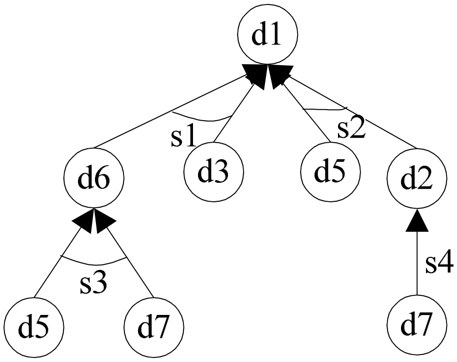


Figure 1. AND/OR tree representing service dependencies

For example, if four services are available in a service repository, such as *s1={In(d3,d6}, Out{d1}}, s2={In{d2,d5}, Out{d1}}, s3={In{d5,d7}, Out{d6}}* and *s4={In{d7} ,Out{d2}},* then the service dependency graph can be represented in AND/OR tree like in Figure 1.

For example, if a user specifies the request in form of desired outputs (or resource) *O={d1}* and the available inputs *I={d3, d5,d7}* then a solution can be found in two ways. The first solution is using a chain of services: *s1* and *s3.* The second solution is the service chain: *s2* and *s4.* Indeed, in large repositories more then one solution can be anticipated. In this case, QoS features become a common model to discriminate and distinguish the different solutions. Therefore, service composition model should allow users to define non – functional demands as well as an important requirement faced in huge service repositories.

V. Measuring QoS of Composite Services

In large repositories with services that overlap in their functionality, QoS is becoming natural discriminating and ranking factor. Quality of service is important in composite services as well, as they should not respond to business complex needs only, but they must perform within the limits of given QoS constraints.

User preferences over quality of service parameters can take different forms. In certain cases, user may express certain QoS requirements over a single variable only, such as the response time solely. But in many situations, users have more complex demands, and the quality of composite service is evaluated based on multiple variables, like response time, cost, throughput, and so on. Moreover, preferences may come in form of constraint satisfaction or constraint optimization problem. In former situation, given the AND/OR tree with QoS data, the quest is to find solutions that satisfy the given constraints. In later case, the composition system should be able to find “the best” solution that maximize or minimize the given objective function.

Based on this, different QoS constraints over composite services can be applied, ranging from single variable as constraint satisfaction problem to multiple variable as constraint optimization problem. The later is considered especially difficult situation. In presence of multiple QoS parameters, it is a difficult task to find the optimal solution and there must be tradeoffs among different quality criterion. It is not always possible to find a solution with minimum execution time, minimum cost and highest availability rate of services.

Moreover, not all QoS parameters follow the same aggregation pattern when doing end – to – end planning. The solution of AND/OR tree is rather sub tree than a path, and this in turn involves combination of parallel and sequential vertices. This implicates different aggregation patterns when calculation the global QoS of the composite service. Indeed, for the computation of the global QoS, four different aggregation functions have been identified, and brief explanation is provided in what follows:

1. No aggregation can be applied: certain QoS can not be aggregated. This is especially true for non – measurable ones, such as the requirement that each service has to support “SOAP v2.0” or “X.509” digital certificate for communication. In this case, such QoS are used on the level of local planning, even when end – to – end QoS analysis is performed.
2. Critical Path Calculation: in parallel structures, for some QoS is taken into consideration the maximum value as valid. Such example is the execution time of services. There is no point to further minimize the value of the lower execution time in parallel structure, since has no affect to global QoS.
3. Sum function: some QoS are aggregated using pure sum function. For example, the price of the composite service is calculated as the sum of all involved services in solution, regardless in parallel or in sequential manner.
4. Average sum: some QoS parameters, such as reputation, are aggregated as average sum of all involved services in solution.

VI. Searching Composite Services

Different searching techniques can be applied to AND/OR trees, depending on the form of QoS preferences. We have mainly considered multiple QoS parameters, and three fundamental searching approaches that can be applied in such situations are discussed in what follows. First approach can be used for QoS constraint optimization solution, and the other two algorithms for multiple QoS constraint satisfaction solutions.

First, in presence of multiple QoS values, we can transform them in a single value using the equation described in:

$$tw = \sum_{i=1}^{N} w_i * QoS_i \qquad (1)$$

where N is the number of different QoS taken into consideration, w_i is the weight that the user gives to QoS parameter *i*. In addition, the following condition should hold: $w_i = [0,1]$ and $\sum_{i=1}^{N} w_i = 1$. Thus, the multidimensional QoS is transformed into one single value. In this situation, the AO* algorithm [10] can be applied directly to find composite services.

The limitations of first method is that not every QoS can be expressed using the Equation 1. Many QoS are not measurable and are not numbers. In this case, the first approach is not applicable. In addition, we may want to express some specific QoS boundaries that we do not want to be exceeded. Thus, the issue is transformed to constrained satisfaction problem, and not optimization one.

A critical issue when attempting to find the solution tree at run time is the ability to calculate QoS aggregate values of the multiple parameters that are presented in Section 5, and how to eliminate nodes that exceed the preset QoS threshold from further expansion. Non – measurable QoS parameters are easy to use, because those services that violate these criteria's are automatically discarded during AND/OR tree expansion. For other patterns, we modify the QoS parameters values of all related nodes in the solution path recursively using the modification formulas as follows:

- For Critical Path Calculation (e.g. execution time):

$$ET(n) = \begin{cases} ET(n) + ET(p),\ n \text{ is AND node} \\ ET(p),\ n \text{ is OR node} \end{cases} \qquad (2)$$

where p is parent node of n.

- For sum pattern(e.g. price):

$$\Pr ice(n) = \begin{cases} \sum_i \Pr ice(c_i),\ n \text{ is AND node} \\ \max_i \Pr ice(c_i),\ n \text{ is OR node} \end{cases} \qquad (3)$$

where c_i are children's of node n.

- For average sum (e.g. reputation) the situation is more complex, and we keep two data, i.e. the reputation of node and how many services contributed to that reputation (N):

$$R(n/N) = \begin{cases} \dfrac{R(n) + \sum_i R(c_i)}{N + \sum_i R(N_i)},\ n \text{ is AND node} \\ \min_i R(c_i/N_i),\ n \text{ is OR node} \end{cases} \qquad (4)$$

where c_i are children's of node n.

Then, two major approaches could be applied directly: breadth – first constrained searching and depth – first constrained searching. The pseudo – code for constrained breadth – first searching in bottom – up fashion is presented in Algorithm 1.

Algorithm 1: Pseudo code for breadth-first AND/OR tree constrained searching

```
1:  Put the start node s (s points to desired outputs) on a list called OPEN
2:  Remove the first node on OPEN and put it on another list, for example called CLOSED and call this node n.
3:  Update QoS aggregate values recursively in expanded tree using Equations 2, 3 and 4.
4:  Check constraints.
5:  If QoS violated, label then the node n as UNSOLVABLE and continue. Otherwise go to step 9:.
6:  Apply the unsolvable – labeling procedure to the search tree.
7:  If the start node is labeled unsolvable, exit with failure; otherwise continue.
8:  Remove from OPEN any nodes having unsolvable ancestors and their influence in the overall QoS values and go to step 2:.
9:  Expand node n, generating all its successors. Put these successors at the end of OPEN and provide pointers back to n. If there are no successors, label n as UNSOLVED and go to step 6:, otherwise continue.
10: If any of the successors are terminal nodes (desired inputs of services), label them as SOLVED and continue; otherwise go to 2:.
11: Apply the solve labeling procedure to the search tree
12: If the start node is labeled SOLVED, exit with the solution tree that verifies that the start node is solved; otherwise continue;
13: Remove from OPEN any nodes that are solved or that have ancestors that are solved
14: Go to 2:.
```

We first create an auxiliary node *s* and connect it to user desired outputs. In breadth – first fashions node expansion, the solution tree is incrementally enlarged by adding more nodes to AND/OR tree structure, and then we continuously update the QoS aggregate values of nodes using Equations 2, 3 and 4. If the QoS contribution of the last node violates the user's preset QoS threshold then this node is removed from further expansion, including its influence to the possible solution tree. We stop the whole procedure when the start node *s* is marked SOLVABLE or when there are no further nodes to expand.

Indeed, the AND/OR tree contains two types of nodes. Data nodes (input and output of services) which are of type OR nodes, and they do not contribute directly to the aggregate QoS values. On the other side, services are AND type of nodes, because all their inputs must be available for their successful invocation. These nodes directly contribute to the overall QoS of composite services.

In following paragraphs we provide an example that illustrates the way the algorithm proposed in the previous section works. Assuming the repository presented in Table I, a simple request given by an imaginary client would be as follows:

- Output: ZipCode, PriceDollar,.
- Input: Book, GoalCurrency, City.
- Constraints: execution time to be less then 7 millisecond, price to be less then 10 dollars ($), and reputation to be greater then 0.80.

TABLE I. SAMPLE SERVICE REPOSITORY

Service	Input	Output	QoS
CurrencyConverter (CC)	PriceEuro GoalCurrency	PriceDollar	[2,3,0.85]
BookpriceFinder (BF)	ISBN	PriceEuro	[1,2,0.87]
ISBNFinder (IF)	Book	ISBN	[3,2,0.80]
ZipCodeFinder (ZCF)	City	ZipCode	[2,2,0.82]
CompositeService (CS)	Book GoalCurrency	PriceDollar	[7,1,0.80]

In Table I, the QoS parameters of services are expressed in form of a vector. The first element denotes its execution time in milliseconds (ms), the second parameter denotes its invocation price in dollars ($) and the third element gives information about its reputation. Figure 2 illustrates fragments of trace of breadth – first searching in AND/OR tree, solving the given problem.

The final solution is found using services ZCF, CC, BF and IF. Another alternative solution can be obtained using services ZCF and CS. However, during the tree expansion, CS exceeds the threshold of execution time to be less than 7 milliseconds.

The best – first constrained search can be applied in similar fashion, by expanding first the recently generated nodes firstly.

Although the implementations details are out of the scope of this paper, a prototype system that serves as proof – of – the concept discussed in this paper is developed. It consists

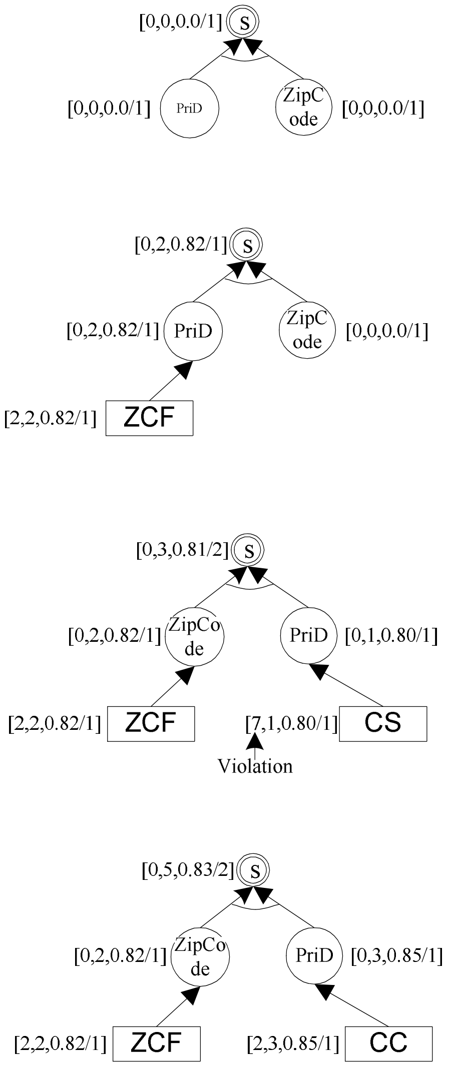


Figure 2. A trace of breadth –first AND/OR tree expansion

of four basic layers. The first layer consist of files and other semantic data used to describe functional and QoS features if available services using the ontology scheme presented in [7]. Second layer consist of the searching function that is able for semantic discovery of individual services that takes into account QoS constraints, and JENA API [14] was selected. The third layer consists in creation of search space representation and searching techniques for finding composition services. The Java API JGraphT [15] primitives and generic infrastructure for graphs has been adopted to represent AND/OR trees and to implement the searching approach explained earlier. Finally, the fourth layer is the user interface through which are entered the QoS constraints and displayed the result.

VII. Conclusion and Future Works

The AND/OR tree represent an elegant formalism to express the Grid service composition problem with QoS constraints. The model has high expressiveness, which consequently allows addressing QoS driven service composition from different perspectives. Constraints can be expressed in all shapes, sizes and flavors. In addition, different searching techniques can be applied to find composite services that fit to complex user requirements. Combined with semantic annotations used to describe functional and non – functional features of services, it provides flexible infrastructure for composition oriented Grid service discovery.

Service composition systems except being effective, they must be able to find composite services in a reasonable time. In effect, this depends on the underlying implementation details, such as the data structures used, searching techniques and cleverness how to combine them in an effective way. Moreover, the evaluation of the efficiency should be conducted on clear benchmarks. In absence of widely accepted benchmarks, the evaluation turns out to be difficult process. Therefore, our future work will be mainly focused on investigating and developing efficient, flexible data structures and searching techniques that address semantic composition discovery of Grid services based on AND/OR graphs not effectively but efficiently as well, and compare them with other approaches on clear benchmarks. The construction of a friendly user interface would also contribute as an improvement.

References

[1] Web Service Resource Framework. Available online at: http://www.oasis-pen.org/committees/tc_home.php?wg_abbrev=wsrf (Last accessed: October 2010).

[2] J. Rao and X. Su. A Survey of Automated Web Service Composition Methods. *In Proceedings of the First International Workshop on Semantic Web Services and Web Process Composition*, SWSWPC 2004, San Diego, California, USA, July 6th, 2004.

[3] Liang, Q.A. and Su, S.Y.W. AND/OR Graph and Searching Algorithm for Discovering Composite Web Services. *International Journal of Web Services Research, vol.2, no.4, pp.* 48 – 67, 2005.

[4] Gu, Zh., Xu, B., and Li, J. Inheritance – Aware Document – Driven Service Composition. *CEC/EEE '07,* p. 513, Tokyo, Japan. IEEE Computer Society.

[5] Yan, Y., Xu, B. and Gu, Z. Automatic Service Composition Using AND/OR Graph. *10th IEEE Conference on E – Commerce Technology and The Fifth IEEE Conference on Enterprise Computing, E – Commerce and E – Service. p.* 335. *2008*

[6] UDDI Specification v.3.0.2. Avaliable online at: http://www.oasis-open.org/committees/uddispec/doc/spec/v3/uddi-v3.0.2-20041019.htm (Last accessed: October 2010).

[7] WSDL Spec. Avaliable online at: http://www.w3.org/TR/wsdl, March 2001. (Last accessed: October 2010).

[8] SOAP Spec. Available online at: http://www.w3.org/TR/2007/REC-soap12-part0-20070427/,

[9] OWL – S: Semantic Markup for Web Services. Available online at: http://www.w3.org/Submission/OWL-S/ (Last accessed: October 2010).

[10] WSMO: Web Service Modeling Ontology. Available online at: http://www.wsmo.org/ (Last accessed: October 2010).

[11] Zenuni, Xh., Ismaili, F. and Raufi, B. Ontology Design and Development for Grid Services. In Proceedings of the Fourth International Conference of Information Systems and Grid Technologies. pp. 105-114. 2010.

[12] Nilsson, N. Problem Solving Methods in Artificial Intelligence. McGraw – Hill 1971.

[13] Martelli, A., and Montanari, U. Optimizing Decision Trees Through Heuristically Guided Search. Commun. ACM, vol 21, no 12, pp. 1025 – 1039, 1978.

[14] JENA: A Semantic Web Framework for Java. Available online at: http://jena.sourceforge.net/ (Last accessed: October 2010).

[15] JGraphT: A Free Java Graph Library. Available online at: http://www.jgrapht.org/ (Last accessed: October 2010).

Customer Concerns in Telecommunications Contact Centers

Information and Systems Factors

Mike Hart, Grant Thavarajoo, Kgolane Thobejane
Department of Information Systems
University of Cape Town
Cape Town, South Africa
mike.hart@uct.ac.za, grant.thavarajoo@googlemail.com, thbkgo001@uct.ac.za

***Abstract*—The contact center is a potentially strategic touch-point between consumers and organizations, but often the service obtained leaves much to be desired. Effective use of the information and technology in our digital society appears to have many inhibitors – organizational, technological and human. This study analyses some 2150 incidents regarding telecommunications service providers reported by customers to a South African consumer portal. Their main concerns were information related issues: inadequate communication channel integration, limited functional integration and customer information inaccuracies. System issues comprised system faults, system limitations and system unavailability. Compounding this in fewer cases were non-information service-related issues: incompetent or rude staff and lengthy delays. Although telecommunications products and contact center technology may be advanced examples of the digital age, unless there is sound management, organization and motivation, customers may not receive the service they expect.**

Keywords-contact center; information systems; channel integration; telecommunications; customer service

I. Introduction

Earlier telephone-based call centers were generally designed to save costs, but as they have moved to become contact centers or more optimistically, customer care centers, they have found themselves occupying an increasingly strategic role [1, 2, 3]. To satisfy today's digital society, they offer customers contact through email, text messaging (SMS), on-line self service, interactive voice recognition (IVR) self service, fax, chat, blogs, Twitter and Facebook amongst others, and have become the first line of customer contact for many organizations [1, 2, 4, 5]. Their management is sometimes torn between satisfying financial and performance criteria such as calls answered per agent per day, measures of customer service such as first call resolution (FCR), and surveys of consumer satisfaction [3, 6]. With the increase in available technology and channels of communication has come an increased need for co-ordination and management of customer information [7]. As consumers become more digitally aware and confident, their expectations of quality service rise accordingly.

Internationally, in many developing as well as developed countries, there has been major recent growth in the use of cellular or mobile phones, for both personal and business use. This has been demand-driven and clearly fulfils a need. Yet the telecommunications industry in many countries attracts a relatively large number of complaints. A recent industry survey of US contact centers [8] shows that TMT (technology, media and telecommunications) was the industry grouping with the highest percentage of inbound calls that were complaints (27%). In South Africa the percentage of complaints about telecommunications service providers is similarly well above the average for all industries. This suggests that while society is being provided with a wider range of digital options, the management and coordination of those still leaves much to be desired.

This raises the research question: *"Given the wide range of available digital and technology options, what are the inhibitors to effective and efficient customer service in telecommunications cost centers?"*

As part of a larger study into customer service in contact centers in South Africa, over 2000 customer related complaints to a consumer portal about their telecommunications service providers were analyzed to determine what the major concerns were. South Africa has recently expanded its service offerings as an offshore business process outsourcing (BPO) and contact center destination, and it is most important that industry levels of service are maintained and improved [9, 10, 11].

This paper continues with a brief summary of some relevant background aspects, and then discusses the research methodology adopted for this study. Following that, an analysis of incidents in telecommunications contact centers uncovers the main themes. These are briefly discussed in relation to earlier literature, and the paper then concludes and makes some practitioner and research recommendations.

II. Background

Telephone-based call centers have existed for more than two decades, their initial primary aim being a cost-saving centralization of the organization's external communications. They allowed for inbound customer calls as well as outbound calls, often for marketing purposes. With the growth of the digital society many more communication options have become available, and call centers have become contact centers, sometimes hopefully also referred to as customer care centers [1, 7, 12, 13]. They may offer customers additional communication media such as email, SMS, on-line self service, IVR self service, fax, chat, blogs, Twitter and Facebook [13].

At the same time contact centers have become a key touch-point for many organizations [2, 14]. The interaction experienced by consumers has become increasingly

ISBN: 978-1-61208-116-8

important in determining whether they maintain an ongoing relationship with the organization, or whether there is “customer churn” [7]. Company management should therefore be allocating the necessary resources and efforts to their contact centers.

For management and their employees (agents and supervisors) there is however often a tension between the motivation of cost saving, and those of customer service and satisfaction [2, 3, 6, 12]. Well motivated, trained and satisfied agents are critical to quality service [1], and customers look to them for empathy, assurance and responsiveness [13].

Contact centers use a wide range of technology, including automatic call distributors (ACD), IVR and speech recognition, call recording, PBX, auto-diallers, workforce optimisation, helpdesk systems, telephony soft-switching, and workflow. Recently there has been a strong move to IP and VOIP [5, 14, 15]. Agents or consultants need information on customers, their products, transactions and past communications with the organization. This is provided (with varying degrees of agent access) by customer databases, knowledge management, customer relationship management (CRM) and ERP systems [5, 16]. While CRM has been pushed as a “technology silver bullet” for managing customer information and relationships, its practical success has been limited. “*it is important to consider where the company might have failed to address the factors most closely associated with CRM success — people, processes and day-to-day customer management activity, while over-investing in the factors which are, after all, only CRM-enabling systems and data*” [17, p. 349].

The variety of communication channels now available creates problems of information management [18, 19, 20]. Multi-channel integration is more talked about than practiced: a 2009 global survey of 554 contact centers found that less than 22% of centers had non-voice channels integrated into their universal queue [5].

A relatively recent phenomenon is that of contact center outsourcing and offshoring [21]. This has primarily been for cost-saving reasons, but over time companies have realized the risks of considering cost alone if, for example, the overall image of the organization will suffer due to inexperienced, inefficient or “incompatible” contact center agents communicating with their customers. First conducting a SWOT analysis of the potential outsourcing situation can be beneficial [11].

A. South Africa in the Digital World

South Africa’s mobile cellular subscriptions increased from 16.8 million in 2003 to 45 million in 2008, and there is now effectively one mobile phone for every person in the country. At the same time the number of fixed lines decreased slightly to 4.3 million [22, 23]. South Africa has the highest ICT Development Index (IDI) in Africa for Skills, coming 3rd after Seychelles and Mauritius for the IDI Access and Use indices [22]. South Africa is rated 9th of 139 global countries by the World Economic Forum for its financial markets and banking systems, 35th for its firm-level technology absorption, but only 76th for overall technological readiness [24], and is often referred to as a mix of 1st and 3rd world economies.

III. Research Methodology

Following recent research into similar issues in the banking sector, this study focused on telecommunications service providers, mainly of mobile or cellular phones. The research was exploratory, aiming to uncover the main information and systems-related concerns of customers dealing with telecommunications contact centers. The consumer portal HelloPeter [25] has in past research [26] proved a useful source of information for customer-related incidents. This portal enables consumers to register online, and then anonymously post a comment (complaint or compliment) in connection with service received from a supplier. Companies who register (for a fee) are able to obtain the contact details of the customer, follow up with them, and respond publicly on the website if they wish. Based on the response received, the customer then has the option of giving their feedback on the supplier’s response and their rating as either “over the moon”, “quite impressed”, “indifferent”, “unimpressed” or “utterly disgusted”. There are currently over 1200 companies registered on the site (up from 650 two years ago), and over 11,000 comments are received per month. On average about 20% of comments are compliments, and the balance are complaints (fairly expected due to human nature). However, amongst the cellular phone providers, the percentage of compliments is typically only between 5% and 12%.

While almost all South Africans are cellular phone owners, far fewer use the Internet. Posting HelloPeter [25] incidents requires on-line access, and the motivation to submit to this website, and so the set of incidents cannot necessarily be considered to be representative of the full South African consumer community. It does however describe a large number of actual incidents experienced by digitally aware telecommunications customers, and an opportunity to gain an idea of the variety and extent of information and system-related consumer issues in contact centres. Incident details posted can be viewed by everyone, without passwords, and in addition the founder and owner of the site has given the author permission for the data to be used for research purposes.

Over 2100 postings between November 2009 and March 2010 to South Africa’s six main telecommunications service providers (90% from four mobile and 10% from two landline operators) were selected for analysis. While a largely qualitative approach was adopted in order to explore the incidents, there were some useful quantitative measures available to give context. Being exploratory, the approach was inductive and cross-sectional, and thematic analysis [27, 28] was used to iteratively derive major themes and subthemes. Incidents were initially screened to ensure they all concerned contact center communication, information, technology and/or systems issues (others were excluded from analysis). Data was extracted from HelloPeter [25] to an Excel spreadsheet, and then imported to NVivo 8 CAQDAS software. A six step iterative process [29] for identifying, analyzing, coding and grouping nodes and themes was followed until the final themes and subthemes shown in Figure 1 emerged. These are discussed in the following section.

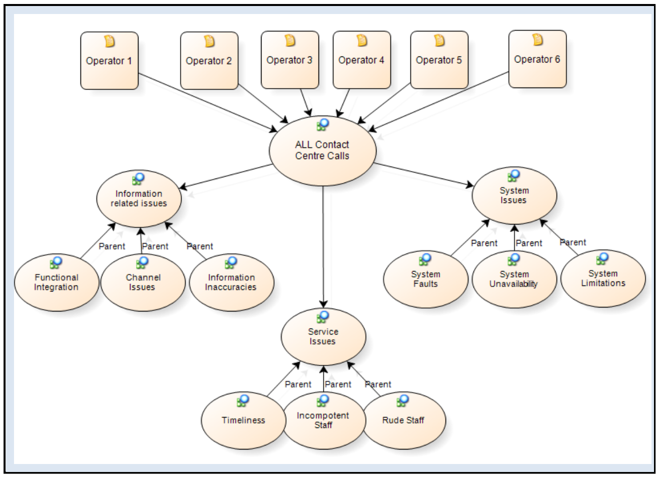


Figure 1. Themes and subthemes.

IV. Analysis of Incidents

The two main themes that emerged were Information-related Issues and Systems Issues. In addition certain Service Issues were often mentioned as compounding features of the incident description, and these have been included in a separate theme. All three were further divided into sub-themes. Each emergent theme and subtheme is now briefly discussed, with limited supporting quotes due to space considerations. It should be noted a few issues were often mentioned during the course of an incident, and many incidents were therefore coded under more than one node. These reflected the perceptions of consumers, whose assessment of the underlying cause might sometimes be inaccurate (although supplier response would often clarify matters). Numbers of incidents reflecting each theme is shown merely to indicate prevalence in the sample, and should not be generalized further.

A. Information Related Issues

1) Channel Issues: multiple communications channels used, but often limited channel integration (1268)

In most cases, customers stated the channel(s) of communications they used to interact with the contact center, including (in order of utilization): Phone, Email, SMS, Website and Fax. Very few used the post, whilst SMS's were primarily used for customers 'to get in touch' with agents with whom they had previously made contact. All complaints where multiple channels were used were coded here. The variety of digital communications media provides consumer flexibility, but demands efficient channel management and integration. Yet channel issues were the major cause of customer dissatisfaction. Agents lacked a 'single view' of customer communication, which greatly limited their ability to resolve incidents. Multi-channel integration was often limited or non-existent, and information sent via one channel appeared to have disappeared into a black hole when customers contacted the organization via a different channel. Some incidents showed repeated use of multiple channels to assist with queries.

> *"...I phoned them, faxed a letter and e-mailed a letter of cancellation. I then phoned to make sure they received it and that it would be cancelled. I was assured that it would be. I then received an sms to confirm this, only to wake up this morning and receive an sms that my airtime was topped up. When phoning them I was told that it still shows as active."* (Q1).

> *"...Called last Thursday and was told to send the necessary letter head with request which I did on Friday. Come Monday I called and was told they had not received it - so I sent it again. Tuesday I called them 4 times!!!!! I eventually had to email them to a X. He said he sent them at 11h00 and it would take about 2 hours to get them. ..... Well one week later I have still not received a thing.....and I am yet to get a call back from any of the agents I spoke to" (Q2).*

These types of interactions lead to a breakdown in communication, and frustration with the operators, and could be caused by bad processes, carelessness and human error, as well as poor channel integration.

2) Information inaccuracies or unavailability of information (593)

The second most common subtheme was inaccurate information stored on the operators' information systems. This relates to the ability of the agent to offer correct information to the customer and have the necessary (correct) information available to assist the customer with their query.

> *"I applied for a new contract with X on the 25th of February 2010 and my contract was approved on the 26th of February 2010. However, due to a mistake in the database records in X (phone description in database did not match phone available in stock), I was only able to obtain my phone on Tuesday 2nd March 2010" (Q3).*

In rare cases the user may not necessarily have been negatively affected by the original issue, but the information accuracy may still lead to unexpected results:

> *"After the second month I realised that they still had not taken the first 2 payments off and I also never even received an invoice. I called to see what the problem is and they said they don't know but will instruct the account department to take the amount immediately... I asked the account department to call back, they never did. I called again a week after and I said to the guy I have never heard of a company that the client actually needs to chase to take the money off his account... Yesterday, my service was suspended" (Q4).*

Information accuracy can be a function of poor processes or data quality management. Information unavailability may indicate that agents are not given access to certain systems, or do not know how to access it.

3) Functional Integration: information not integrated across organisational units or functions (256)

The different organizational functions of telecommunications operators can include billing departments, performance management, retail stores as well as other partner sites. A query handled by different departments in an organization was often devoid of supporting information captured at previous points of contact. This often results in users being subjected to numerous call transfers in order to locate the resolving party, for example:

> *"I've been trying to get through to the X upgrades department since last week Thursday already. I've been dialling the call centre number. Most of the time I get cut off and if I do get to speak to a consultant I then get cut off when they try to put me through to the relevant department, at one stage I was transferred 3 times, held on for ages each time only to get cut off yet again." (Q5).*

This highlights the difficulty users can endure when being transferred to alternate departments or divisions instead of the initial agent resolving their call. Functional integration issues may lead to customer dissatisfaction either through lack of information integration within systems or due to restrictive company protocol:

> *"Phoning X resulted in nothing - apparently the admin department requested legal department to close the account. Legal referred me to admin and vice versa. I went to the nearest branch (X), only to be told that the accountholder (deceased) must sign to transfer the number!!!! Eventually they insisted that I fill in forms and supply bank statements, even though I am a customer for the last 10 + years!" (Q6).*

A common cause for functional integration issues can be incompatible systems used between different business units, and unimaginative business processes. Lack of co-ordination and distributed systems, and limited technological infrastructure may also hinder the ability for an organization to share its information.

B. System Issues

1) System Faults: technical faults or errors experienced during interaction with the customer and the agent's information systems (527)

In many cases customers complained about the technical state of the contact centers' hardware or software and its impact on the resolution of the query.

> *"They promised that this would all be corrected and the money I paid for the services would be refunded - fat chance as I am still waiting for this! In addition my contract would be cancelled on the 5 January 2010. Funny how I ended up getting a bill for February!!! this month I applied for finance on a car I wanted to get only to be told that I have a negative credit score and that no one will finance me - X blacklisted me for their error!!!! Useless." (Q7)*

In situations like these, customer dissatisfaction is further compounded and could result in subscriber churn.

2) System Limitations: restrictions or policy directives that prevent the system behaving in a manner that the customer expects (417)

System limitations refer to the lack of perceived functionality of information systems to address customer complaint resolution. In particular several cases emerged that highlighted policy directives or the inability of the information system to address the customer query. The statement below is a case in point:

> *"I called in several times requesting my bill via email and every time they sent me some link to download and I keep trying to tell them I can't download it please sent it via pdf or something and when I check it's the some link again. Friday I spoke to 10 different people, the bill has 8 pages she sends 1 page is this a joke there's something called 1st call resolution ..."(Q8)*

3) System Unavailability: downtime associated with contact centre information systems (464)

Customers expect that information systems will be available during their interaction with the contact center or its related functions. Failure may be due to inefficient technical infrastructure or resources and in many cases leads to severe inconvenience and dissatisfaction.

> *"Apart from the fact that they were offline and couldn't accept payment and I had to go to Y to pay my cellphone account for which they felt no need to apologize..... In 5 months time my contract comes up for renewal. It will be insane to renew with X." (Q9)*

In some cases users are also unable to make contact due to excessively long waiting times:

> *"I tried calling back that night and they were offline so could not help me once again. I then tried calling again from yesterday morning and my calls have gone unanswered. I have held on for an average of 30 minutes each time I call (before I hang up due to frustration!) and no one has answered the phone" (Q10).*

Issues of this nature demonstrate the mismatch between expectations of the user and the ability to offer customer service by the contact centre.

C. Service Issues

1) Rude Staff: dissatisfaction from customers interacting with rude or bad mannered agents (199)

In addition to the above systems and information-related problems, a few customers specifically mentioned perceived rudeness of behavior or attitude to them by the agent. This usually compounded the negativity of the complaint. Reasons included perceived unwillingness to hear the customers' side of the story, or to escalate a problem to more senior staff, and a lack of respect or empathy.

2) Incompetent Staff: agents that cannot perform their expected functions effectively (183)

Examples included failure of agents to perform tasks properly, e.g., form submissions, return phone calls, valid explanations etc. A common complaint was the poor response from contact center agents to requests to diagnose technical issues with the customer's devices.

3) Timeliness: ability of the agent to address, respond to, or follow up a customer complaint timeously (143)

Customers expect queries, and their follow ups, to be handled in a timely, effective manner. However there were a number of incidents where operators took exceptionally long to address or resolve queries. In many cases the delay was attributed to system disruptions or agent incompetency, as well as functional integration issues, where calls had to be routed to another agent for resolution.

These results will be discussed in the following section.

V. Discussion

The dominant themes that telecommunications customers complained about concerned incorrect, missing or delayed information about themselves and their past communications with the organizations. It was very clear that use of more than one communication channel often caused problems due to a lack of information integration. The challenges of multi-channel integration mentioned in [18, 19 and 20] were clearly not adequately addressed in most cases. Different channels may be based on varied technologies and generally appear to be managed separately, if at all. CRM software, while dealing with some customer information, does not seem to be integrating it satisfactorily across different channels. A common consequence is multiple submissions or requests for customer information, with resultant delays and customer irritation.

Ensuring correctness of information should be part of a general drive for data and service quality in any organization [1, 13], and examples show there is clearly room for improvement here. One of the other major causes of consumer unhappiness was the inability of the agent to achieve first call resolution [3], for a number of reasons. One may be that the organizational systems themselves are not adequately integrated – the main aim of ERP software. Alternatively, the agent may only be given access to certain modules, processes or functions. There can be valid situations where the agent is required to obtain information from another person, or where the request needs to follow a controlled process through other business units. This then needs to be well managed within the organization, with one person taking responsibility for resolution of the call, processes being re-engineered if necessary [15], and not sending the customer from pillar to post and cutting them off in the process. Ideally the agent should have access to as much customer-related information as possible, through a user-friendly portal, and receive sound training in the effective use of ERP, CRM and knowledge management software [1, 2, 6, 7, 15, 16, 17].

Many customers also experienced system issues – faults, limitations and unavailability. Because these are "customer-facing" systems they deserve above-average attention, with meticulous design and meaningful service level agreements [5]. If they cannot operate and be available to service customers 24x7, this should be clearly publicized to customers, and top management should be involved in decisions of this nature.

The first two service issues – incompetent and rude staff – are largely a function of agent training (classroom and on-the-job), coaching, mentorship and motivation. Agents need an empowering environment [2, 6], with supportive follow-up and quality monitoring of their communication. Apart from training in product, organizational, systems and process knowledge, agents may require additional language, communication, life skills, math or computer literacy skills. Training in "customer care", empathy [13] and customer responsiveness is currently recognized to be of major importance.

The final service issue, timeliness, was only coded in more serious cases, but is the end result of most of the other problem areas noted, and a major customer irritant. With proper measurement processes and criteria in place [3, 5], this should be greatly improved.

It should be noted again that these incidents should be regarded as illustrative, and not necessarily representative.

VI. Conclusion

The digital society is making wider use of a range of communication channels. As contact centers provide more of these for their customers they need to ensure that customer communications are effectively and efficiently managed and fully integrated. There is no point giving customers the "channel of their choice" if their past communications cannot be synchronized with others received and sent through the contact center. Customers also often require information from a number of corporate systems. In theory this should be facilitated by ERP and CRM software, but in practice some systems are not cleanly integrated or may not be accessible by the agent, and the concept of a "single view of the customer" remains very elusive for many organizations.

The telecommunications industry might be expected to have a relatively strong ability to cater effectively for customers in the digital society. For whatever reason, possibly due to its rapid growth and regular technological change, it seems to engender an above average percentage of customer complaints, and come short in a number of technical, information and systems areas. It is interesting that the local service provider with the smallest percentage of complaints is one which has had to "play catch-up" to the two much bigger established players (granted licenses much earlier), and currently appears to be making a concerted, and reasonably successful, effort to use customer service as a competitive weapon.

Telecommunications companies should learn from their past interactions with customers [17] and use technology such as CRM more effectively in future. But they should also realize that increased digital opportunities imply increased organizational and managerial effort and resources, and adopt a stronger customer focus.

Further research could be carried out to obtain the perspective of the telecommunications organizations and the challenges they face in supplying better customer service. It could also examine customer concerns with telecommunication contact centers globally.

ACKNOWLEDGMENT

The financial support of the South African National Research Foundation is acknowledged.

REFERENCES

[1] R. Kantsperger and W. H. Kunz, “Managing overall service quality in customer care centers: Empirical findings of a multi-perspective approach”, *International Journal of Service Industry Management*, vol. 16, no.2, pp. 135-151, 2005.

[2] H. J. Richardson and D. Howcroft, “The contradictions of CRM – a critical lens on call centres”, *Information and Organization*, vol. 16, pp. 143-168, 2005.

[3] M. L. Hart, B. Fichtner, E. Fjalestad, and S. Langley, “Contact centre performance: In pursuit of first call resolution”, *Management Dynamics*, vol. 15, no. 4, pp. 17-28, 2006.

[4] Z. Aksin, M. Armony, and V. Mehrotra, “The Modern Call Center: A Multi-Disciplinary Perspective on Operations Management Research”, *Productions and Operations Management*, vol. 16, no. 6, pp. 665–688, 2007.

[5] Merchants, Dimension Data’s Global Contact Centre Benchmarking Report 2009. Gauteng, South Africa: Dimension Data, 2009.

[6] M. L. Hart, Y. H. A. Chiang, and M. Tupochere, “Balancing agent performance and customer service in contact centres”, *Journal of Contemporary Management*, vol. 6, pp. 151-169, 2009.

[7] M. Meltzer, “A customer relationship management approach: integrating the call centre with customer information”, *Journal of Database Marketing*, vol. 8, no. 3, pp. 232-243, 2000.

[8] ContactBabel, US Contact Center Decision-Makers' Guide (2009/2010 - 3rd edition), 2010. Retrieved December 17 2010 from http://www.contactbabel.com.

[9] C. Benner, “South Africa on-call: information technology and labour market restructuring in South African call centres”, *Regional Studies*, vol. 40, no. 9, pp. 1025-1040, 2006.

[10] Deloitte and CallingTheCape, Contact Centres and Business Process Outsourcing in Cape Town: 2007/2008 Key Indicator Report. Cape Town: Deloitte Touche Tohmatsu, 2008.

[11] S. Derakhshani and M. L. Hart, “Outsourcing Contact Centres to a Developing Country”, Proc. 16th Americas Conference on Information Systems (AMCIS), 12-15 August 2010. Available at http://aisel.aisnet.org/amcis2010/2.

[12] R. Batt and L. Moynihan, “The viability of alternative call centre production models”, *Human Resource Management Journal*, vol. 12, no. 4, pp. 14-34, 2002.

[13] P. Ramseook-Munhurran, P. Naidoo, and S. D. Lukea-Bhiwajee, “Employee perceptions of service quality in a call centre”, *Managing service quality*, vol. 19, no. 5, pp. 541-557, 2009.

[14] H. Bernett and M. Jaramillo, “Assessing web-enabled call center technologies”, *IT Professional*, vol. 3, no. 3, pp. 24-30, 2001.

[15] I. J. Chen and K. Popovich, “Understanding customer relationship management (CRM) People, process and technology”, *Business Process Management Journal*, vol. 9, no. 5, pp. 672-688, 2003.

[16] M. Xu and J. Walton, “Gaining customer knowledge through analytical CRM”, *Industrial Management & Data Systems*, vol. 105, no. 7, pp. 955-71, 2005.

[17] L. T. Wright, M. Stone, and J. Abbott, “The CRM imperative - Practice vs theory in the telecommunications industry”, *Journal of Database Marketing*, vol. 9, no. 4, pp. 339-349, July 2002.

[18] H. Cassab, “Investigating the dynamics of service attributes in multi-channel environments”, *Journal of Retailing and Consumer Services*, vol. 16, pp. 25-30, 2009.

[19] S. A. Neslin, D. Grewal, R. Leghorn, V. Shankar, M. L. Teerling, J. S. Thomas, and P. C Verhoef, “Challenges and opportunities in multichannel customer management”, *Journal of Service Research*, vol. 9, no. 2, pp. 95-112, 2005.

[20] A. Payne and P. Frow, “The role of multichannel integration in customer relationship management”, *Industrial Marketing Management*, vol. 33, pp. 527-538, 2004.

[21] D. Holman, R. Batt, and U. Holtgrewe, The global call center report: international perspectives on management and employment, 2007. Retrieved December 17, 2010, from http://www.ilr.cornell.edu/globalcallcenter/

[22] International Telecommunication Union, “Information Society Statistical Profiles 2009: Africa”, ITU: Geneva, 2010

[23] S. Esselaar and A. Gillwald, 2007 South Africa Telecommunications Sector Performance Review. University of Witwatersrand, 2007. Retrieved December 17, 2010, from: http://link.wits.ac.za/papers/TSPR2007.pdf

[24] World Economic Forum, “The Global Competitiveness Report 2010-2011”, Editor: Professor Klaus Schwab, WEF: Geneva, 2010.

[25] www.HelloPeter.com

[26] M. L. Hart, M. Mannya, and D. New, “Customer Information Concerns in Banking Contact Centres”, Proc International Conference on Information Management and Evaluation, (ICIME’10), Cape Town, 25-26 March 2010.

[27] J. Fereday and E. Muir-Cochrane, “Demonstrating rigor using thematic analysis: A hybrid approach of inductive and deductive coding and theme development”, *International Journal of Qualitative Methods*, vol. 5, no. 1, pp. 1-11, March, 2006.

[28] D. R. Thomas, “A general inductive approach for qualitative data analysis”, University of Auckland, Auckland, New Zealand, 2003. Retrieved December 17, 2010 from http://www.fmhs.auckland.ac.nz/soph/centres/hrmas/_docs/Inductive2003.pdf

[29] V. Braun and V. Clarke, “Using thematic analysis in psychology”, *Qualitative Research in Psychology*, vol. 3, pp. 77-101, 2006.

Motivating the University Teachers to Involve in e-Learning through Engagement in Information Technology

Dershing Luo
Department of Information Management
National Taiwan University
Department of Information Management,
China University of Science and Technology
Taipei, Taiwan (R.O.C.)
dsluo@cc.cust.edu.tw

Jen Wel Chen
Department of Information Management
National Taiwan University
Department of Business Administration
Chinese Culture University
Taipei, Taiwan
jwchen@sce.pccu.edu.tw

Ching Cha Hsieh
Department of Information Management
National Taiwan University
Taipei, Taiwan (R.O.C.)
cchsieh@im.ntu.edu.tw

***Abstract*—There is already a large body of literature on the e-learning (electronically supported learning and teaching). However, the incentive to encourage university teachers to involve in e-learning is still need to be studied earnestly. This study aims to discover the incentive problems that hinder university teachers developing and using the e-learning systems. E-learning is promoted in most countries prosperously now. However, it is still hard to find schools that can use e-learning effectively in helping their teachers uses. Therefore, there are various perspectives of incentive that need to be studied. For example, what are the motivating factors that can encourage teachers to use this new method of teaching? What techniques can help teachers develop an e-learning system? What policies can facilitate the process of e-learning? Based on a pilot study, we employ a depth interview to find out the answers of the questions mentioned above. In general, it is not easy for teachers to implement e-learning courses by themselves. School administrators and governments need to set policies, build facilities, and find motivators to help teachers use e-learning techniques to improve their teaching methods, and hence help students learn in an easy and convenient way.**

Keywords-information technology; incentive; value; e-learning

I. INTRODUCTION

The Massachusetts Institute of Technology (MIT) established the open course ware (OCW) in 2001. All the courses in MIT can be accessed on the web for the public to learn free in 2010. This new initiative reflects MIT's institutional commitment to disseminate knowledge across the globe. Nowadays, a growing number of universities promote e-learning. According to Elliott Masie's (The Masie Center) definition, e-learning is the use of information and network technology (IT) to design, deliver, select, administer, and extend learning. Furthermore, Cisco Systems defines e-learning as "internet enabled learning, components can include content delivery in multiple formats, management of the learning experience, and a networked community of learners, content developers and experts. E-learning provides faster learning at reduced costs, increases access to learning, and clear accountability for all participants in the learning process." In general, e-learning takes place over the internet rather than in a physical classroom. Moreover, some e-learning may combine a portion of traditional classroom teaching. The development of e-learning in higher education is now a hot topic for most of the universities. However, there are many problems that still hinder teachers to implement e-learning courses in their colleges.

Despite the wealth of studies on technology and education, questions about the incentives for teachers to use e-learning remain unaddressed. In one much-cited research commentary, Alavi and Leidner call for increased research on technology-mediated learning (TML). They recommend that researchers explore "the explicit relationships among technology capabilities, instructional strategy, psychological processes, and contextual factors involved in learning" [1]. How to encourage the teachers to use the e-learning smartly is an urgent question.

E-learning has also been usually included in the lists of school evaluation items. Therefore, school administrators believe that e-learning is a trend that can not be ignored. Currently, a lot of universities have distance-learning centers. However, only a few of them have successful e-learning programs. People may ask is it difficult to implement e-learning programs? We argue that it depends on how it is done. At first, let us consider the problems that hinder most colleges. Based on our pilot interview with several school administrators, e-learning is still focused on the few teachers who have the necessary computer skills to develop their own e-learning courses. For most teachers, this is a difficult process even though the IT is friendlier than ever. Secondly, although it seems that colleges have already found some ways of encouraging teachers, these motivators, however, are not strong enough to motivate teachers sufficiently. For example, most colleges pay the teachers US$300 to 1,500 for each course to encourage them. For those teachers who have the ability to do so, the subsidies are not their main concern. For those teachers without the ability, they are unable to participate. Therefore, the problems that we have mentioned above are related to people's readiness and ability.

There are many problems that are related to people factors. The faith of teachers [5], the value system, achievement, self actualization, role changing, technology affordance, and even the culture are critical factors [3, 5, 9]. Especially, although most school administrators believe that e-learning is critical to their school's performance. But they do not have policies required to support their actions. Without these policies, there is no way of distinguishing the difference between teachers who have developed e-learning courses and those who have not. When this situation occurs, there will be no one who is willing to use e-learning to teach. Assuming that schools have the policies to support their actions, school administrators will find other problems are easier to counter comparatively.

The structure of this study is as follows: Section 2 briefly depicts the theories of motivation. Section 3 describes the method used. Sections 4 and 5 show the results and suggestions for involving teachers in e-learning. The conclusion is made in the last section.

II. THEORY DEVELOPMENT

Through comparing the motivation theories provides a well development of the questionnaire needed for a former pilot survey. There are three perspectives of motivation.

(1) Content perspective: Motivating people from internal motivation, such as hierarchy of needs from Maslow [8], ERG theory from Alderfer (existence, relationship, and growth) [2], theory X and theory Y [11], dual factor theory from Frederick Herzberg (hygiene factors and motivators factors) [6], and McClelland Achievement Motivation Theory (need for achievement, need for power, and need for affiliation) [9].

(2) Process perspective: Emphasize the internal reaction process or factors influence one's efforts. The alike theories are equity theory by Adams, Goal-setting theory from Edwin Locke, Expectancy theory by Victor Vroom (expectancy, instrumentality, and valence), and reinforcement theory by Skinner.

(3) Integrated perspective: Emphasize both the internal and institutional factor, also give a vivid description of the impact process and relationship of human's effort behaviors. Such theory as Porter and Lawlerthe's integrated expectancy model is depicted in Figure 1.

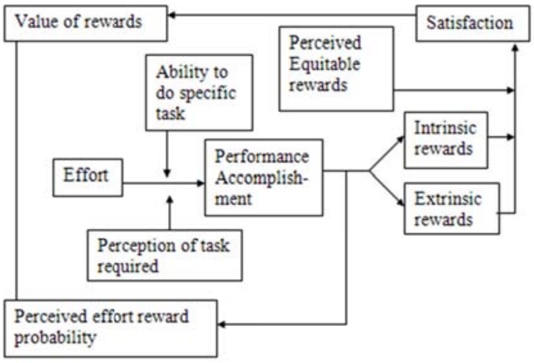

Value of rewards
Satisfaction
Perceived Equitable rewards
Ability to do specific task
Intrinsic rewards
Effort
Performance Accomplish-ment
Extrinsic rewards
Perception of task required
Perceived effort reward probability

Figure 1. Integrated expectancy model developed by Porter and Lawlerthe in 1968

Moreover, as [8] depicted, the salary is not the most important incentive. The first three incentives are good job development, better opportunity for education and self development, and the witness of self achievement.

In the construction of the questionnaire, the theoretical base mainly employed is the Porter-Lawler model of motivation [7]. It assumes that behavior is directed by a conscious expectation that we have about our own behavior leading to the achievement of a desired outcome. Therefore, there is no motivation, unless both of these conditions (expectation and value) are met. This theory redefines motivation as a cognitive, decision-making process through which the individual chooses desired outcomes, and sets in motion the actions appropriate to their achievements.

To sum up, the integrated expectancy model considers both the internal and institutional factor. It can be a more complete perspective when design an incentive system in universities. The pilot survey is based on the integrated expectancy model. This study is based on the results of the pilot survey; we want to achieve a better understanding the deep motivation when the above incentives are set.

III. METHODS

After the analysis of motivation theories, we developed the questionnaire in the pilot study. The questionnaire is divided into two groups: one is for the development of e-learning materials, the other one is for the use of e-learning system. Except the demographic data, the questionnaire concludes the following: internal reward, expectation of external reward, efforts and willingness, capability, awareness of role, work performance, feeling of internal reward, actual external reward, equity, and job satisfaction.

This study aims at an in-depth understanding on the incentives in developing and using e-learning systems. We use a questionnaire survey and in-depth interview in this study. Followed the survey, we interview the teachers and administrators in depth. Through the in-depth interview, we found some valuable incentives for encouraging university teachers to engage in e-learning. The subjects and survey in the pilot study are as follows.

Subjects: In the survey, the subjects are selected from ten universities at the northern part in Taiwan. There are totally 300 subjects.

Survey of questionnaire: We invited several experts and designed such factors to generate motivation. We have also carried out a pilot test in a selected institute to testify its consistency and validity. However, the questionnaire response rate is only 21% in the pilot study.

We use the in-depth interview in this study. Based on the result of the pilot study, we selected dozens of teachers to our non-structured in-depth interview. Moreover, we performed a semi-structured in-depth interview of chiefs from two colleges in the northern district of Taiwan. By doing so, we have a deep understanding of the relevant problems in developing and using e-learning system,

IV. Results

In the in-depth interview, we aim to compare the survey of experienced and inexperienced as depicted in Table 1. The results from the questionnaire survey are:

To those whom are experienced in e-learning, they still have doubt about the performance of e-learning. Both 'willingness' and 'role consciousness' were ranked low even though teachers are capable of developing and using e-learning systems. They expect the outer payments much, additionally, they are willing to develop and use e-learning systems due to their self expectations. So the incentives required by experienced teachers are that they are given more self-direction and enhancements.

To those whom are not experienced in e-learning, they also have doubts about its effectiveness. Due to the reason of 'willingness', 'role consciousness', and 'self-expectations' being ranked low, these results show that they are not ready to develop and use e-learning systems. So the incentive needed by inexperienced teachers is the role of "helper". That is they should be given encouragement and help but not only focusing on the financial incentives. In other words, through the various incentive mechanisms, if the teachers know their own limitations and abilities, then the development and use of e-learning systems may improve.

Table 1 The comparison of experienced and inexperienced in e-learning

Group	**experienced**			**Inexperienced**		
Variables	Mean	Min	Max	Mean	Min	Max
internal reward (7)	27.04	0	40	22.41	0	32
expectation of external reward (7)	31.85	0	42	27.18	0	32
efforts and willingness (7)	26.77	0	36	22.35	0	30
capability (6)	31.96	21	42	25.29	0	35
awareness of role (5)	19.35	10	30	14.47	0	24
work performance (10)	36.12	7	55	17.71	0	30
feeling of internal reward (3)	13.00	0	18	10.47	0	16
actual external reward (7)	26.27	0	42	20.75	0	39
equity (4)	12.77	0	20	9.81	0	20
job satisfaction (8)	25.15	0	43	23.06	0	40

p. s. the bracket after the variable is the number of items; the number 0 in each column means that respondent has no idea.

In the in-depth interview, we discuss each factor concerning by the teachers. Here, we discuss these factors divided into the following aspects.

In policy aspect: The results show that the policy is very important in the development and use of e-learning systems. E-learning curricula can be successfully completed only with strong policy. Once the policy is determined, further coordination will become easier. Take one of the interviewed colleges for example. Its policy for the e-learning of all general curricula in the first two years of college is determined. Hence other factors like salary, work load and evaluation are also formed. The obvious effect is that almost all teachers are urged to develop and use e-learning systems. This is quite different from the past.

In quality aspect: The results show that students benefit from those aspects such as schedules, times and ways of learning. Both complementary learning and enhancement learning are benefited. One of the interviewed colleges designed a new style of instruction—either style A (1/3 traditional instruction plus 2/3 e-learning) or B (2/3 traditional instruction plus 1/3 e-learning) depending on the information literacy of the teacher.

In motivation aspect: The results show that possible incentives are as follows: (1) Salary: rewards are developing fee, bonus, or over-time pay, while penalties are in the form of over-time works. (2) Work load: the common style is to adjust the teaching style or reduce the teaching hours of those who developed the e-learning system. (3) e-learning techniques: assistants or some orientation training may aid the teachers' willingness to develop and use e-learning systems. (4) Job: the accomplishment of the development of e-learning itself is a reinforcement of teaching. Some promotion may follow. (5) Rights or honor: more using rights or granting honor for those who develop and use e-learning systems would be a great encouragement.

V. Suggestions for Implementing E-learning System

Current studies noted valuable suggestions from various perspectives: student perspective in the teacher–student dynamics, teaching paradigm [10], and institutional environment [9], etc. This study echoes the requirements of information literacy [2, 4]. Based on this study and the experiences that we have accumulated in implementing e-learning systems at schools, some suggestions are listed below:

1. Setting e-learning policies: In order to create appropriate policies to support e-learning systems, schools should have the policies to distinguish the differences between teachers who have developed learning courses and those who have not. For example, the differences could be seen as financial award, job security, recognition, and annual evaluation.

2. Finding motivating factors: As we all know, someone is willing to do something that attracts him/her. Currently, most schools use financial means to motivate teachers. We can not deny that this is a motivating factor. However, it depends on how we use it. As we have mentioned, some colleges provide US $1,000 dollars as a cash award to motivate teachers. This is a one time cash award. However, there is another way that can provide a much more attractive reward for teachers. For example, when schools embrace e-learning as one of their teaching methods, this new technology can reduce their costs by lowering their teacher numbers. If the school can take part of the savings from every semester and give it to the teachers as their long term cash rewards, this will be a huge motivation for teachers to participate. Moreover, a lot of teachers are concerned about promotion. All teachers are climbing the ladder (from instructor, assistant professor, associate professor to professor). Therefore, there should be policies that can help these teachers climb higher. In general, there are many things that need to be done before teachers can really implement an e-learning system. Teachers, therefore, need to be motivated.

3. Building e-learning facilities: Most schools fail to implement e-learning systems because they are short of the relevant facilities for teachers to use in a simple and convenient way. Currently, most schools provide e-learning systems that allow teachers to use their computer to record and edit their materials. It seems to be the right way to do it because every school uses the same method. However, this is also the reason that most schools have the same problems when implementing their e-learning systems. As mentioned above, if schools use this method, then only the teachers who have the required techniques can produce e-learning materials. Under this situation, e-learning courses will be restricted and fail. However, there is another method that can solve this problem. One college has built studio rooms to help teachers produce their e-learning courses. In these studio rooms, college provides the entire editing (Pre and Post authoring) services from the time teachers work in the studio until the courses are uploaded to the system. This kind of service gives every teacher the opportunity to produce their e-learning courses. When all teachers have the ability to produce their e-learning, it can increase the chance of the e-learning process.

4. Providing training courses: Basically, when schools want to implement any program, relevant training is required for successful implementation of the program. For example, teachers need to know what specific skills, knowledge, and system requirements are needed. If schools do not have studio rooms to help teachers, then the training will focus on how to help teachers use their computers to put together e-learning courses. E-learning system is usually composed of two parts. One is to produce the courses; the other is to management the courses. Training classes should also include these two parts and beheld every semester since schools will have new teachers needing to be trained.

VI. Conclusion

The development of an e-learning system is an important policy in Taiwan’s higher education. Most schools are trying to make it. However, most of schools are facing the problem that only a few of the teachers have the ability to produce e-learning courses resulting in the fact that the quality of these courses is not as good as expected. Therefore, the Ministry of Education in Taiwan (MOE) has urged universities to unite their resources (equipments, courses, techniques, etc.) and figure out a better solution for this situation. Much research and many projects have been conducted and focused on this area. Our research found that if schools want to implement an e-learning system, they need to give all the teachers the ability to produce e-learning courses. If schools just provide an e-learning system and expect teachers to automatically pick up the system, the miracle of e-learning will not bear fruit. Therefore, as far as schools are concerned, policies, facilities, training, and the motivating of teachers are the key issues for implementing the system.

Acknowledgment

Part of this study is under grant NO 92-2511-S-157-001-X3 funded by National Science Council in Taiwan.

References

[1] Alavi, M. and D. E. Leidner (2001). "Research Commentary: Technology-Mediated Learning—A Call for Greater Depth and Breadth of Research." Information Systems Research 12(1): 1-10.

[2] Alderfer, Clayton P., An Empirical Test of a New Theory of Human Needs; Organizational Behaviour and Human Performance, volume 4, issue 2, pp. 142–175, May 1969

[3] Ali, Radwan and Irvin R. Katz (2010). Information and Communication Technology Literacy: What Do Businesses Expect and What Do Business Schools Teach? Educational Testing Service Research Report, ETS RR-10-17.

[4] Coppola, N. W., S. R. Hiltz and N. G. Rotter (2002). "Becoming a virtual professor: Pedagogical roles and asynchronous learning networks." Journal of Management Information Systems 18(4): 169-189.

[5] Juhdi, Nurita, Ahmad Zohdi Abd Hamid, and Mohd Saeed Bin Siddiq (2010). The Impact Of Technology On Job Characteristics And Internal Motivation: A Study Of Instructors In Institutions Of Higher Learning In Malaysia. International Journal of Arts and Sciences 3(14): 327 – 350.

[6] Herzberg, Frederick (1959), The Motivation to Work, New York: John Wiley and Sons

[7] Lim, C. P. and C. S. Chai (2008). "Teachers' pedagogical beliefs and their planning and conduct of computer-mediated classroom lessons." British Journal of Educational Technology 39(5): 807-828.

[8] Maslow, A. H. (1943). A Theory of Human Motivation, Psychological Review 50(4): 370-96.

[9] McClelland, D. C. (1987). Human Motivation. Press Syndicate of the University of Cambridge. New York.

[10] McFarlane, K. (2001). Just another Electric Circus? Meeting Standards for K to 12 E-Learning Classroom Resources. Education Canada. V.41 n.3 p.26-27 ERIC_NO. EJ643770

[11] McGregor, Douglas (2002). Theory X and Theory Y. Workforce; Jan 2002, 81 (1), p32.

[12] Porter. L. W. and Lawler, E. E. (1968). Managerial attitudes and performance, Dorsey Press.

[13] Sun, P. C., H. K. Cheng and G. Finger (2009). "Critical functionalities of a successful e-learning system - An analysis from instructors' cognitive structure toward system usage." Decision Support Systems 48(1): 293-302.

[14] Tahir M. Nisar, (2002) "Organisational determinants of e-learning", Industrial and Commercial Training, Vol. 34 Iss: 7, pp.256 - 262

[15] Tao, Yu-Hui (2008). Typology of college student perception on institutional e-learning issues – An extension study of a teacher’s typology in Taiwan. Computers & Education, 50 (4). pp. 1495-1508.

Storage-based Location Tracking in Mobile Networks

Madhuraj Mishra and Aditya Trivedi,
Indian Institute of Information Technology and Management, Gwalior, India
madhurajmishra@rediffmail.com, atrivedi@iiitm.ac.in

Abstract—**with the increasing number of mobile terminals, it is a challenge how to reduce the cost and provide fast and efficient call delivery to the mobile terminals. In the existing mobile networks, the call connection between the two terminals is based on the registration of their identity in the databases known as home location register and visitor location register. Conventional registration strategies will incur a high volume of signaling traffic. These strategies can work well up to a certain level of call to mobility ratio. In this paper, we propose a storage-based location tracking scheme based on the storage of the location of a mobile terminal at the repository database, which efficiently reduces the location updates and searching cost in the mobile networks.**

Keywords-***Repository; Visitor location register; Home location Register; Call to mobility ratio; mobile networks.***

I. INTRODUCTION

Mobile communication is one of the emerging fields in the area of communication. In the mobile networks, the location of a mobile terminal changes frequently so it is difficult to trace a mobile terminal's location. There are three main strategies proposed for location tracking in the hprevious work; these are local anchoring strategy (LAS) [1] [5] [12], IS-41[5], and group registration (GR) [1] technique for location tracking in mobile networks.

In the previous studies, several strategies such as local anchoring strategy (LAS), forwarding strategy (FS) [7] [11] [12][16] and replication strategy (RS) [17] were proposed to reduce the burden on the HLR. Since, a user may change his mobility patterns frequently, any single strategy can not cope with such time-varying mobility patterns efficiently. In this paper, an efficient strategy based on storage of information about the location of mobile terminal known as storage-based (SB) location tracking strategy is proposed which works well with a good level of call to mobility ratio (CMR). To reduce cost, we applied a repository database which will result in low cost and also have low complexity of database handling.

The remainder of this paper is organized as follows: In Section 2, we explain the SB strategy. This section also includes the details of location registration and updating procedure in SB strategy. In Section 3, we describe the analytical model. In Section 4, numerical results are described. Finally, conclusion is given in Section 5.

II. PROPOSED STRATEGY

A. Basic approach for SB location tracking

The proposed SB location tracking approach is based on the Group Registration Technique [1], which is modified to reduce the cost of mobile communication. In this strategy it is proposed to attach a database for storage which has a good capacity and the information from it can be easily accessed and provided for further processing. Each location area (LA) maintains a registration waiting list (RWL) to keep the newly arrived MTs identities (IDs). However, before the location of the newly moved in MTs is updated at the HLR, a mechanism should be placed so that any incoming call for these MTs can be successfully delivered to their current LA. For this purpose, either forwarding or local anchoring can be used to set up a forwarding pointer from a MT's old LA or local anchor to its new LA as the MT changes its LA. The main factor which reduces the cost of mobile tracking is the repository database which stores the information about the mobile terminal and updates this information in the neighboring visitor location register (VLR).Two methods were introduced [1] to report the location change to the HLR: 1) Static local anchoring (SL): an MT's local anchor is changed to its new VLR upon the arrival of its next incoming call. 2) Dynamic local anchoring (DL): in addition to method 1, the MT's new VLR becomes its local anchor if such a change results in a lower cost.

B. Location Update Procedure in SB Strategy

When a MT changes its LA, the following procedure is performed (Figure 1) [1]

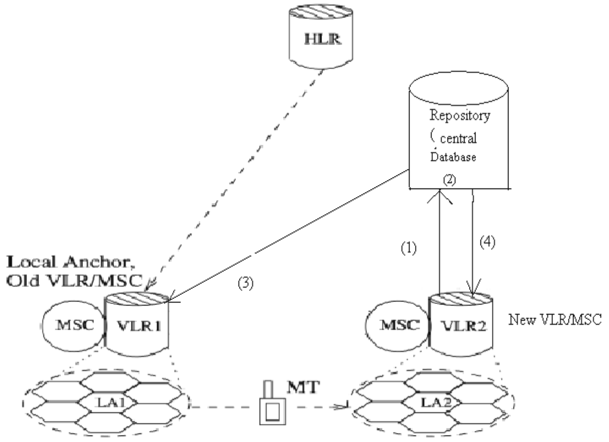


Figure 1. Location Update Procedure in SB Strategy old VLR is the Local Anchor

1. The new MSC detects that the MT enters an LA in its area and sends a message to inform to repository (central database) that MT has entered in it.

2. The repository updates the location of MT in its memory from old MSC to new MSC.

3. Repository sends a message to the old mobile terminal about the current location so that it can find it easily.

4. Repository sends an acknowledgement to the new MT that it has updated the location of MT in the database. The Old MSC checks if it's associated VLR is the local anchor of the MT. If yes (Figure 1), the local anchor is updated to point to the MT's current VLR, and the location update procedure is complete, otherwise (Figure 2) [1], go to the next step.

5. The old VLR removes the MT's ID from its old LA's RWL. A location update message is sent to the MT's local anchor, which then updates itself to point to the MT's new VLR.

6. The local anchor sends back an acknowledgment message to the MT's old MSC. The location update procedure is complete.

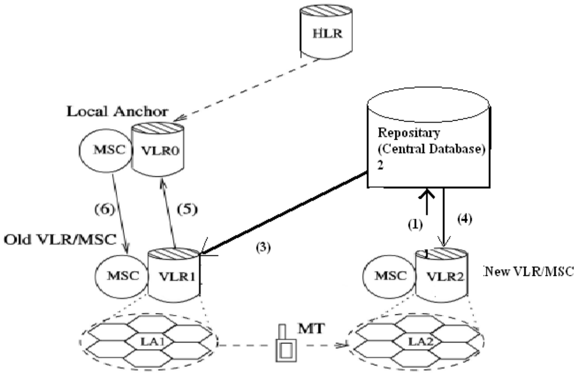


Figure 2 Location Update Procedure in SB Strategy old VLR is not the Local Anchor

C. Call Delivery Procedure

Local anchor changes occur only during the call delivery procedure. The call delivery procedure proceeds as follows:

1. When a call for an MT is originated (the caller can be a wire line or mobile phone), a location request message is sent to the MT's HLR.

2. The HLR obtains the ID of the called MT's local anchor and sends a route request message to the local anchor.

3. If the local anchor is the current VLR of the MT (Figure.3), go to the next step; otherwise (Figure. 4), the route request message is forwarded to the current VLR/MSC.

4. The called MSC searches for the called MT. If the MT is found, a temporary local directory number (TLDN) is allocated to the MT.

5. If the RWL of the called MT's current LA is not empty, all MT IDs in the RWL are sent to the HLR in the route response message along with the TLDN.

6. The current VLR registers the mobile terminal's location information to the repository which stores it in the database.

7. Meanwhile, except for the called MT, for each other MT in the RWL, the current VLR sends a deregistration message to its local anchor, which removes the MT's forwarding pointer entry. The RWL is then emptied. The current VLR becomes the local anchor of all MTs in the RWL.

8. After receiving the route response message from the VLR, the HLR forwards the TLDN to the calling MSC. If the route response message contains any to be- updated MT's ID, the HLR changes these MTs' local anchors to the current VLR.

9. After receiving the TLDN, the calling MSC can set up a connection to the called MSC

III. ANALYTICAL MODEL

A Location Update Cost of the Proposed Strategy

Assume that the MTs arrive at an LA according to a Poisson process; the incoming calls to an MT follow a Poisson Process, and an MT's residence time in an LA follows an exponential distribution [1].To evaluate the cost of proposed strategy following cost notation is used

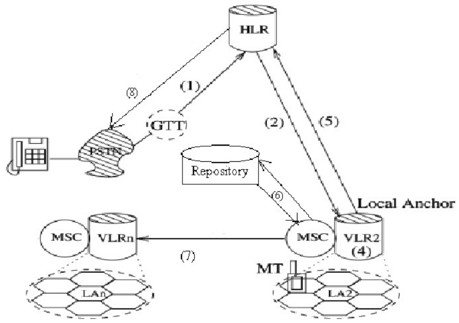


Figure 3Call Delivery Procedure Current VLR is the Local Anchor [1]

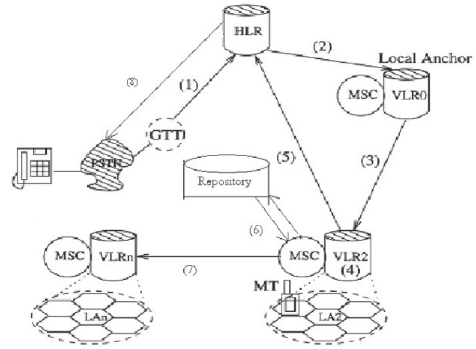


Figure 4 Call Delivery Procedure Current VLR is not the Local Anchor [1]

Cv Cost for a query or an update of the VLR.

Ch Cost for a query or an update of the HLR.

Cvv Cost for transmitting a signaling message between two VLRs.

Chv Cost for transmitting a signaling message between a VLR and the HLR.

As seen from call update procedure, in the proposed strategy, as an MT changes its LA, a location update is performed. However, the location update procedure may be different [1].

There exist two cases that incur different location update costs:

Case 1. The old VLR of the moved-out MT is its local anchor (Figure. 1).

Case 2. The old VLR of the moved-out MT is not its local anchor (Figure. 2).

Location updates cost in Case 1. During the time period that a MT stays in an LA, if there is at least one incoming call arriving for any MT in the LA, then the MT's current VLR becomes its local anchor and its location is updated at the HLR during the call delivery procedure for the first incoming call to this LA. In this case, when the MT moves to another LA, the cost incurred by the LA change, U1 is (Figure. 1)

$$U1 = 2(Cv) \qquad (1)$$

Where the cost for the old VLR deleting the user profile and creating a forwarding pointer is Cv,, and the cost for the new VLR creating the user profile and adding the MT's ID to the RWL is Cv. All the data are centrally updated in the database.

Location updates cost in Case 2. If there is no incoming call to any MT in an LA during a MT's residence period in the LA, steps 5 and 6 in the location update procedure needs to be executed when the MT moves to another LA. In this case, the cost incurred by the LA boundary crossing, U2 is (Figure. 2)

$$U_2 = 4(Cv) \tag{2}$$

where in addition to the costs in, the old VLR removing the MT's ID from its old LA's RWL incurs Cv, the old local anchor updating the forwarding pointer incurs Cv, Given MT, let p_1 be the probability that there is no incoming call to any MT in m_0 LA during m_0's residence period in the LA (i.e., Case 2). Then, the expected location updates cost incurred by an LA change in the proposed Strategy [1]

$$\begin{aligned} Cu &= (1-p_1)U_1 + p_1U_2 \\ &= 2(1+p1)Cv \end{aligned} \tag{3}$$

In the following, we calculate the expected location update cost per call arrival for MT m_0. Let $f(y_0)$ be the density function of m_0's LA residence time with mean 1/λo and g (x_0) be the density function of m_0's inter call interval with mean 1/ μ_0 [1], i.e.,

$$f(y_0) = \lambda_0 e^{-\lambda_0 y_0}, \quad g(x_0) = \mu_0 e^{-\mu_0 x_0}$$

The probability that n LA boundary crossings occur between two call arrivals is

$$\alpha(n) = \begin{cases} 1 - \frac{1}{\rho}[1 - f^*(\mu_0)] & n=0 \\ \frac{1}{\rho}[1 - f^*(\mu_0)]^2 [f^*(\mu_0)]^{n-1}, & n>0 \end{cases} \tag{4}$$

where $\rho = \frac{\mu_0}{\lambda_0}$ is the call-to-mobility ratio (CMR) of MT m_0 and f*(s) is the Laplace-Stieltjes transform of $f(y_0)$.

Then, the expected location update cost per call arrival is

$$C_{u,s} = C_u \sum_{n=1}^{\infty} n\alpha(n) \tag{5}$$

B. *Call Delivery Cost of the Proposed Strategy*

There are two types of calls for which the call delivery cost may be different:

The first incoming call to an MT after the MT moved into a LA.The rest of the incoming calls to the MT during the period that the MT stays in the LA.

We will compare the proposed strategy to the IS-41, local anchoring strategy and Group Registration Strategy. Since the cost for message exchanges between the caller and the callee's HLR is the same for these four strategies, this cost is not included in the cost estimation in this paper. Note that the call delivery cost can be calculated in the same way for both handoff calls and non handoff calls, while the cost triggered by the handoff process can be omitted, since this cost would be the same for all the four strategies when the same handoff mechanism is adopted. When the first incoming call at an LA is for another MT in the LA since MT $m_0$moved into the LA, the HLR is pointing to the current VLR of $m_0$an incoming call arrives for m_0. In this case, the call delivery procedure shown in Figure.3 is performed to deliver an incoming call to m_0.The call delivery cost, V1, is

$$V1=2Chv+Cv+\theta Ch+RL*Cv \tag{6}$$

where the cost for the HLR sending the location request message to the called MT's current VLR is Chv, the cost for the current VLR retrieving the location of the MT's current LA is Cv, the cost for the current VLR sending the TLDN and the MT IDs in the RWL, if any, to the HLR is Chv, and the cost for the HLR to update the MTs' service profiles is Ch. The last term represents the cost of step 6 for local anchor deregistration θ is the probability that there is at least one MT ID in the RWL and RL is the average number of MTs in the RWL [1].

When an MT receives the first incoming call to an LA after the MT moved into the LA, the MT's HLR is still pointing to the MT's local anchor, and the call delivery procedure shown in Figure. 4 is followed. The call delivery cost, V2 is

$$V2=2Chv+2Cv+Ch+\Upsilon Cv \tag{7}$$

where the cost for the HLR sending the location request message to the called MT's local anchor is Chv, the cost for the local anchor retrieving and removing the forwarding pointer is Cv, the cost for the current VLR retrieving the location of the MT's current LA is Cv, the cost for the current VLR sending the TLDN and then to-be-registered MTs' IDs to the HLR is Chv, and the cost for the HLR to update the MTs' service profiles is Ch. In addition, the cost incurred by step 6 of the call delivery procedure is Cv for each MT in the RWL (except for the called MT whose local anchor is deregistered in step 3 of the call delivery procedure). ϒ denotes the average number of MTs in the RWL (excluding the called MT)[1].

Let p_2 be the probability that m_0 receives the first incoming call to m_0's LA since m_0 moved into the LA. The average cost of the first incoming call since an MT moved into its current LA, V is

$$\begin{aligned} V &= (1-p_1)V + p_2V_2 \\ &= 2C_{hv} + [p_2(1-\theta) + R_L]C_v + C_v \end{aligned}$$

$$+[p_2+(1-p_2)\theta]C_1 \tag{8}$$

After the first incoming call, any subsequent incoming call to m_0 while it remains in its current LA invokes cost V1. Thus, the expected cost per call arrival is [1]

$$C_{c,s} = \begin{cases} \mathrm{V} & \rho \le 1 \\ \frac{1}{\rho}(\mathrm{V} + \lfloor \rho - 1 \rfloor \mathrm{V1}), & \rho > 1 \end{cases} \tag{9}$$

Note that, only when ρ > 1, there are likely more than one incoming calls to the MT during its residence at an LA.

Therefore, the total cost per call arrival for the proposed strategy, denoted by CT,s is

$$\mathrm{C}_{\mathrm{T,S}} = \mathrm{C}_{\mathrm{u,s}} + \mathrm{Cc}_{\mathrm{,s}} \tag{10}$$

C. Tracking Cost of the IS-41 Strategy

The location update cost incurred by each LA change in the conventional IS-41 strategy is 4Chv + 2Cv +Ch. Thus, the expected location update cost per call arrival in the conventional strategy is

$$C_{u,c} = (4C_{hv} + 2C_v + C_h)\sum_{n=1}^{\infty} n\alpha(n) \tag{11}$$

The call delivery cost per call arrival of the conventional strategy is

$$C_{c,c} = 2C_{hv} + C_v \tag{12}$$

Note that, as mentioned earlier, the cost of message exchanges between the caller and the callee's HLR is not included in the preceding equation.

Therefore, the total cost per call arrival for the conventional strategy, denoted by $C_{T,c}$ is

$$C_{T,c} = 2C_{u,c} + C_{c,c} \tag{13}$$

D. Tracking Cost of the Local Anchoring Strategy

The local anchor of an MT in the LAS strategy could be the current VLR of the MT or a different VLR. In the former case, the location update cost per LA change is U1, while in the latter case the location update cost per LA change is U2-Cv.(No RWL operation is needed here.) The average location update cost incurred by each LA change in the LAS strategy is

$$\begin{aligned} C_l &= (1 - P_r\{E_0\})U_1 + P_r\{E_0\}(U_2 - C_v) \\ &= \left(\frac{2\mu_0}{\lambda_0 + \mu_0}\right)(C_{vv} + C_v) + \frac{\lambda_0}{\lambda_0 + \mu_0}(4C_{vv} + 3C_{vv}) \end{aligned} \tag{14}$$

where Pr{E0} is the probability that an MT receives no call while residing in an LA [1].

Then, the expected location update cost per call arrival in the LAS strategy is

$$C_{u,l} = C_l \sum_{n=1}^{\infty} n\alpha(n) \tag{15}$$

The delivery cost of the first incoming call in the LAS strategy is

$$\begin{aligned} C_{l,1} &= 2C_{hv} + C_{vv} + 2C_v + C_h + (C_{vv} + C_v) \\ &= 2C_{hv} + 2C_{vv} + 3C_v + C_h, \end{aligned} \tag{16}$$

where Cvv +Cv represents the cost of local anchor deregistration.

The delivery cost of any subsequent incoming call to an MT after its first incoming call is 2Chv+Cv. Thus, the expected cost per call arrival in the LAS strategy is

$$C_{c,l} = \begin{cases} C_{l,1}, & \rho \le 1 \\ \frac{1}{\lfloor \rho \rfloor}[C_{l,1} + \lfloor \rho - 1 \rfloor (2C_{h\,v} + C_v)], & \rho \le 1 \end{cases} \tag{17}$$

Therefore, the total cost per call arrival for the LAS strategy, denoted by

$$C_{T,l} = C_{u,l} + C_{c,l} \tag{18}$$

E. Call Delivery Cost in the Group Registration Strategy

Call delivery cost in GR strategy is very much similar to the SB strategy the main difference lies in the updating and delivering cost with the interaction of two VLRs. Here is the cost of GR strategy is given [1].

$$U_1 = 2(C_{vv} + C_v) \tag{19}$$

$$U_2 = 4(C_{vv} + C_v) \tag{20}$$

$$\begin{aligned} Cu &= (1 - p1)U_1 + p1U_2 \\ &= 2(1 + p1)Cv \end{aligned} \tag{21}$$

$$C_{u,g} = C_u \sum_{n=1}^{\infty} n\alpha(n) \tag{22}$$

All the symbols used in this strategy is same as used in the SB strategy and the total cost calculated below as.

$$V_1 = 2C_{hv} + C_v + \theta C_h + R_L(C_{vv} + C_v),$$

$$V_2 = 2C_{hv} + C_{vv} + 2C_v + C_h + \gamma(C_{vv} + C_v)$$

$$V_2 = (1-p_2)V_1 + p_2 V_2$$

=2Chv +[p2 (1-θ)+RL](Cvv +Cv)
+ Cv +[p2 +(1-p2)θ]Ch (23)

Total Cost in GR strategy is

$$C_{T,g} = C_{u,g} + C_{c,g} \tag{24}$$

IV. NUMERICAL RESULTS AND DISCUSSIONS

In this section, performance comparison studied. First, several critical parameters of the proposed strategy, namely, p_1, p_2,θ, and β(k) are taken [1] and their impact on the proposed strategy is discussed, after that the proposed strategy is compared with the IS-41, LAS, GR strategies under different scenarios. The proposed strategy and the simulations in this section are applicable to any user movement patterns such as random walk, moving back and forth across adjacent LAs, etc. In Figures 5, 6, 7, 8, and 9, it is assumed that $\lambda_{n,i}$ (i=1,2,……..,8) are uniformly distributed over (0.1,3) and the incoming call arrival rates ϕ_i (i=1,2,3,…….,M) at all MTs in the LA are uniformly distributed over (0.2, 3), which implies that the $\mu_{ni,i}$(i=1,2,3,……8) are also uniformly distributed over (0.2, 3)

A. Parameter Evaluations

Figure. 5 shows the graph of probabilities p_1 and p_2 versus λ_0 where Ni =20 (i= 1, 2 … 8) are used for demonstration purposes. From Figure. 5, we can see that p_1 and p_2 are very small. As Ni (i =1, 2……., 8) increases, p_1 and p_2 decrease further p_1 increases as λ_0 increases. This is true since as the mean LA residence time, ($1/\lambda_0$), of MT decreases, the probability that there is no incoming call to an LA during the MT's stay at the LA increases. It is also noted that the impact of μ_0 on p_1 is negligible. On the other hand, p_1 decreases as μ_0 decreases and the impact of λ_0 on p_2 is negligible. That is, as the mean of an MT's intercall interval, $1/\mu_0$ increases the probability that the MT receives the first incoming call to its current LA decreases. In real mobile networks, Ni (i=1,2,…………. ,8) may be much larger than 20,therefore, p_1 and p_2 would become much smaller. Thus, the location updates process shown in Figure. 1 and the call delivery process shown in Figure. 3 are invoked by the proposed strategy most of the time, resulting in less cost than those processes shown in Figure. 2 and Figure 4 Figure. 6 studies the impact of the MT arrival rate at a LA on the probability θ [1] that the RWL is not empty when an incoming call arrives to the LA, under different numbers M of MTs registered at an LA. From Figure 6, it can be seen that θ decreases as η decreases or M increases. This is true because a smaller MT arrival rate to an LA increases the probability that the RWL is empty upon the arrival of a call. On the other hand, more MTs in an LA will make the LA receive calls more often and the RWL will be emptied more frequently, thus the probability that the RWL is not empty becomes smaller, resulting in a smaller cost for the proposed strategy. Figure.7 studies the probability β(k) that there are k MTs in the RWL, which determines the cost of piggybacking the RWL in the route request acknowledgment message as well as the memory requirement for maintaining the RWL in the VLR. Three sets of (M,η) i.e., (200,100),(200, 50), and (400, 50), are considered. It is observed from Figure. 7 that the first set of (M,η) results in the greatest β(k) (k= 1, 2 …). As the MT arrival rate at the LA decreases or the number of MTs M in the LA increases, β(k) k=1,2,3……) decreases. It can be seen that β(k) approaches zero when k ≥ 4 for all given parameter sets.

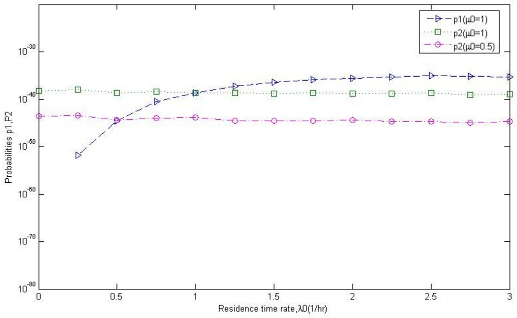


Figure 5 Probabilities p_1 , p_2 versus residence time rate

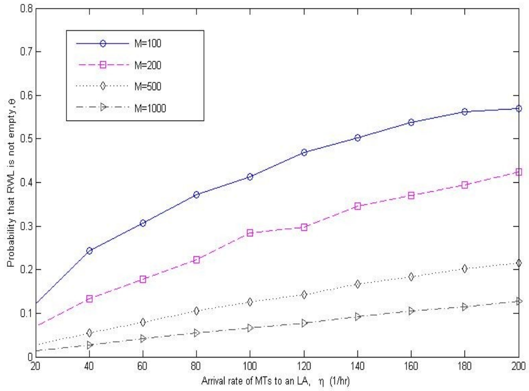


Figure 6 Probability of the RWL not empty versus arrival rate of a MT to a LA

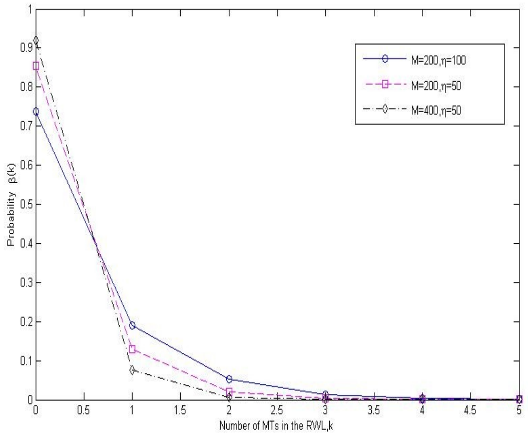


Figure 7. Probability of the k MTs in the RWL, β(k) versus number of MTs in the RWL

The International Mobile Subscriber Identity (IMSI) is usually used to identify an MT in location management and its length is no more than 15 digits. If one digit is 1 byte long, then N MT IDs in the RWL require a space of 15 × N bytes. For N=4, the required space is 60 bytes. As indicated in Figure. 7, usually a much smaller space is needed for the RWL.

B. Performance Comparisons

Figure. 8 compares the total cost per call arrival of the proposed strategy with those of the conventional IS-41 strategy, GR strategy and the LAS strategy under different λ_0 values, where (M,η) is set to (200,100).In Figures. 8 and 4.5, the following cost values are used: Cv =1, Ch = 1.5, Cvv =1, and Chv = 2. It is assumed that a query or an update at the HLR incurs a larger cost than that at the VLR, and the message exchanges between the HLR and a VLR incur a larger cost than those between two VLRs.

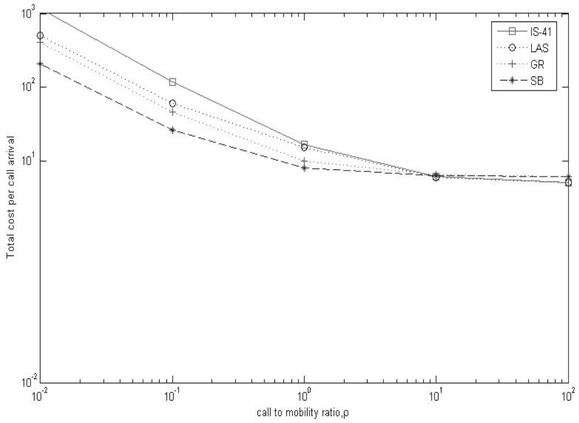


Figure 8 (a) Total cost per call arrival versus CMR for different strategies with λ_0 =0.5 per hour

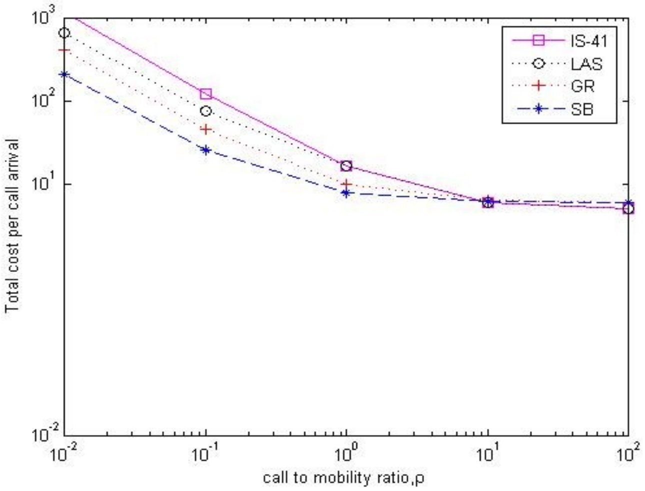


Figure 8 (b) Total cost per call arrival versus CMR for different strategies with λ_0 = 5.0 per hour

This reflects the generic view that the resources at the HLR are usually more expensive to consume than those at the VLR and the distance between the HLR and VLR is larger than that between two VLRs. From Figure 9, it is observed that the proposed SB strategy incurs a smaller total cost per call arrival than three other conventional strategy and the LAS strategy when an MT's CMR ρ < 10, while for ρ> 10, the proposed strategy incurs a slightly Equivalent cost to the other three strategies. Moreover, the impact λ_0 on the total cost per call arrival of the proposed strategy is negligible.

The preceding observations can be explained as follows Compared to the IS-41 strategy, the SB strategy has a smaller location update cost by reporting its location changes to the local anchor. Compared to the LAS strategy, the SB strategy incurs both smaller location update and call delivery costs. The SB strategy also has low cost than the GR strategy in both the cases.

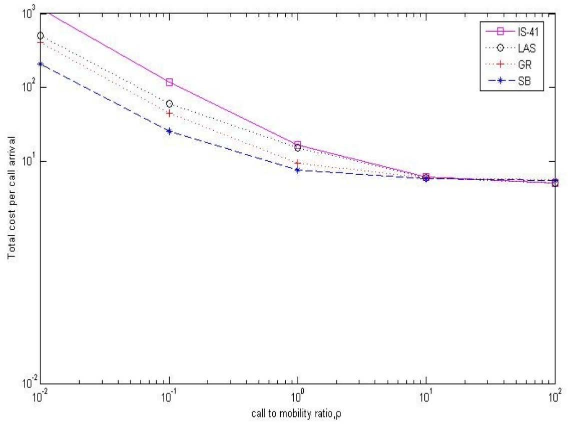


Figure 9 (a) Total cost per call arrival versus CMR for different strategies with M=200 and η=50

In Figure. 9, we compare the proposed strategy to the IS-41 strategy, LAS strategy and the GR strategy under two different sets of (M,η): (200, 50) and (400, 50)We observe that, as M increases, the proposed strategy outperforms the IS-41,GR and LAS strategies over a wider range of ρ. When comparing Figure. 8a to Figure. 9a, we observe that as η decreases, the proposed strategy results in a smaller cost for high ρ values. These observations are readily understood in that a larger number of MTs in the LA or a smaller MT arrival rate to the LA results in a smaller RWL (i.e., RL), thus reducing the call delivery cost of the proposed strategy (which especially benefits high ρ values). Furthermore, a larger M results in smaller p_1 and p_2 , reducing both the expected location update cost and calls delivery cost in the SB strategy.

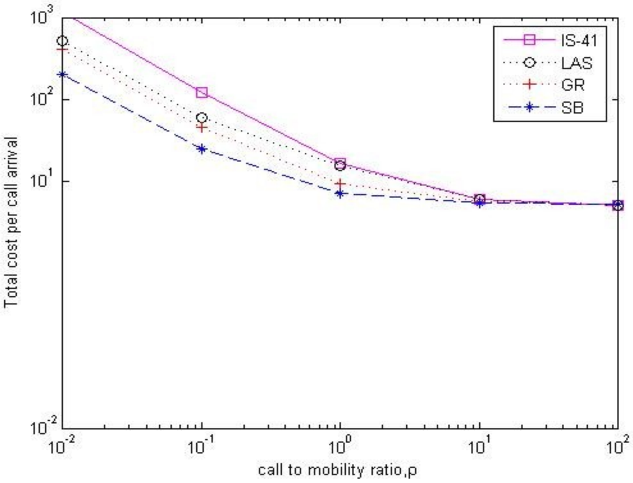


Figure 9(b) Total cost per call arrival versus CMR for different strategies with M=400 and η=50

V. CONCLUSIONS

A SB location tracking strategy has been proposed in this paper which is based on the modified version of group registration technique. It uses one repository as a central database which stores the information about the mobile terminal and also the updated location of it. An analytical model of the strategy is described and numerical result is presented for the performance evaluation. The proposed strategy is compared to IS-41, the LAS and the GR strategies and it is observed that the proposed strategy can achieve a cost reduction over a wide range of CMRs. Moreover, the proposed strategy is based on the concept of central database and storage of information about MT deployed in existing mobile systems and does not require the system to collect the mobility and calling statistics for individual mobile terminals. The proposed strategy work well for low CMR.

REFERENCES

[1] Zuji Mao and C. Douligeris, Senior Member "Group Registration with Local Anchor for Location Tracking in Mobile Networks" IEEE Trans. on mobile computing, vol.5, no.5,pp. 583 – 595, MAY 2006

[2] I.F. Akyildiz, J.S. M. Ho, and Y.-B. Lin, "Movement-Based Location Update and Selective Pagings for PCS Networks," IEEE/ACM Trans. Networking, vol. 4, no. 4, pp. 629-638, Aug. 1996.

[3] E. Cayirci and I.F. Akyildiz, "User Mobility Pattern Scheme for Location Update and Paging in Wireless Systems," IEEE Trans. Mobile Computing, vol. 1, no. 3, pp. 236-247, July-Sept. 2002.

[4] I.-R. Chen and B. Gu, "Quantitative Analysis of a Hybrid Replication with Forwarding Strategy for Efficient and Uniform Location Management in Mobile Wireless Networks," IEEE Trans.Mobile Computing, vol. 2, no. 1, pp. 3-15, Jan.-Mar. 2003.

[5] J.S.M. Ho and I.F. Akyildiz, "Local Anchor Scheme for Reducing Signaling Costs in Personal Communications Networks," IEEE/ACM Trans. Networking, vol. 4, no. 5, pp. 709-725, Oct. 1996.

[6] Y.-B. Lin, "Reducing Location Update Cost in a PCS Network," IEEE/ACM Trans. Networking, vol. 5, no. 1, pp. 25-33, Feb. 1997.

[7] W. Ma and Y. Fang, "Two-Level Pointer Forwarding Strategy for Location Management in PCS Networks," IEEE Trans. Mobile Computing, vol. 1, no. 1, pp. 32-45, Jan.-Mar. 2002.

[8] Z. Mao and C. Douligeris, "A Location-Based Mobility Tracking Scheme for PCS Networks," Computer Commun., vol. 23, no. 18, pp. 1729-1739, Dec. 2000.

[9] Z. Mao and C. Douligeris, "An Integrated Strategy for Reducing Location Management Cost," IEEE Comm. Letters, vol. 8, no. 1, pp. 33-35, Jan. 2004.

[10] Z. Mao and C. Douligeris, "A Distributed Database Architecture for Global Roaming in Next-Generation Mobile Networks," IEEE/ACM Trans. Networking, vol. 12, no. 1, pp. 146-160, Feb. 2004.

[11] S. Tabbane, "An Alternative Strategy for Location Tracking," IEEE J. Selected Areas Comm., vol. 13, no. 5, pp. 880-892, June 1995.

[12] J. S. M. Ho and I. F. Akyildiz, "Local anchor scheme for reducing location tracking costs in PCN's," in Proc. ACM MOBICOM'95, Nov.1995, pp. 181-194. 131

[13] R. Jain and Y. B. Lin, "An auxiliary user location strategy employing forwarding pointers to reduce network impact of PCS," ACM-Baker J.Wireless Network, vol. 1, no. 2, pp. 197-210, July 1995.

[14] Y. B. Lin, "Determining the user locations for personal communicationis services networks," IEEE Trans. Veh. Technol., vol. 43, no. 3, pp.466,473, Aug. 1994.

[15] I.R. Chen and B. Gu, "A Comparative Cost Analysis of Degradable Location Management Algorithms in Wireless Networks," The Computer J., vol. 45, no. 3, pp. 304-319, 2002.

[16] I.R. Chen, T.M. Chen, and C. Lee, "Performance Evaluation of Forwarding Strategies for Location Management in Mobile Networks," The Computer J., vol. 41, no. 4, pp. 243-253, 1998.

[17] N. Shivakumar, J. Jannink, and J. Widom, "Per-user profile replication in mobile environments: algorithms, analysis, and simulation results," ACM-Baltzer J. Mobile networks and Applications, vol. 2, no. 2, pp 129–140, Oct. 1997.

Surveillance: A (Potential) Threat to Political Participation?

Maria João Simões
Department of Sociology,
University of Beira Interior, Covilhã
Researcher at Research Centre of Social Sciences (CICS),
University of Minho
Braga, Portugal
mariajoaosimoes@sapo.pt

***Abstract* - Despite efforts of several authors, surveillance is nowadays yet sparsely understood, although surveillance has increasing impacts in our lives. The purpose of this paper is to point out, from a theoretical point of view, the threats that surveillance presents to political participation (and by consequence, to democracy) in digital societies. Current researches present the threats of surveillance to democracy focusing mainly in democracy-privacy trade-offs. Such debate, on the one hand, circumscribes the issue to a great extent to choosing the rulers and the kind of political regime, which does not allow a broader analysis of citizen participation in all spheres of public life in their daily life. On the other hand, the current debates seem put a little aside from the main issue: it is not the loss of privacy, but the loss of autonomy that challenges participation (and by consequence, democracy); although nowadays the threats to autonomy proceed mainly from the loss of privacy.**

Keywords - surveillance, political participation, autonomy, risk

I. INTRODUCTION

Digital societies are those that bet on a knowledge-based development, through triggering processes that enables permanent creation, updating, diffusion, transference and sharing of knowledge. On that account knowledge societies are a better concept. On the other hand, concerning the predominant political and ideological conceptions, the knowledge-based development can also be inclusive, sustainable and participative. One chooses also by a more all-encompassing conception of digital city that are linking to all aspects of social, economic, political and cultural life. Another crucial theoretical proposition is that ICT don't determine the digital cities' emergency and development per se; they are only a device to achieve this goal.

Simões [1][2][3] pointed out the social, political and cultural conditions, not only technological ones, that stimulate the participation through the use of ICT but also those which constrain that participation. This paper is the background to the challenge we are facing today and to which we intend to answer: what are the challenges that participation is dealing within digital societies considering the spreading and deepening of ICT and other technologies as surveillance devices?

Although current studies on surveillance, namely those by Lyon and Marx, present the threats of surveillance to democracy focusing mainly in democracy-privacy trade-offs, our proposal is quite different. On the one hand, we focus the debate on a more transversal issue: the political participation. Held's concept of politics was adopted [4]; according the author, politics is power; it's the capacity of individual and collective social actors to change or to maintain their social and physical environment. Politics concerns areas that demand appropriate public actions and choices from which one expects public consequences.

On the other hand, the analysis of the threats to political participation (and thus to democracy) is centred on autonomy and not on privacy, as in current researches.

Firstly, this paper presents crucial requirements for political participation. Secondly, the major theoretical propositions on surveillance are emphasised and the next point deals with the factors contributing and stimulating surveillance in digital societies. Before the final considerations, the paper looks upon surveillance risks to political participation. Finally, some considerations will be presented.

II. REQUIREMENTS FOR POLITICAL PARTICIPATION

Marshall [5] identifies three kinds of citizenship – civil, political e social – that, in his evolutionist perspective, have emerged in a sequential way in the 18th, 19th and 20th centuries. The first established the necessary rights for the exercise of individual freedom (namely personal freedom, freedom of thought, speech and religion, the right to justice, to private property and to establish contracts. Political citizenship acclaims the right to vote and of association and the right to participate in politics, whether as a voter or an elected member (thus with political authority).

This gradual evolution of citizenship should conclude at the welfare state in the mid-20th century, through the attribution of social rights that have established social and economic security rights, i.e., education, health, employment, social security rights, among others.

Marshall [6] had already made a clear distinction between formal rights and their effective application, defending that the effective exercise of citizenship was only possible since where the social rights were also assured. Without social rights, i.e., a certain level of well-being, people would not have material grounds to participate as equals in the political and social life, because they would live economically and politically subordinated to others. In this sense, the impact of the formal equality of civil and political rights was scarce.

Theories of political participation, notwithstanding voluntarist action theories, start with a crucial assumption: an unequal distribution of economic, social and political resources limits the possibility of autonomous choices, judgement and political actions.

To Simões [7], the most important issue is not if people act, but how they act, which became crucial to restate, on the one hand, the effectiveness of civil, political and social rights, and enunciate, in the other hand, the concept of autonomy and the conditions to autonomy effectiveness.

How Roche [8] referred the citizens are autonomous when they are rational agents and free moral individuals. Autonomy «connotes the capacity of human beings to reason self-consciously, to be self-reflective and to be self-determining. It involves the ability to deliberate, judge, choose and act upon different possible courses of action in private as well as public life» [9]. Its effectiveness is closely dependent on material and cognitive resources, among others, that people are able to achieve, and on facilities that allow (or prevent) access to these resources. It also depends on the liberation of constraints concerning relationships of economic, political and social domination.

To Oldfield [10], actions are autonomous when they show characteristics as self-determination and authenticity. An action is self-determined when it is a product of a person's will. This requires the skills of not being constrained by others or by the demands of institutions where one is included. It is authentic when it is built and chosen by each person, not by others, and rationally presented.

The main issue about expanding surveillance is whether autonomy is being threatened, therefore endangering the political participation within digital societies. The debate around the autonomy concept is more heuristic to a further research than the privacy one. As most people when facing the claim around privacy-security trade-offs, rapidly choose security, saying they have nothing to hide. But are they aware of the consequences brought to their autonomy by losing privacy? Are they aware of the loss on their faculty and power to make political choices and to participate in decision-making within the public domain? Thus, when we face threats to political participation, the central issue is the autonomy.

Certainly, the autonomy is mostly threatened nowadays by the invasion of privacy enabled by ICT and other technologies used as surveillance devices. In this sense, it is crucial to analyse how autonomy intertwine with privacy (invasion) and also to deepen the research about privacy in the digital era, as this era has some specificities and it is also a very open-ended theoretical domain, where there are already many contributions to the research, namely of Nissenbaum [11] and Stalder [12]. The first author gives crucial contributions for this matter when she presents a theoretical account of the right to privacy as it is applied to information on people and she proposes an analysis of privacy in terms of contextual integrity, differentiating norms of appropriateness and norms of distribution. The second author when emphasises that the theory of privacy – based on concepts of separation and individualism – is unworkable in an environment characterized by a myriad of electronic connections.

III. SURVEILLANCE, AMBIVALANCE AND SOCIAL INTERESTS

Surveillance is not a new phenomenon, but it increased exponentially in the end of the 20th century, largely because until then it was predominantly restricted to administrative, productive and military spheres. After that period, surveillance broadened extensively to all spheres and fields of social activity (commercial, health, public spaces, and so on), and it is also becoming global, intensively entering the routines of our private and daily lives.

Surveillance has two faces, how Lyon [13] [14] referred to; surveillance is both enabler and constrainer to our action. A good example of this ambivalent process is how surveillance spread alongside the development of democracy and the emergency of social rights.

It's wrong that we focus only on one surveillance face. Its positive and/or negative dimensions can be tightened more in one direction than another, depending on what interests and purposes surveillance devices are designed, created and used for. We should also look upon the historical context in which those are inserted because the effects of technological systems aren't the same in all contexts in which they are used. On the other hand, knowing people's and organizations' ideologies in each context helps to understand the goals for which surveillance devices are aimed and how they are used; it also helps to identify specific resistance forms that allow to change or reducing the more negative effects of surveillance.

As Lyon [15] pointed, the systematic surveillance actually developed along the cities, corporations, governmental administration and military organization growth, all in all with modernity, particularly increasing since the 19th century. The reinforcement and spreading of surveillance practices can not be understood, like in Marxist approaches, as a product of a capitalist conspiracy [16]. They result from the complex way we organize our society, our social, political and economic relations. We live in societies that value consumption freedom, speed, mobility, efficacy, productivity, efficiency and security. In this sense, in modern societies, most organizations use ST in systematic ways, in order to reduce uncertainties and to control production outcome but mostly to prevent risks in security grounds as opposing to threatening behaviours and obtaining people agreement.

The surveillance technologies (ST) used for risk prevention were intensified after September 11th, 2001. The belief in diminishing potential risks and controlling outcomes through technology enhances the pressure to move towards more and more sophisticated surveillance means. More and more specialized agencies are increasingly using more and more sophisticated means to collect people's data in a customary way, make all of them target groups of monitorization and suspicion.

The way how ST are designed and programmed as the intended data outcome aren't neutral. As Lyon [17] highlights, ST intended effects aiming to reinforce regimes that they were designed and programmed for. These effects have only recently began to be analysed in a systematic way and are yet barely understood.

Surveillance was always a source of power and today even more. But it would be simplistic to think that ST reinforce the position of the most powerful, as according to the structuralist perspective of Foucault, read in the panopticon metaphor, where individuals act passively in accordance to the established rules, their behaviours being determined by settled surveillance systems. It would also be simplistic to think, according to the voluntarist theories of action, that surveillance implies no constraints to individual action, as people are able to choose their actions freely.

Burns and Flam [18], in the field of synthesis theories, reject both perspectives, as humans, despite the rigidity of social structures, are cognitive beings and endowed with opportunities to resist, reduce, change and reshape the constraints imposed by dominant groups. As Giddens [19] says, in the field of dialectic control, power is not always absolute.

Technologies are both socially shaped and with social consequences; some of them can overhaul the purposes for which they were created. Besides, even if they were created and used with good purposes, they might have undesirable and unintended consequences.

IV. FACTORS THAT STIMULATE AND CONTRIBUTE TO SURVEILLANCE

Surveillance is thus today the result of a process that has been spreading and being refined throughout history. An important milestone was the 19th century, as the development of statistics was used by the states to count, categorize, classify and administer their citizens. Its emergency already set the framework and models to how surveillance should develop in the next centuries.

With the introduction of statistics, Castel [20] emphasises a change from the "observable" to the "deduced" in the construction of individual selves. The administrative practices were thus focusing more on risks than on dangers. The gathering of data, which focused only on specific suspected groups or individuals to thwart danger, widened to monitor everyone to prevent risk. Risk management is associated to the use of statistic techniques.

These techniques were set to create personal profiles that weren't under the immediate glance of those who watch. Preventive policies are no longer interested in individuals; with statistic techniques, profiles are built starting with the subject deconstruction followed by a construction based on a combination of statistical correlations of heterogeneous elements and facts that allow verifying if individuals are susceptible or not of producing risk.

When paper surveillance was replaced by electronics, it gained «the potential to erode liberties and freedoms because those technologies changed the balance of power in our society» [21]. In the late 1980's, before the increasing sophistication of ICT and the emergency of others ST, Gary Marx [22] highlighted that computers had not only spread the surveillance scope but also allowed more regular and deepened forms of surveillance, emphasizing also that the way ST are to be used in contemporary societies reaffirms an increasing totalitarian potential, regardless who controls these surveillance means.

Commercial surveillance is one of the major contributors to the erosion of the boundaries between public and private spaces and to the intensification of surveillance that begins to embrace the routines of our daily life and our privacy [23]. Let us just consider credit cards, data from insurance companies or frequent costumer cards provided by all kinds of shops, including bookshops and cinemas.

Certainly, commercial surveillance devices are aimed to persuade people as consumers, but they can hold other purposes, i.e., political, as it happened with other technological devices. These cards allow knowing where, when and what we consume, thus revealing our political and ideological preferences: precious information to detect political activists in some democratic countries and even more in (potential) authoritarian and totalitarian regimes.

The spreading of CCTV «over city-centre streets represents the most visible sign of the "dispersal of discipline" from the prison to the factory and school, to encompass all the urban landscape» [24].

About liberty-security trade-offs, after the terrorist attacks in the London underground, people of UK feel safer with the CCTV, but in this ambivalent social world, these devices created for safety reasons can stretch its action to the political sphere, recording encounters and gatherings that participating people didn't want to be watched and recorded.

If surveillance spread in fixed places, the same happened in online and offline mobilities area [25]**.** If credit cards and mobile phones allow detecting where people are or were, on the other hand, GPS (Global Positioning Satellite) allow locating and monitoring drivers and mobile phone users; other localization devices as intelligent transportation systems, are referred by Bennet, Raab and Regan [26], namely the automatic systems of highway tolls and embedded chips in vehicles.

All these devices allow tracing where we are, where we were, where we are going or where we went and clearly with whom we talk or meet. On the other hand, online mobility also allows recording not only what we search, what we read, with whom we speak and what is said, which is invasive concerning civil and political rights.

Biosurveillance, research domain namely of Ploeg [27] and Nelkin [28], is another growing field of surveillance through the collection of DNA, eyes, face, hands, fingers, voice and body data. The possibility to link each name and each number to a body allows distinguishing people therefore “enriching” citizen categorization, classification and profiling in order to reinforce social control.

Such processes cause not only inequalities accessing social and other rights thus reducing necessary resources to political participation, but such data can also be used more directly to achieve political goals, namely to identify people who were in specific places involved in political activities.

The implementation of ID cards has increased from September, 11th on, particularly the most recent cards, which became the most important and the most sophisticated one from a technological point of view to perform social classification and control [29]. These devices, as those authors refer, combine, on one hand, traditional characteristics (namely data from the health system or social security databases) with more advanced characteristics of identification as biometric data. On the other hand, in ID cards we can find programmable chips to collect more data and that can be easily linked to remote authentication mechanisms [30].

However, ICT remained central because they alone have turned possible the collection of personal data and the construction of widened databases that allow the categorization and the upcoming classification of individuals for purposes of social control.

We face a set of circumstances in which interest groups stimulate the development and deepen of technologies; these technologies strengthen the extensive and intensive advances of surveillance and also permit their spreading to unforeseen areas when they are to be conceived and designed.

Marx [31], among other authors, pointed out that the control practice and culture is changing. According to him, “hard” forms of control aren’t reducing simultaneously as “soft” forms are expanding in different ways. Although, his “soft” designation doesn’t seem the more adequate term, because although some of them are less visible they are more invasive and challenge even more our citizenship rights. As the author refers we deal with data collection processes that encompass the use of misleading information, with benefits offered in return of information, with gathering of “false” volunteers appealed to good citizenship and even with the use of hidden or disguised collection techniques.

V. SURVEILLANCE, RISK AND POLITICAL PARTICIPATION

The major part of surveillance nowadays takes place within the digital sphere, thus we don’t see it and many times we don’t perceive it in our daily life. Moreover too much information about citizens flow more and more without their knowledge and permission. On the other hand, in most of the situations digital surveillance doesn’t work clandestinely, but in those regular moments and places of our daily life.

This is even more worrying, because large population sectors are more willing to give away their personal data, believing more in the benefits of surveillance than in its potential risks, thinking they have nothing to fear or hide [32].

Governments, companies or the media contribute to this situation because news focuses on the impressiveness of robbery, crime, war and terrorism facts. This contributes to increasing feelings of insecurity on people, making them to accept the surveillance devices.

We can yet look at cultural changes characterized by the fascination of exposing aspects of private lives and public exposition, namely on facebook and TV, which implies that people cooperate with surveillance means [33]. Such assumptions allow rejecting the simplistic view that the available data only serve the interest groups that have the technological devices to collect data.

Moreover, the ICT made possible erasing data, either without trace, or traceable only by experts, which raises crucial issues concerning data reliability [34].

Both of these aspects have great implications in the political sphere.

As Marx [35] emphasizes, we are interested as citizens in avoiding discrimination, manipulation and an inappropriate classification that could be the result of an inadequate combination of our personal data.

This inappropriate classification can even be obtained for political ends, namely inducing certain political behaviour.

As it has already referred, the problematic issue is that citizens profile are constructed connecting decontextualized data which don’t translate the variety of personal and social contexts, as these are reduced to limited number of variables: the necessary for prevision and scientific generalization capacities. Given that any data can be classified regarding a statistic model as comprising higher or lower risk, it was assumed that risk is understood and consequently controllable in a probabilistic ground. In this process identities are constructed with decontextualized data and subsequently not equal to our own identities. For example, we can be considered political activists or oppositionists when fact we are not. The worst is that these identities were never questioned and, despite this fact, they can prescribe our political behaviour.

As Marx [36] says, it can work for management and medical decision-making, but not for liberties and democratic rights; many times people know little about these databases and their consequences and how their identities are thus constructed.

The registration, categorization and classification made possible by ST trigger processes of inclusion and exclusion concerning participation opportunities that have impact upon people’s life trajectories, depending on the categories in which they are inserted, even though these categories do not match up to our own identities. In short, surveillance impacts on political participation, on our life opportunities, on our privacy and also on social control and democracy.

So, we are facing nowadays a paradoxical situation: the use of ST for preventing risk has increasingly become a risk [37]. A risk to be taken into account for the exercise of citizenship; where we face also the threat of totalitarianism. The question is whether we are witnessing either an increase in their negative dimensions or a growing imbalance of power between the "vigilant" and the "monitored" and thus, a subsequent thickening of the raising risks of a totalitarian and/or unequal society, composed by increasingly “transparent” citizens.

VI Final Considerations and Future Work

With the enlargement of the surveillance scope and its increasing intensity, if we render ourselves into more “transparent” citizens and because it is impossible to know whether we are being monitored or not, like in Foucault’s panopticon, ST have the capacity to induce a state of permanent and conscious visibility, as felt by the panopticon prisoner, that assures the automatic functioning of power, which can threaten our autonomy, constrain our political activity or determine political participation according to the *status quo*.

The increasing and deepening surveillance referred above occur most of the times with none or insufficient public debate about established policies, but also with such haste that legal and political efforts that could safeguard certain social and political implications are limited.

As Lyon [38] and other authors say, can we be facing the emergency of a more totalitarian society, a prison society?

Yes, we do, but before going down that road, it is essential to centre the debate on two previous issues. Firstly, we need to focus the debate on a more transversal issue: political participation. Secondly, we need to centre the research primarily on the challenges to autonomy instead of those regarding privacy. It is the reduction of autonomy that can contribute to the conditioning of even more public actions as well as choices of the citizens (namely in the spheres of education, work and health), but can also lead to a totalitarian society. So, it is important to put the following question: do “transparent” citizens have autonomy to participate in the political life of their cities, regions, countries or even at a global level?

Political participation depends primarily on autonomy. Certainly, the invasion of privacy is nowadays the major process of reducing autonomy, adjoining to other forms of domination, as economic domination. Thus, the ways how privacy intertwines with autonomy are crucial issues.

The concept of autonomy has also a more heuristic potential for research on the perceptions of people about how surveillance threats political participation. This can be a first step to a public debate.

On the other hand, by centring the debate on autonomy it becomes easier to mobilize citizens to participate in the public debate concerning expanding surveillance and the threats to their political participation. As we saw in other section, when citizens are facing the security-privacy trade-offs, they choose security, saying that they have nothing to hide. But are they aware of the consequences that the loss of their privacy brings upon their autonomy (and thus in their political participation)? Then, the research and the public debate have to focus on the issue of autonomy.

The public debate will be one of the paths to face the imbalance of power between the "vigilant" and the "monitored". Through ICT, citizens can trace many platforms not only to debate but also to mobilize other citizens who aren’t aware of surveillance threats. The public debate can lead to another design, other programming of ST and even to someone’s refusal. More than that, citizens have to participate in decision making processes regarding the design and use of surveillance devices in order to assure our freedom and democracy, which are only possible with our autonomous participation.

Sociological research is also crucial to our self-knowledge of social reality and a path to achieve more responsive surveillance practices, that is, practices that didn’t threatened the political citizenship.

From a theoretical point of view, it would be crucial to deepen even more the links between autonomy and privacy in digital societies. In future works we will consider these issues.

References

[1] Simões, Maria João (2005), " Política e Tecnologia - Tecnologias da Informação e da Comunicação e Participação Política em Portugal " [Politics and Technology - Information and Communication Technologies and Political Participation in Portugal], Oeiras: Celta.

[2] Simões, Maria João and Santos, Domingos (2008), "Challenges in the Digital Cities and Regions in Portugal" in Paul Cunningham and Miriam Cunningham (eds.), Collaboration and Knowledge Economy - Issues, Applications, Case Studies, Amsterdam: IOS Press, Vol. 5, Part 1, pp. 546-554.

[3] Simões, Maria João and Araújo, Emília (2009), “A sociological look at e-Democracy” in in Patrizia Bitonti and Vanessa Carrieri (eds.), E-Gov. 2.0: pave the way to e-Participation, Rome: Eurospace S.r.l; pp. 155-161.

[4] Held, David (1996), “Models of Democracy”, Cambridge: Polity Press (2ª ed.).

[5] Marshall, Thomas (1998), “Citizenship and Social Class”, em Gershon Shafir (ed.), The Citizenship Debates, Minneapolis, University of Minnesota Press, pp. 93-111.

[6] Marshall, Thomas (1998), “Citizenship and Social Class”, em Gershon Shafir (ed.), The Citizenship Debates, Minneapolis, University of Minnesota Press, pp. 93-111.

[7] Simões, Maria João (2005), " Política e Tecnologia - Tecnologias da Informação e da Comunicação e Participação Política em Portugal " [Politics and Technology - Information and Communication Technologies and Political Participation in Portugal], Oeiras: Celta.

[8] Roche, Maurice (1998), “Citizenship, social theory and social change” in Bryan Turner and Peter Hamilton, Citizenship: Critical Concepts, London, Routledge, (2ª ed.), pp. 80-110.

[9] Held, David (1996), “Models of Democracy”, Cambridge: Polity Press (2ª ed.), p. 300.

[10] Oldfield, Adrian (1998), "Citizenship and Community: Civil Republicanism and the Modern World", Londres: Routledge (2º ed.).

[11] Nissenbaum, Helen (2004) “Privacy as Contextual Integrity”, Washington Law Review, vol. 79, Dec. 2010, pp. 101-139.

[12] Stalder, Felix (2002), “Opinion. Privacy is not the antidote to surveillance”, Surveillance & Society 1 (1), Dec. 2010, pp. 120-124.

[13] Lyon, David (1994), "The electronic Eye – The rise of surveillance Society", Cambridge: Polity Press.

[14] Lyon, David (2001), "Surveillance Society – Monitoring every day life", Buckingham: Open University Press.

[15] Lyon, David (1994), "The electronic Eye – The rise of surveillance Society", Cambridge: Polity Press.

[16] Lyon, David (2001), "Surveillance Society – Monitoring every day life", Buckingham: Open University Press.

[17] Lyon, David (2001), "Surveillance Society – Monitoring every day life", Buckingham: Open University Press.

[18] Burns, Tom e Flam, Helena (2000), "Sistemas de Regras Sociais – Teorias e Aplicações", Oeiras: Celta.

[19] Giddens, Anthony (1987), "The Nation-State and Violence", Cambridge: Polity Press.

[20] Castel, Robert (1991), "From Dangerousness to Risk" in G.Bruchell, C. Gordon e P. Miller (eds.), The Foucault Effect: Studies in Governamentality with Two Lectures by and Interview with Michel Foucault, Chicago: University of Chicago Press, pp. 281-298.

[21] Davies, Simon (1992), "Big Brother: Australia's growing web of surveillance", Sydney: Simon & Schuster, p. IV.

[22] Marx, Gary (1988), "Undercover: Police Surveillance in America", Berkeley: University of California Press.

[23] Lyon, David (1994), "The electronic Eye – The rise of surveillance Society", Cambridge: Polity Press.

[24] Norris, Clive (2003),From personal to digital: CCTV, the panopticon, and the technological mediation of suspicion and social control" in David Lyon (ed.), Surveillance as Social Sorting – Privacy, Risk and Digital Discrimination, London: Routledge, pp. 249-281; p. 249.

[25] Bennett, Colin; Raab, Charles, and Regan, Priscilla (2003), "People and place: patterns of individual identification within intelligent transportation systems" in David Lyon (ed.), Surveillance as Social Sorting – Privacy, Risk and Digital Discrimination, London: Routledge, pp. 153-175.

[26] Bennett, Colin; Raab, Charles and Regan, Priscilla. (2003), "People and place: patterns of individual identification within intelligent transportation systems" in David Lyon (ed.), Surveillance as Social Sorting – Privacy, Risk and Digital Discrimination, London: Routledge, pp. 153-175.

[27] Ploeg, Irma Van Der (2003), "Biometrics and the body as information: normative issues of the socio-technical coding of the body" in David Lyon (ed.), Surveillance as Social Sorting – Privacy, Risk and Digital Discrimination, London: Routledge, pp. 57-73.

[28] Nelkin, Dorothy; Andrews, Lori (1993), "Surveillance creep in the genetic age" in David Lyon (ed.), Surveillance as Social Sorting – Privacy, Risk and Digital Discrimination, London: Routledge, pp. 94-110.

[29] Lyon, David (2009), "Identifying Citizens – ID Cards as Surveillance", Cambridge: Polity Press.

[30] Stalder, Felix; Lyon, David (2003) "Electronic identity cards and social classification" in David Lyon (ed.), Surveillance as Social Sorting – Privacy, Risk and Digital Discrimination, London: Routledge, pp. 77-93.

[31] Marx, Gary (2008), "Vigilância Soft" in Catarina Frois (orgª.) A Sociedade Vigilante – Ensaios sobre identificação,vigilância e privacidade, Lisboa: ICS, pp. 87-109.

[32] Lyon, David (2001), "Surveillance Society – Monitoring every day life", Buckingham: Open University Press.

[33] Marx, Gary (2008), "Vigilância Soft" in Catarina Frois (orgª.) A Sociedade Vigilante – Ensaios sobre identificação,vigilância e privacidade, Lisboa: ICS, pp. 87-109.

[34] Lyon, David (1994), "The electronic Eye – The rise of surveillance Society", Cambridge: Polity Press.

[35] Marx, Gary (2008) "Vigilância Soft" in Catarina Frois (orgª.) A Sociedade Vigilante – Ensaios sobre identificação,vigilância e privacidade, Lisboa: ICS, pp. 87-109.

[36] Marx, Gary (2008) "Vigilância Soft" in Catarina Frois (orgª.) A Sociedade Vigilante – Ensaios sobre identificação,vigilância e privacidade, Lisboa: ICS, pp. 87-109.

[37] Lyon, David (2001), "Surveillance Society – Monitoring every day life", Buckingham: Open University Press.

[38] Lyon, David (1994), "The electronic Eye – The rise of surveillance Society", Cambridge: Polity Press.

Pharmaceutical Care: A Challenge for EAI in Pharmacies

Juha Puustjärvi
University of Helsinki
Department of Computer Science
Helsinki, Finland
juha.puustjarvi@cs.helsinki.fi

Leena Puustjärvi
The Pharmacy of Kaivopuisto
Helsinki, Finland
leena.puustjarvi@kolumbus.fi

***Abstract*—A new trend in pharmacy practice is to move away from its original focus on medicine supply towards a more inclusive focus on patient centered care. This new trend is named pharmaceutical care. It emphasizes the responsible provision of drug therapy for various purpose of achieving definite outcomes that improve or maintain patient's quality of life. In order to achieve these goals pharmacists are expected to assume many different functions including caregiver, communicator, teacher, life-long learner and manager. These functions set new challenges for pharmacy's information systems usability as well as pharmacist's skills to use them. A problem here is that existing pharmacies' information systems are isolated in the sense that they have their own data stores that cannot be accessed by other information systems used in pharmacies. As a result a pharmacist is burdened by accessing many systems inside a user task, e.g., in giving drug therapy for a patient. In order to alleviate this problem we have designed a knowledge-oriented EAI (Enterprise Application Integration) strategy for pharmacy's information systems. Our key idea is to revolve all pharmacy systems and applications around the shared knowledge base, and in this way avoid the problems of isolated non-interoperable systems and heterogeneous replicated data. In this paper, we present how knowledge-oriented EAI strategy can be exploited in contributing pharmaceutical care. In particular, we present how Semantic Web technologies such as RDF and OWL are intertwined in our developed solutions.**

Keywords - internet; semantic web; OWL; RDF; web-based applications; EAI; knowledge managememnt; e-health; ontologies

I. INTRODUCTION

During the last years there has been a trend for pharmacy practice to move away from its original role on drug supply towards a more inclusive focus on patient care [1, 2, 3]. This new trend in pharmacy practice is named *pharmaceutical care* [4, 5, 6]. It emphasizes the responsible provision of drug therapy for the purpose of achieving definite outcomes that improve patient's quality of life [7, 8].

The provision of pharmaceutical care sets significant challenges on pharmacy's ICT and pharmacists' skills as healthcare is a field where the fast development of drug treatment and the introduction of new drugs require specialized skills and knowledge that need to be renewed frequently [9, 10].

In particular, the amount of new information concerning medication increases rapidly. Pharmacies receive this information from a variety of sources [11], e.g., from medical authority, medicinal wholesalers, educational organizations and pharmaceutical companies. These information entities arrive in variety formats, e.g., by paper mail, e-mail, and fax. The nature of the information entities may vary, e.g., an information entity may be a learning object, a regulation, a guide or a bulletin [12]. Further, some of the information gives rise for a new business rule or changing prevailing rules and practices.

Two interesting questions arising from this reality are:

- How pharmacy systems should interoperate in order to support pharmaceutical care?
- How medicinal information should be organized in order to ensure ready access for pharmacists?

In ICT (Information and Communication Technology) this kind of issues are analyzed in the disciplines of *Enterprise Application Integration* (EAI) [13] and *Knowledge Management* (KM) [14]. The former is a strategic approach to binding many information systems together and supporting their ability to exchange information and leverage processes in real time while the latter concerns with acquiring, organizing and retrieving information within an organization.

Although current pharmacy systems have proven to be valuable and powerful systems in traditional pharmacy's role [15], they fail in contributing pharmaceutical care. This is due to the fact that existing pharmacy systems are monolithic isolated systems that cannot be easily modified to be able to interoperate with relevant external sources.

Providing pharmaceutical care implies that pharmacy applications should be able to cross organizational boundaries, and in this sense they should be open. In open systems the configuration of the whole system can change dynamically. For example in supporting drug therapy, by the permission of a patient, the pharmacist should be able to access patient's PHR (Personal Health Record) [16, 17] stored in a server in internet.

Nowadays, a common practice in pharmacies information management is that the incoming information entities are not stored at all, or are stored in a variety of systems such as in Document Management Systems, Learning Content Management Systems, Content Management Systems, Database Systems and Customer Relationship Management Systems [14].

The problem here is that the same information may be stored in separate systems, and each system is hardcoded to only work with its own data. Using such systems is overly complicated and thereby does not contribute to pharmaceutical care.

We have analyzed various EAI and KM strategies' suitability for supporting pharmaceutical care. Our key idea has been the integration of information oriented EAI strategy [18] and knowledge management strategies, which exploits semantic web technologies. We call such a strategy *semantic EAI.*

Through semantic EAI we can avoid the common problems of the legacy systems, i.e., the problems of the systems that are built based on proprietary solutions, developed in piecemeal way, and tightly coupled through ad hoc means, and which thereby have many duplicated functions, and which are non-interoperable.

In this paper, we first in Section II give an overview of the notion of pharmaceutical care. Then, in Section III, we motivate our proposed solutions by considering the interoperability problems of existing pharmacy systems. In Section IV, we consider EAI strategies focusing on semantic EAI. We first describe our used way for integrating heterogeneous data into the format that is compatible with the pharmacy's knowledge base. Then we consider portal-oriented application integration, where a pharmacist can view a multitude of systems through a single user interface. Finally, Section V concludes the paper by discussing from pharmacy's point of view the advantages and disadvantages of semantic EAI.

II. PHARMACEUTICAL CARE AND THE ROLES OF PHARMACISTS

The International Pharmaceutical Federation (FIP) defines pharmaceutical care to be the responsible provision of drug therapy for various purpose of achieving definite outcomes that improve or maintain a patient's quality of life [1].

The introduction of pharmaceutical care sets many new requirements on pharmacy's information systems as well as for pharmacists' skills and knowledge [4]. In this context the notion of "seven-star pharmacist" is taken up by FIP in its policy statement on Good Pharmacy Education Practice. It covers the following seven roles of a pharmacist [1]:

- *Caregiver*: Pharmacists have to view their practice as integrated and continuous with those of the health care system and other health professionals.

- *Decision maker*: The appropriate use of resources such as medicines, chemicals, equipments and practices should be the foundation of the pharmacist's work.

- *Communicator*: The pharmacist provides a link between prescriber and patient, and so he or she must be knowledgeable and confident while interacting with other health professionals.

- *Manager*: Pharmacists must be able to manage effectively various resources such as human, financial and information resources. In particular ICT will provide challenges as pharmacists assume greater responsibility for sharing information about medicines and related products.

- *Life-long learner*: The role of continuing education and lifelong learning is becoming still more important as the fast development of technologies requires specialized skills that need to be renewed frequently.

- Teacher: The pharmacist has a responsible to assist with the education and training of future generation of pharmacists and the public. Teaching as well as pharmacist life-long learning assumes the exploitation of modern e-learning technologies.

- *Leader*: In areas where other care providers are in short supply or non-existent the pharmacist is obligated to assume a leadership position in the overall welfare of the patient.

For each of these roles there are many specific ICT-based systems and disciplines including ERP (Enterprise Resource Planning), CRM (Customer Relationship management), SCM (Supply Chain Management), LMS Learning Content Management) and CM (Content Management System) that are developed for improving the management of these roles. We next consider such systems and their interoperability requirements in supporting pharmaceutical care.

III. SEGMENTATION IN PHARMACY SYSTEMS

Integrating information in an ever growing internetworking world is likely to be the most urgent need for any kind of business, trade or science. Daily work in supporting pharmaceutical care is not an exception: a huge and increasing amount of heterogeneous medicinal data is distributed over network of computing system, and the usage of this heterogeneous data requires the introduction of interoperability and integration technologies.

Basically the term *integration* refers to the idea of putting diverse concepts together to create an integrated whole. Instead *interoperation* refers to making applications work together by sharing the appropriate messages but without any single conceptual integration. Further, *semantic integration* means that a new ontology is derived from existing ontologies such that the new ontology facilitates the interoperability of the systems [19].

Initially the problem of software and data segmentation was increased with the introduction of commercial off-the-shelf applications such as ERP, SCM and CRM systems

[18]. In particular the early versions of these systems were designed as self-contained boxes with no means for accessing internal data or processes. Though the new versions of the pharmacy systems provide better access to their data, integrating them with other systems in pharmacies is still problematic.

As a result, many pharmacies, as well as other organizations, have an environment of disparate legacy systems and applications, which typically interact by a number of connections that are poorly documented and difficult to maintain. Such systems also have their own isolated data stores and user interfaces. Such architecture of a pharmacy system is illustrated in Fig. 1.

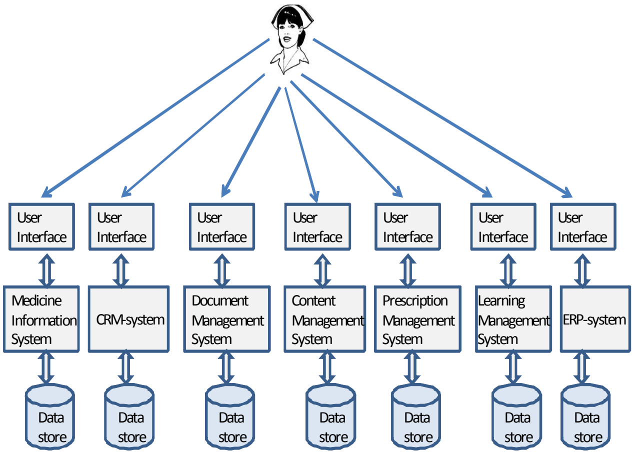


Figure 1. A segmented architecture of pharmacy system.

If the data needed in a user task is not included in one system, then the user has to log and access data from various systems. Further, if the systems do not support SSO (Single Sign-On) this may require many user activities.

To illustrate this difficulty let us assume that a pharmacist is dispensing a medicinal product, say *Diovan*, for a patient. First, in order to store the dispensation of the prescription, the pharmacist opens the prescription system. Then from customers social insurance card the pharmacist recognizes that the patient is a veteran. The pharmacist remembers that the discount of ten per cent is granted for veterans of some drugs, but she does not remember whether it concerns *Diovan.*

By accessing the business rule management system by the keywords "discounts" and "veterans" the system return the statement "discount of 10 per cent of the prescription based partially repayable drugs is granted for veterans". Next the pharmacist opens and finds from the medicinal system that *Diovan* is not a partially repayable drug. Then she retrieves from the pricing system the medicinal products that are cheaper and are substitutable with *Diovan* and are partially repayable. After this, by the permission of the client the pharmacist change *Diovan* to a substitutable medicinal product *Valsartan* and grants the discount of 15 per cent for the customer. Finally, from the drug therapy system the pharmacist prints the instructions concerning *Valsartan* and gives them to the patient.

The reason for the complexity of the pharmacist user task is that the data needed by a pharmacist is stored in separate systems, and their data cannot be accessed by other systems.

IV. SEMANTIC EAI

The need of interoperation within organization's systems led to the evolution of EII (Enterprise Information Integration) and EAI. The early work on addressing the challenges of heterogeneity involved integrating different data sources. Such a work was called EII. Later on there has been work on EAI where information flows between applications is addressed. The EAI infrastructure allows systems and applications throughout an enterprise to seamlessly communicate with one another in realtime.

EAI solutions can exist at many levels such as database level, process level, portal level and method (function call) level. Hence, the principal distinction between Information-oriented, Process-oriented and Service-oriented and Portal-oriented application integration is been done [18]:

- In Information-oriented approaches applications interoperate through a data store.
- In Process-oriented (also called workflow–oriented) approach the interoperation is controlled through a process model that binds processes and information within many systems.
- In Service-oriented interoperation applications share methods (e.g., through Web service interface) by providing the infrastructure for such method sharing.
- In Portal-oriented application integration a multitude of systems can be viewed through a single user interface, i.e., the interfaces of a multitude of systems are captured in a portal that user access by their browsers.

Further, in the emergence of many new technologies based on Web services and Semantic Web, there are still more changes for EAI each change having its limitations and opportunities.

In our developed architecture we use the Information-oriented approach in achieving the interoperability between the pharmacy's systems and applications. That is, the systems and applications interoperate through sharing a knowledge base, and the ontology is developed by integrating the ontologies of the interoperable systems and applications.

From user's point of view our used application integration strategy follows the Portal oriented approach as the multitude of pharmacy's systems and applications can be viewed through a single user interface.

In the following two subsections we give a more detailed description of these strategies.

A. *Information-Oriented Integration Using Knowledge Management System*

Knowledge management system refers to a computer based system for managing knowledge in organizations [14]. A knowledge base is a special kind of database for knowledge management. It provides the means for the computerized collection, organization, and retrieval of knowledge for various applications.

Today an ever expanding set of knowledge management systems are using the technologies of the Semantic Web [20]. That is, knowledge is organized according to ontologies, and automated tools are used in accessing and maintaining knowledge [21].

The idea of sharing pharmacy's knowledge base is illustrated in Fig. 2.

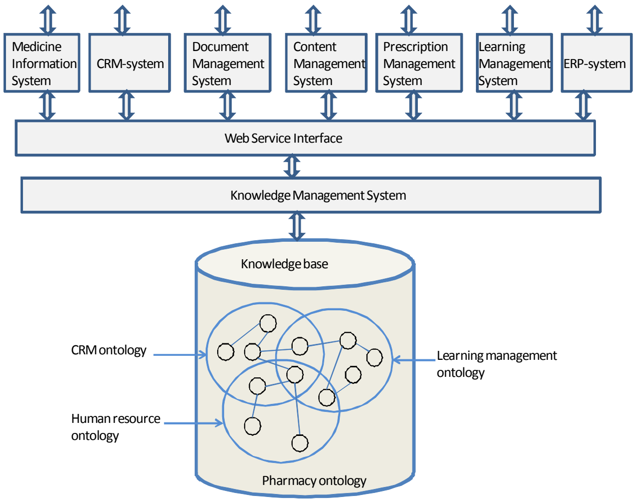

Medicine Information System
CRM-system
Document Management System
Content Management System
Prescription Management System
Learning Management System
ERP-system
Web Service Interface
Knowledge Management System
Knowledge base
CRM ontology
Learning management ontology
Human resource ontology
Pharmacy ontology

Figure 2. The architecture of semantic EAI in a pharmacy.

We follow the idea of knowledge centric organizations [14], where, the key idea is to revolve all applications around the shared ontology. In our case, as illustrated in Fig. 2, it means the integration of the data repositories of the pharmacy's systems such as business rule management system, content management system and learning object management system. To the integrated ontology we refer by the term *pharmacy ontology*. So the pharmacy's systems can seamlessly interoperate through accessing the shared pharmacy ontology.

The data of the pharmacy ontology is received and gathered up from a variety of health care organizations. However before the data can be inserted into the knowledge base (pharmacy ontology) it must be transformed into RDF/XML-format [22] that is compatible with the pharmacy ontology. Such transformations require that a specific stylesheet [14] is developed for each input data that is not compatible with the pharmacy ontology. A language associated with stylesheets is XSLT (Extensible Stylesheet Language) [23]. It is a markup language that uses template rules to specify how a style sheet processor transforms an XML document.

In order to illustrate this transformation assume that the graphical ontology of Fig. 3 is a subset of the pharmacy ontology. In this graphical ontology, ellipses represent classes and boxes represent properties.

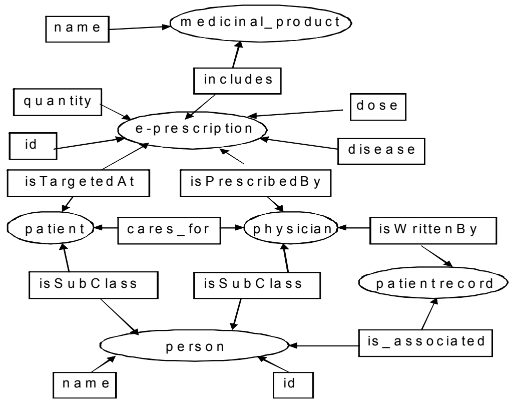


Figure 3. A subset of the pharmacy ontology in a graphical form.

Assume now that the pharmacy system receives the XML coded document presented Fig. 4. As the document is in XML it does not contain any semantics, and so it is transformed by a stylesheet engine into RDF/XML-format that is compatible with the ontology of Fig. 3. Such an RDF document is graphically presented in Fig. 5, and in RDF/XML format in Fig. 6.

```
<prescription>
      <prescription_id>abc123</prescription_id>
      <patient>
                <name>John Smith </name>
                <id> 1465766677</id>
      </patient>
      <medicinal_product>Panadol</medicinal_product>
      <disease>fever</disease>
      <quantity>30</quantity>
      <dose>One tablet three times a day</dose>
      <physician>
                <name>Lisa Taylor </name>
                <id> 98765432</id>
      </physician>
</prescription>
```

Figure 4. A precription in XML format.

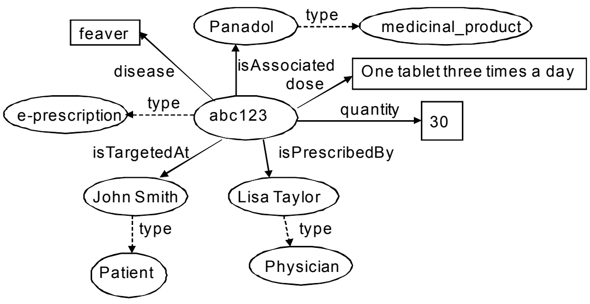


Figure 5. A precrition presented by RDF-graph.

```
<rdf:RDF
    xmlns : rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns : xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns : mo="http://www.lut.fi/ontologies/montology#"
    <rdf:Description rdf:about="abc123">
        <rdf:type rdf:resource="&mo;e-prescription"/>
            <mo : dose>One tablet three ti mes a day</mo : dose>
            <mo : quantity rdf:datatype="&xsd;integer">30</mo : quantity>
            <mo: includes>Panadol</mo: includes>
    </rdf : Description>
    <rdf:Description rdf:about="1465766677">
       <rdf:type rdf:resource="&mo;patient"/>
            <mo : name>John Smith</mo : name>
    </rdf : Description>
    <rdf:Description rdf:about="98765432">
    <rdf:type rdf:resource="&mo;physician"/>
            <mo : name>Lisa  Taylor</mo: name>
    </rdf : Description>
</rdf:RDF>
```

Figure 6. A prescription in RDF/XML format.

B. *Portal-Oriented Integration Using ASP*

In Portal-oriented application integration a multitude of systems can be viewed through a single user interface, i.e., the interfaces of a multitude of systems are captured in a portal that user access by their browsers (Fig. 7).

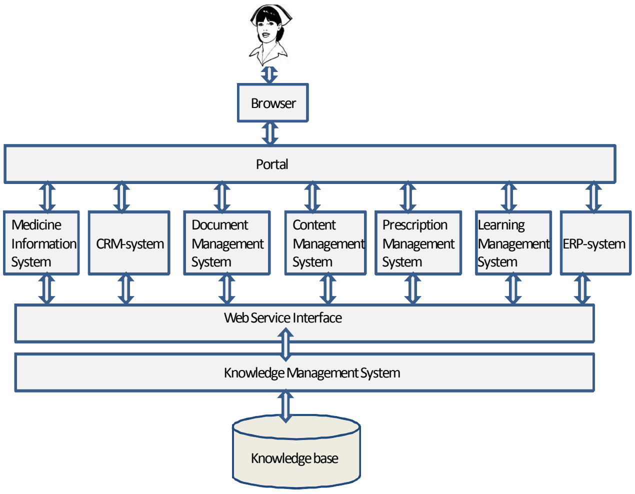

Browser
Portal
Medicine Information System
CRM-system
Document Management System
Content Management System
Prescription Management System
Learning Management System
ERP-system
Web Service Interface
Knowledge Management System
Knowledge base

Figure 7. The architecture of portal-oriented EAI .

User (pharmacist) interacts with the pharmacy's systems and applications by a browser, and so all the data for a pharmacist are presented in HTML. However, all the content in the knowledge base are stored OWL (Web Ontology Language) format [24], i.e., represented by RDF/XML syntax.

A significant gain of accessing the portal through a browser is that instead of the proprietary pharmacy systems the whole pharmacy system can be implemented by exploiting the notions of ASP (Application Service Provider).

An ASP is a business that provides computer-based services to customers over a network. The application software resides on the vendor's system and is accessed by users through a web browser using HTML or by special purpose client software provided by the vendor.

The need for ASPs has evolved from the increasing costs of specialized software that have far exceeded the price range of small to medium sized businesses such as pharmacies. Also the growing complexities of software have led to huge costs in distributing the software to end-users. Through ASPs, the complexities and costs of pharmacy's software can be decreased.

On the other hand ASP strategy is criticized since users cannot modify the software they use. However, in pharmacies this is not a problem as they do not have personnel for software development. In some cases organizations have rejected ASP strategy since organizations data is being stored, and controlled, by third parties, thus increasing its attack surface. Within pharmacies the critical data are patients' prescriptions that are received from third parties such as prescription holding stores and physicians, and so the control of third parties cannot be avoided.

V. CONCLUSIONS

Pharmaceutical care sets significant challenges on pharmacy's ICT and pharmacists' skills as healthcare is a field where the fast development of drug treatment and the introduction of new drugs require specialized skills and knowledge that need to be renewed frequently.

On the other hand, pharmaceutical care does not exist in isolation from other health care services: it should be provided in collaboration with patients, physician, nurses and other health care providers. As each of these occupational groups has their own information systems also the medicinal information systems and applications developed for pharmaceutical care should be able to co-operate with these systems.

In ICT these kinds of issues are analyzed in the disciplines of Enterprise Application Integration and Knowledge Management. In the emergence of many new technologies based on Web services and Semantic Web, there are new chances for EAI and KM that are appropriate for providing pharmaceutical care. The key point in our presented semantic (ontology based) EAI is that all parties have a common understanding of the pharmacy ontology on which the exchanged messages (documents) are based on.

In order to exchange documents the parties need a common language through which to exchange documents between their computer systems. Though XML is rapidly becoming the key standard for data representation and transportation, XML-documents themselves do not capture any semantics. Thereby the introduction XML-messaging requires that communicating applications must be hard-coded meaning that the semantics of the message are coded in the communicating applications.

A more flexible way for achieving consensus on exchanged messages is to develop appropriate domain ontology, and use it as a vocabulary in exchanging RDF/XML coded documents. Essentially the developed ontology must be shared and consensual terminology among the communicating parties as it is used for information sharing and exchange. Hence, the corner stone of our presented solutions is that all the communicating parties must commit to the same ontology, e.g., to our developed pharmacy ontology.

REFERENCES

[1] K. Wiedenmayer, R. S. Summers, C. A. Mackie, A. g. Gous, M. Everard, and D. Tromp, “Developing pharmacy practice”, World Health Organization and International Pharmaceutical Federation 2006

[2] van Mil JW, Schulz M, Tromp TF. Pharmaceutical care, European developments in concepts, implementation, teaching, and research: a review. Pharm World Sci. 2004 Dec; 26(6):303–11.

[3] C. Hepler and Strand L., Opportunities and responsibilities in pharmaceutical care. Am J Hosp Pharm 1990;47:533–43.

[4] The role of the pharmacist in the health care system. Preparing the future pharmacist: Curricular develop -ment. Report of a third WHO Consultative Group on the role of the pharmacist, Vancouver, Canada, 27–29 August 1997. Geneva: World Health Organization; 1997. WHO/PHARM/97/599. Available at: http://www.who.int/medicinedocs/

[5] Hepler C., Clinical pharmacy, pharmaceutical care, and the quality of drug therapy. Pharmacotherapy. 2004 Nov; 24(11):1491–98.

[6] Berenguer B, La Casa C, de la Matta MJ, Martin- Calero MJ. Pharmaceutical care: past, present andfuture. Curr Pharm Des 2004;10(31):3931–46.

[7] Shumock GT, Butler MG, Meek PD, Vermeulen LC, Arondekar BV, Bauman JL. Evidence of the economicbenefit of clinical pharmacy services: 1996–2000. Pharmacotherapy 2003; 23(1): 113–32.

[8] Strand L, Cipolle R., Morley PC, Frakes MJ. The impact of pharmaceutical care practice on the practitioner and the patient in the ambulatory practice setting: twenty-five years of experience. Curr Pharm Des 2004;10(31):3987–4001.

[9] Tietze K. Clinical skills for pharmacists. A patient-focused approach, Mosby Inc. USA, 1997. Medicines Partnership UK, www.medicines-partnership.org

[10] Walker R. Pharmaceutical public health: the end of pharmaceutical care? Pharm J 2000; 264:340–341.

[11] Puustjärvi J. and Puustjärvi L., Managing Medicinal Instructions. In the proc. of the International Conference on Health Informatics (HEALTHINF 2009). pp. 105-110. 2009.

[12] Puustjärvi J.,and Puustjärvi L., The role of medicinal ontologies in querying and exchanging pharmaceutical information. International Journal of Electronic Healthcare, Vol. 5, No.1 pp. 1 – 13. 2009.

[13] Singh, M. and Huhns M., Service Oriented Computing: Semantics, Processes, Agents. John Wiley &Sons, Ltd. 2005.

[14] M. Daconta, L. Obrst, and K. Smith. The semantic web: a guide to the future of XML, web services and knowledge management. Indianapolis: John Wiley & Sons. 2003.

[15] Puustjärvi, J. and Puustjärv L.,. Towards Semantic Exchange of Clinical Documents, International Journal on Advances in Life Sciences (IJALS). Vol. 1, No.2&3 pp. 69 – 76. 2009.

[16] Raisinghani M.S. and Young, E., Personal health records: key adoption issues and implications for management, International Journal of Electronic Healthcare. Vol. 4, No.1 pp.67-77. 2008.

[17] Lewis, D., Eysenbach. G., Kukafka, R., Stavri P.Z., and Jimison, H., Consumer health informatics: informing consumers and improving health care. New York: Springer. 2005.

[18] D. Davenport. B2B Application Integration. Addison-Wesley.2001.

[19] J. Puustjärvi, and L. Puustjärvi. Application Integration and Semantic Integration in Electronic Prescription Systems. International Journal of Computer Science Issues (IJCSI). Vol. 7, Issue 3, No 2, pp. 1-8. 2010.

[20] J. Davies, D. Fensel, and F. Harmelen, Towards the semantic web: ontology driven knowledge management. West Sussex: John Wiley & Sons.2002.

[21] G. Antoniou, & F. Harmelen. A semantic web primer. The MIT Press. 2004.

[22] RDF – Resource Description Language. Available at: http://www.w3.org/RDF/

[23] E. Harold and W. Scott Means W., XML in a Nutshell. O’Reilly & Associates, 2002.

[24] OWL – WEB OntologyLanguage. Available at: http://www.w3.org/TR/owl-features/.

ISBN: 978-1-61208-116-8

Value for Users in Social Media Services - a Framework Walkthrough

Sampo Teräs
Department of Computer Science and Engineering
Aalto University
Espoo, Finland
e-mail: sampo.teras@tkk.fi

***Abstract*— This paper presents several key values for users in social media services. Based on Kujala and Väänänen-Vainio-Mattila's value framework, a walkthrough was conducted on ten social media services of different types. We also analyzed which elements are needed in the service to produce these values. The main findings are that social media services mainly produce social life, creativity, and emotional values. Key elements that the services required were contacts, media sharing, an opportunity to provide feedback, and profile or character creation. By enabling these features, value can be created for users. The perceived value still depends on how the user utilizes the particular service, and how the other users interact with him/her.**

Keywords - Value; value-in-use; social media services; walkthrough.

I. INTRODUCTION

The popularity of the social media services is startling. Currently, there are over 400 million Facebook users around the world [1]. There are over three million articles in English in Wikipedia, and millions more in other languages [2]. The social media services have raised the Internet up to a new level. Rather than merely gathering and using information, users are now also producing and sharing information. The audience is also now much broader. Instead of being limited to technology driven youth, the Internet is used almost by everybody. The average Internet gamer is no longer an Internet-savvy youngster, but a middle-aged parent instead [3].

Why are the social media services so successful and popular? What is the value that these leisure time services provide to millions of users who access their services several times per day? For example, Google is the main advertising service on the Internet, but Facebook is gaining on Google due its vast number of members [4]. Why do people spend several hours per day playing online games and adding information to Wikipedia without getting paid?

Based on these phenomena, two research questions were formulated. 1) What forms the value in the social media services? 2) How is that value provided to users in social media services?

Since the service creation and use are interlinked, the value-in-use is linked to the value's formation [5]. In this research, the elements that create value in social media services will be examined. These elements can be functions or properties of the system, but they can also be demands from the community, users, or the use itself.

The popularity of social media services, and the fact that they are mainly used voluntary during leisure time, suggests that the services provide a significant amount of value to the users. However, it is not entirely clear what these values are.

Because some value is provided to the user, there must also be some elements within the services that enable this value creation. These elements can be functions in the system, but they also can be demands from either the user or the community.

The related research is presented in section II. The research methods are described in detail in section III. Section IV contains the results of the research, and section V presents the discussion and conclusion.

II. RELATED RESEARCH

Cockton wrote that, "we should judge systems by what endures beyond the interaction" [6]. This means that it is important to find the reasons for the use, rather than only focusing on the use itself. Instead of measuring the quality of the system, the system's worth should be considered. In this paper, the system is a software centered service system, a system that focuses on supporting customer's activities and processes [5]. Here, the value does not only mean the economical return from the use, but other aspects as well (for example, social benefits and knowledge increase, or simply having a pleasant time).

Grönroos noted that service creation is partly customer-driven, partly competition-driven, and partly technology-driven [5]. Clearly, the technology is built on Internet service and computers. Web 2.0 and social media are not strictly described, but involve several central elements [7]. The service improves when there are more users [7]. For example, in Facebook, the information increases when users have more contacts – other users that are connected to the user. The web pages are no longer static sources of information. Instead, they are dynamic [7]. Information changes constantly, so there is a practical reason for visiting the pages more than once. The users should be treated as co-writers of the information [7]. They produce most of the content in the Web 2.0 services, while the service provider mainly supplies the tools to share the information. Most of the social media services are quite simple [7]. The service needs to be simple to use so that it will reach a wide audience and can be accessed on a daily basis. Usually, the social media services give something valuable for free, something that used to be chargeable [7].

Value is created during the use of these services. For example, there is value for the customer when a shop assistant helps him or her. There is no value from the assistant's existence to the customer while he or she is not helping the customer. In addition, the value comes from a long time customer-provider relationship [5]. Rather than a single purchase, it is much more profitable to both parties if the customer pays the provider for a longer period of service, while knowing that he or she will get quality service anytime or anywhere it is needed.

There are several definitions for the word value. The Webster dictionary has included the following definitions for value: “1) a fair return or equivalent in goods, services, or money for something exchanged, 2) the monetary worth of something: market price, 3) relative worth, utility, or importance <a good value at the price> <the value of base stealing in baseball> <had nothing of value to say>, 7) something (as a principle or quality) intrinsically valuable or desirable <sought material values instead of human values — W. H. Jones>” [8]. Pietarinen defined value as a formula V(X,S,P), where X has a value V to S because of P [9]. Thus, the value is dependent upon the receiver. Others might find different, or even no value from functions that are extremely valuable to others. He also stated that values can be divided into two general categories, final values, and vicarious values that lead to these final values [9].

Maslow defined the hierarchy of human needs [10]. The physiological needs, such as air and nourishment, are at the base of this hierarchy. Above these are safety needs and social needs, such as security and friends. After that comes the esteem needs, such as reputation. Finally, self-actualization is needed. Later, Alderfer proposed the ERG theory (Existence, Relatedness, and Growth) to create a hierarchy of behavior influences [10]. The three categories in ERG theory are existence needs, relatedness needs, and growth needs. However, value can be considered more than just fulfilling the needs of users.

Boztepe divided user value into utility, social significance, emotional, and spiritual categories [12]. Utility is divided into convenience, quality and performance, and economy. It defines the effectiveness and efficiency similarly to the definitions in the ISO 9421-11 standard [13]. It takes the ergonomic sides, such as accessibility and physical comfort, into consideration. There is also the economic viewpoint for value, defined as use economy.

Social significance is divided into social prestige and identity [12]. These define the social relationships in the use, such as impression management, role fulfillment, and group belongingness. Humans tend to ascribe importance to the community to which they belong. They want to be a part of it, and hopefully an important part. Nevertheless, they still want to have their identity and uniqueness. A good example of meaning of the identity is that the customer’s commitment to a certain brand is dependent on the identity values it provides [14].

The emotional category is divided into emotion and sentimentality [12]. These values define the fun, enjoyment, and memories of the service or system. Along with this, there is also a spiritual value that is important for those who have strong faith. For example, some religious groups do not use certain machines for religious reasons.

There are several other definitions and categorizations of user value. Kujala and Väänänen-Vainio-Mattila developed a combined framework based on the ones mentioned above and several other publications [5]. This framework is based on seven value divisions that are shown in Table I, with keywords selected from the values description. The framework was chosen because it is rather novel, and it contains the information collected from several different authors over several decades. The framework was also light and straightforward, permitting a walkthrough of ten services.

TABLE I. SIMPLIFIED VERSION OF THE KUJALA AND VÄÄNÄNEN-VAINIO-MATTILA’S VALUE FRAMEWORK. SEE [15] FOR THE FULL VERSION.

Value	Keywords
Social values	Relatedness, social and external esteems status, power, control and dominance, achievement, conformity, equality, helpfulness, honesty, loyalty. Respect, influence, power, social achievement, conformity.
Emotional, hedonic values	Aroused feeling or affective states, pleasure, fun, sensory, enjoyment. Positive feelings, pleasure and enjoyment, increase in emotional experiences, support in handling experiences and emotions and saving emotional occasions.
Stimulation and epistemic values	Excitement, experienced curiosity, novelty and gained knowledge. Increase in excitement.
Growth and self-actualization values	Self-actualization, creating, independent thought, and action. Creating new things, achieving internal esteem.
Traditional values	Respect, commitment, and acceptance of customs and ideas (projected to culture or religion). Task support, maintaining customs and ideas.
Safety values	Security, social order, healthy, comfort, freedom from fear. Protection and alarms, ease of use, familiarity.
Universal values	Understanding, appreciation, tolerance, and protection for all, welfare, nature. Ecological soundness, improving equality.

The seven value categories are: social values, emotional/hedonic values, stimulation and epistemic values, growth and self-actualization values, traditional values, safety values, and universal values. Social values describe the social side of the user’s life. These include having friends and other contacts, being appreciated and respected by the community, having some influence and power over others, and having some honesty and loyalty around. For example, when a person comes to a new work environment he or she needs to get related to the working community.

Emotional and hedonic values are the user’s feelings. These feelings include fun, enjoyment, and other pleasurable emotions. The positive feelings, and effectively handling their related experiences, are also major part of the emotional needs. For example, watching comedy is fun, although it might not be so productive.

Stimulation and epistemic values make the user aroused and excited. Finding something new and increasing knowledge are the key points in stimulation values. For example, many people are interested in traveling to other countries and seeing unfamiliar cultures.

Growth and self-actualization values describe the output of the user's inner self. Mainly, this refers to creativity, but it also means having individual thoughts and actions. For example, a child wants to put the clothes on by himself/herself, because he/she wants to show that he/she can do it without help.

Traditional values are related to the cultural habits of a group. People in a community might keep something quite valuable for historical reasons, although people outside of the community might not see it in the same way. For example, people might still support the home team, although they continue to lose.

Safety values describe the needs for security and comfort. These include mental security (for example, freedom from fear), and physical security (for example, being healthy). For example, people pay a significant amount of money to feel safe from burglars, fires, car accidents, and diseases.

Universal values describe the common good, and include: public tolerance, welfare, protecting nature, understanding others, and having equality. For example, people donate food to the poor because they believe it is the right thing to do.

The value created for users can be from one or several of these categories. Furthermore, the impact might vary. Others find some categories to be more important than others.

III. RESEARCH METHODS

The research method involved employing a value framework and analyzing several social media services reflected by this framework. Below are the descriptions of the chosen services and the framework used in the walkthrough.

A. *Services*

According to Cavazza's Social Media Landscape, the chosen services are in the following groups: social platforms, content sharing, discussion expressing, publication expressing, business to business networking, casual gaming and multiplayer online gaming [16]. The chosen services were: YouTube, Wikipedia, Facebook, LinkedIn, Combat Arms EU, Kongregate, Picasa, Habbo, Last-fm, and several different phpBB-forums. The services are introduced in Table II. There were five different phpBB-forums used in the research, varying from 400 to 96000 members under the topics of sports, games, and politics.

TABLE II. LIST OF SERVICES AND THEIR DESCRIPTIONS. THE DESCRIPTION IS RETRIEVED FROM THE SERVICE PROVIDER'S WEB PAGES (8.4.2010) AND THE CATEGORY IS BASED ON THE CAVAZZA'S SOCIAL MEDIA LANDSCAPE [16]

Service	**Description**	**Category**	**URL**
Combat Arms EU	"Free to play intense multiplayer action."	MOG gaming	http://combatarms.nexoneu.com/Intro.aspx
Facebook	"Facebook helps you connect and share with the people in your life"	Social platform	http://www.facebook.com/
Habbo	"Make friends, join the fun, get noticed." (Note, the Finnish version of service was reviewed)	3D discussion expressing	http://www.habbo.com/
Kongregate	"Thousands of the best free games, and our community features mean that playing them here is more satisfying than anywhere else. If you do well in our games you'll earn badges, level up, and gain the respect of your friends."	Casual games gaming	http://www.kongregate.com/
Last-Fm	"Last.fm recommends music, videos and concerts based on what you listen to."	Music content sharing	http://www.last.fm/
LinkedIn	"Over 60 million professionals use LinkedIn to exchange information, ideas and opportunities."	BtoB networking	http://www.linkedin.com/
phpBB-forums	"The most widely used Open Source forum solution."	Forum discussion expressing	http://www.phpbb.com/
Picasa	"Picasa is free photo editing software from Google that makes your pictures look great. Sharing your best photos with friends and family is as easy as pressing a button! "	Photo content sharing	http://picasa.google.com/
Wikipedia	"The free encyclopedia that anyone can edit."	Wiki publication expressing	http://en.wikipedia.org/wiki/Main_Page
Youtube	"YouTube is the world's most popular online video community, allowing millions of people to discover, watch and share originally-created videos."	Video content sharing	http://www.youtube.com/t/about

B. *Walkthrough*

The research process contained two steps. In the first step, the researcher created an account and added some content to the service. The purpose was to actually see the value that the service provides. It takes some time to get used to a new service and find all of the elements that it provides. Moreover, the value does not actually exist until a certain amount of time has passed, or until enough contacts are added. For example, it is not possible to receive information about what is occurring before something actually happens. Further, chatting with another user is impossible before the other user logs on to the service. During this step, several observations about the value were made, but the findings were uncategorized. At this point, findings from other services were noted, but they were not emphasized as significantly as the ones from the reviewed services.

Once a basic understanding of the system was achieved, it was time to proceed to the next step. At that time, the services were formally reviewed with Kujala and Väänänen-Vainio-Mattila's value framework, which was described in the previous chapter. For each service, the value was reflected through each division of the value framework. The main focus was to find properties from the service to support the division's value creation. There were additional findings that reflected the researcher's use of the service. For example, in some services, the emotional values were not directly linked to clear properties, but assumptions had to be made. It was also hard to describe what type of content each user was looking for; thus, the researcher's subjective perspectives are included. There was an attempt to keep these at a minimum, but in this kind of walkthrough, it is rather impossible to remain completely objective. For example, it is rather personal to find something funny.

Findings are partly based on the users' observed behavior and comments. For example, a comment like, "That was funny," leads to an assumption that at least one user found the content to be humorous or emotionally stimulating. Several services were targeted to some user groups (for example, children) that the researcher could not really relate to. This led to some assumptions about what is valuable for the users, although not necessarily valuable for the reviewer.

IV. Results

After analyzing the services within the framework, the following results were achieved.

A. *Values*

Here are the phenomena listed by each value-category.

1) *Social values*

The main issues for social values involve having friends or contacts in the service, and getting respect or showing off one's own status. In several services, having an extensive number of friends or contacts has a functional effect. For example, in LinkedIn, the service is based on having connections. Further, in Facebook, the content shown on one's own homepage (or wall) is increased when there are more friends. In several games, having more allies makes it much easier to win. Nevertheless, in some contexts, having more friends can be seen as a status symbol of popularity.

In order to gain contacts or friends, the service must provide the networking property. Searching or browsing must be enabled. In many services, there are also suggestions based on the connections. There must be a function that adds users to one's own contacts, but there also has to be a function that blocks the invitations from strangers. Users usually acquire their friends from the real world, (for example, childhood friends, classmates and co-workers), but they can also be gained through the service. For example, joining a team in a game, or a group in Facebook, connects the group members together. This means that the service should provide some sort of grouping property.

Gaining respect and status can be provided by the users, the system, or with money. To gain status from other users, the service should provide some sort of appreciation function. With this function, the other users can inform that they like the actions of a certain user to the entire community. For example, one can like the posts in Facebook, one can rank the videos with stars in YouTube (this changed into thumbs during the research project), or add them to favorites, and the users can increase respect for other users in Habbo. It is extremely important that the user can gain recognition from their fellow users, not just from the system.

Status can also be elevated by the system. Many services provide achievements or levels; games in particular have these properties. Gaining enough experience points, kills, money, or other parameters can permit a character to move up a level or allow access to some special features. In other services, these types of status symbols also exist. For example, in many discussion forums, the amount of posts changes the user's status.

Finally, status can be improved with money. Many of the services are free, but special features are gained with purchases. Money can be changed into service currency, like Habbo-coins in Habbo, or NX-credits in Combat Arms EU, but special properties can also be achieved with money. For example, the user can bring his/her profile to the front page of the service. The service currency can then be changed into virtual property, such as clothes, furniture, and weapons. The usual methods for payment are PayPal, Credit card, money transfers, and SMS. This raises certain security issues, since real money is involved.

2) *Emotional values*

In social media services, sharing emotions and having fun are the two top emotional value creators. Sharing emotions, feelings, and experiences is quite common in micro blogging. Users tend to tell their friends if they are happy, sad, or irritated, and they like to report what happened during the day, or what will happen in the near future. Pictures, and sometimes, even videos, are used for sharing emotions and experiences. After sharing information, friends usually give the user feedback. They relate to the feeling and share their own thoughts about their experiences. To provide this, the service must have some channel to exchange information, whether it is text or another form of media. The possibility to comment or provide other feedback is essential to the creation process.

Most social services are related to leisure time, and aside from sharing information, they are used for having fun. One's own experiences and comments on such services are usually humorous, as are some shared links and images. Games can be also considered fun to play, and many of the services have games included. For example, Picasa has a

"Where in the World" guessing game, Facebook has several games, and in HABBO, there is the option of role-playing, such as a doctor or a parent. To provide these activities, the services require a level of real-time interaction or turn based sequences. There must also be enough graphics for the users to boost their imagination.

3) Stimulation

There are three major elements in social media services to increase stimulation for the user. The greatest one is information gaining, followed by excitement and meeting new people. There are several areas of interesting information. Users are usually interested in their friends, hobbies, work, and other information related to their lives.

The essential question is who is providing all of the necessary information? In many social media services, users themselves create the information. For example, in Wikipedia, the articles are written and edited by the users. In YouTube, users upload the videos, but they do not necessarily create the content for the videos. This might lead to copyright violations. Some users are quite active in terms of providing information, while others are not. For example, in one of the phpBB-forums, there were over 1300 registered users, but only 27 thousand messages; thus, it is likely that the majority simply read the forum.

One important aspect is the amount of information. If there is too much information, it is hard to find the essential information [17]. Fortunately, in the Web 2.0 world, the information is tagged rather than structured in tree form [7]. Thus, the provider of information can set a series of keywords to help others locate the correct information. This also leads to false advertising, where users add popular keywords when they are not related to the information.

Since the social media services are usually quite personal, the privacy factor is notable. When a user provides information about himself/herself, family, and friends, it is necessary to filter the information to certain groups. In Facebook, only registered members can see profiles and those who are not friends can see only the public part of a user's profile. In several services, it is also possible to send private messages to certain user(s).

Excitement is usually gained through achievements, or more specifically, by trying to reach them. As noted before, the status can be measured by these factors, and reaching a new level in a game or discussion forum is usually challenging. There are several other achievements involved, such as finalizing one's profile in LinkedIn or gaining subscribers in YouTube.

Other stimulation effects include: meeting new people and finding some other interesting things, such as new bands of one's favorite music genre. Usually, these are achieved with suggestions, browsing, or mere chance. For example, users can gain new friends while entering a random room.

4) Growth and self-actualization

Creativity is well supported in social media services. The user's self image is also elevated by profiles and characters. There are several different ways to share media in the services. Some even allow users to edit media. The biggest creativity factor is creating, mixing, and sharing media. The use in itself might be considered creative behavior. For example, in YouTube, the users can create playlists from their favorite videos. Last but not least, the users are able to give opinions and other comments related to the media. For example, in YouTube, viewers can suggest that amateur musicians make certain cover songs, or in discussion forums, users can open a new thread for discussion without knowing a great deal about the topic.

Self-actualization can be achieved through the profile. Usually, users try to learn about their strengths, which are then presented in text and images. The question is how honest the information actually is. Particularly, in games and virtual worlds, the users are not trying to copy themselves. Rather, they create some imaginary combination or use a common alias. Although there may be profiles or character information, the users are evaluated by their actions and behavior. For example, in Combat Arms EU, the player with the best stats might not be so popular if he/she cheats.

5) Traditional values

The only traditional values creator in social media services is acceptance. Users tend to use services in their own way. For example, they post things that they find interesting. Without feedback, the user cannot receive acceptance from the community; thus, the service should produce a channel for other users to inform their level of acceptance to the user.

6) Safety values

There are only a few safety values creators in social media services. Since most of the actions are electronic and virtual, physical safety is not really supported. Safety is primarily gained through awareness and information about the family and friends. For example, knowing that user's child is doing okay on a trip creates a calm feeling.

There are several safety issues related to social media services, such as privacy, family filters, and moderation. Since there are often various users on the service and some of them are minors, it is essential to have some protection from other users. Although the users create the content, there should be a moderator in the service. This is particularly true on open discussion forums, where the conversation might end up quite personal and hostile. Since the users are usually more or less anonymous, they can easily use foul language and make threats or insults.

7) Universal values

For universal value creation, social media services offer electronic material and equality. It is often considered an environmental deed to use digital material, and digital material can be easily shared. As a consequence, information can be easily accessed around the world and it will be stored to several locations.

In social media services, all of the users are practically equal. Anyone can create an account despite age, gender, or race. Every user has essentially the same possibilities at the beginning, and the users are usually graded by their deeds, not their backgrounds. Unfortunately, in many services the money can give additional value to the service, so the underprivileged to affluent do not benefit from the services as much as the others do.

B. Types of services

For media sharing services (for example, YouTube and Picasa), the social network was clearly weaker, but the feedback had much more value than in social networking services (for example, LinkedIn or Facebook). In more graphical services such as Habbo or Combat Arms EU, the character's appearance was somewhat important. In Habbo, it was a clear status symbol to invest in clothing and equipment, whereas in Combat Arms EU, the weapons and gear have more functional properties.

C. User groups

The value is also dependent on the user group, as the formula V(X,S,P) suggests. For example, LinkedIn is geared toward professional use and there are not a significant number of emotional values. Nevertheless, it seems that LinkedIn has taken several steps toward Facebook-type service properties, such as micro blogging, and hence, toward broadening the user group.

Habbo targets young children, and for these users, value creation is somewhat different. The main finding was that children are less capable of protecting themselves. This creates a need to have a higher level of safety than in other services. Foul language and teasing of other users seemed to be a problem in services targeting younger individuals. Moreover, the amount of information that children share might cause problems, since they often do not realize what the meaning of giving personal information to strangers is.

Although the users range in terms of age and computer skills, the experienced users most likely gain more from the services. They are more likely able to adjust settings to suit their preferences, and find relevant supporting services for their needs. Further, they are more likely to recognize trolls, spammers, and other misusers on the service.

D. The framework

The framework was easy to use, and sufficiently versatile to observe phenomena within the services. Nevertheless, some divisions did not produce extensive findings. It might be that the services were too socially oriented, and thus, the weight was more on the social values. For this research study, it might have been wise to break the social values into subcategories, such as status and relatedness.

V. Conclusion and discussion

In this paper, the value for user in social media services was analyzed. The research was based on Kujala and Väänänen-Vainio-Mattila's value framework. Ten services of different categories were analyzed.

To create value for user in a social media service, the service must have a significant number of users. The users are necessary, as they create content for the service. Media sharing should also be supported. With text, images, and video, the users can share their experiences and emotions. Subsequently, there must be the option for other users to provide feedback, so they can relate to these experiences and provide social connection for others. Finally, the service should have a profile or character creation. This allows the user to share his or her basic information as much as he/she wishes. Each profile or character should be unique enough to boost the sense of individuality and personalization. These profiles also help to initiate contact with other users.

In the future, the same research study could be conducted with the help of actual users. In this study, several assumptions had to be made, and it would have been beneficial to have end users to verify those assumptions. It is quite likely that the results can be generalized to all social media services, but value creation is likely different with other types of web based services. Yet, it might be interesting to explore whether the non-social values are the same. Finally, it would be interesting to determine the level of value required for successful and prosperous social media services.

References

[1] http://www.facebook.com/press/info.php?statistics, retrieved 14.4.2010

[2] http://wikipedia.org/, retrieved 14.4.2010

[3] Lehdonvirta, V.; "Virtual Consumption" [Doctoral thesis]. Publications of the Turku School of Economics (A-11:2009), Turku, 2009

[4] http://www.ft.com/cms/s/2/fd9b57ac-a879-11df-86dd-00144feabdc0.html, retrieved 14.10.2010

[5] Grönroos, C.; "Service Management and Marketing, Customer Management in Service Competition", 3rd edition, John Wiley & Sons, ISBN: 978-0-470-02862-9, 2007

[6] Cockton, G.; "Designing worth is worth designing". In Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles, NordiCHI '06, pp. 165-174, 2006

[7] O'Reilly, T.; "What is web 2.0?", 2005 http://oreilly.com/web2/archive/what-is-web-20.html, retrieved 14.9.2009

[8] http://www.merriam-webster.com/dictionary/value, retrieved 12.3.2010

[9] Pietarinen, J.; "Ihmislähtöiset luonnonarvot ja luonnon omat arvot". In: Arvot ja luonnon arvottaminen, edited by Haapala & Oksanen, Gaudeamus, Helsinki, ISBN: 9789516627949, 2000

[10] Maslow, A.; "Motivation and Personality". New York: Harper, 1954

[11] Aldefer, C.P.; "Existence, Relatedness, and Growth; Human Needs in Organizational Settings", Organizational Behavior and Human Performance, Volume 4, Issue 2, pp. 142-175, 1969

[12] Boztepe, S.; "User Value: Competing Theories and Models", International Journal of Design 1(2), pp. 55-63, 2007

[13] ISO 9421 - 11 part 11: "Guidance on the usability specification of measures", part of ISO 9241, Ergonomic requirements for office work with visual display units (VDT's), 1997

[14] Saariluoma, P.; "Käyttäjäpsykologia". Porvoo, Finland: WSOY, ISBN: 9789510291849, 2004

[15] Kujala, S. & Väänänen-Vainio-Mattila, K.; "Value of Information Systems and Products: Understanding the Users' Perspective and Values", Journal of Information Technology Theory and Application (JITTA), 9, 4, 2009.

[16] Cavazza, F.; "Social Media Landscape Redux", 10 april 2009 http://www.fredcavazza.net/2009/04/10/social-media-landscape-redux/ , retrieved 14.4.2010

[17] Endsley, M.R.; "Designing for Situation Awareness in Complex System. In: Proceedings of the Second international workshop on symbiosis of humans, artifacts and environment, Kyoto, Japan, 2001

A Microcontroller-based HF-RFID Reader Implementation for the SD-Slot

Andreas Loeffler, Andreas Deisinger
Chair of Information Technologies
Friedrich-Alexander-University of Erlangen-Nuremberg
Erlangen, Germany
Email: loeffler@like.eei.uni-erlangen.de, Andreas.Deisinger@e-technik.stud.uni-erlangen.de

***Abstract*—This work describes an RFID reader system based on an emulated file system to be used in SD-capable systems. Off-the-shelf SD-compliant HF-RFID readers, generally, use the SDIO interface for connecting to computers and PDAs. Therefore, the usage of SDIO-compatible SD card readers is essential to assure correct operation. In contrast to the adoption of SDIO, this work shows an approach to be used with every SD card reader, not necessarily requiring the SDIO specification. This leads to an HF-RFID system to be applied in computer environments (PDA, PC, etc.) where independence of operating systems and special drivers is of great importance. Adding RFID functionalities to existent systems could help to minimize the gap between real and digital world. This approach offers this particular functionality to any SD-compliant host device.**

***Keywords*-Radiofrequency identification, file systems, microcontrollers, smart cards, emulation.**

I. Introduction

The market for RFID applications is still expanding [1]. Therefore, the need for RFID readers [2], particularly mobile RFID readers, is increasing.

There are plenty of RFID readers for nearly every possible kind of interface, like Ethernet [3], SPI, serial port, Bluetooth, USB, etc. Interfaces mainly used by mobile devices are, apart from wireless interfaces, USB, SecureDigital (SD), and some proprietary interfaces. The USB interface on mobile devices is usually driven in device mode not in host mode. Moreover, there exist various USB socket implementations and particular derivatives. These are some reasons why various RFID readers are connected to (primarily mobile) computers and PDAs using the SD interface [4]. Unfortunately, common RFID reader implementations (e.g., [5], [6]) usually prefer the SD Input/Output (SDIO)-interface. Using SDIO instead of standard SD has several disadvantages. The first drawback is the need for an SDIO-compatible SD card reader. The second disadvantage is the need for various drivers to be installed using such an SDIO-capable RFID reader. As some readers only support drivers for one specific operating system, there is rather no reason in upgrading existing systems with RFID functionalities. Besides, the costs of such SDIO readers are usually higher compared to ordinary RFID readers.

The subject of this paper describes an approach to realize an SD-capable (not SDIO) HF-RFID reader to be used in every ordinary SD card reader. It is important to outline the advantages of the implementation. There are two major issues to be regarded. First, full operating system independence, and, second, no additional drivers need to be loaded as the computer's operating system will recognize the reader as an emulated SD card.

This paper will focus on the hardware part and the realization of such a system and is organized as follows. Section II gives a short description of the system. Section III shows the verification and Section IV the limitations of the current implementation of the system. A conclusion and references for future work are given in Section V.

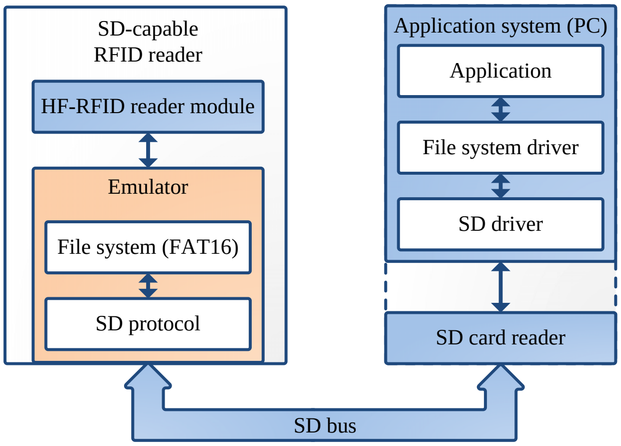


Figure 1. The block diagram of the whole system, including the SD-capable reader (left block) linked via SD bus to the application system (right block, e.g., a PC or PDA)

II. System

This section will provide an overview of the system as a whole. Figure 1 shows the concept of the SD-capable HF-RFID reader. On the left side, there is the current SD-capable RFID reader prototype connecting several blocks including one block for the emulation of the SD card protocol and one block for emulating a File Allocation Table (FAT)-16 file system [7]. The content of the file system is represented by mapped objects; these objects include links to the RFID reader's firmware itself, providing the RFID transponders' data (e.g., the Unique Identifiers (UIDs) of several transponders within the read range of the RFID reader).

The right side of Figure 1 is represented by an ordinary SD card reader, either externally connected, e.g., via USB, or an internal one. The goal of this system is to provide the SD-capable application system (i.e., a PC, PDA or smartphone) with the data of the RFID transponders read out. This is realized by creating a file within the emulation of the FAT file system, which is subsequently read out by the application system. This file includes the data of the transponders, i.e., the output of the RFID reader. This is also known as RFID uplink channel. The downlink channel, i.e., data from the reader to the tags, may also be realized using a file-based approach. By writing information into an emulated file, the SD-capable reader may notice the change and forward the commands to the underlying RFID reader hardware. Because of simplicity reasons, the system described in this work exclusively provides an RFID uplink channel, i.e., the connection from the the RFID transponders to the SD card reader or application, respectively.

A. Hardware

This subsection describes the hardware of the SD-capable RFID reader. Figure 2 shows the basic blocks the reader is built upon. Peripheral parts include a *Debugging interface*, an *RFID interface* to connect to an earlier developed HF-RFID reader module, a *Programming interface* to be able to program the microcontroller (µC) of type Atmel AT32UC3A1512 [8] using either the JTAG interface or USB, and a power supply module providing primarily the µC with 3.3 V.

The main part of the hardware is a 32 bit µC of the AT32UC3A family from Atmel. The main task of the device is the communication with the SD card reader connected over the SD bus. Figure 3 shows the prototype of the SD-capable HF-RFID reader. The microcontroller is in the center of the figure as well as the SD interface (right hand side) that may be connected to an SD card reader. The RFID interface for connecting the RFID module is shown at the bottom of the figure, whereas the JTAG interface serves as programming and debugging interface for the µC's firmware. The USB and RS-232 connections at the top are used as programming (USB) and system debugging (RS-232) sources. System debugging includes the retrieval of system status and error messages, which are generated by the firmware of the µC.

B. Firmware

The next two paragraphs give further details of the SD protocol and its realization within the reader.

1) Basics of the SD Protocol: The SD protocol is a Master-Slave protocol, which means that every step during communication is triggered by the master (SD card reader) followed by the slave's (SD-capable HF-RFID reader) response. After inserting the RFID reader into an SD card reader, the card reader tries to access the *emulated* card. The SD card reader checks the type of card inserted. This could be either an SD or a Multimedia Card (MMC) [9], a predecessor of the SD card. The communication speed at this point has a maximum clock rate of 400 kHz (from the SD card reader), which is defined by the standard. So far, only one communication line on the SD bus is used. The firmware could be implemented in such a way, that it would support both standards. However, due to less strict timing issues and less complexity the focus is on the MMC protocol, which is downward compatible to the SD standard.

Following the assignment as MMC card, some features like supply voltage, maximum speed, number of communication lines on the SD bus, capacity of the card, etc. are requested by the SD card reader. Due to several issues (see Section IV) the supply voltage is determined as 3.3 V, the maximum speed of the emulated card is set to 400 kHz, the number of communication lines is set to one and the capacity of the card is determined as 128 MBytes. These setup values are processed within the so called *Card Identification mode*

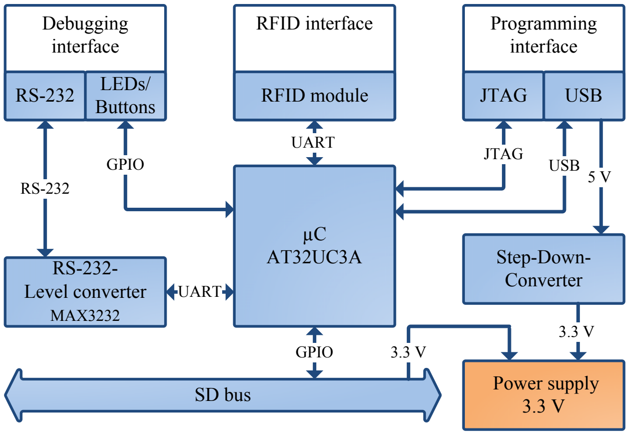


Figure 2. Block diagram of the reader's hardware

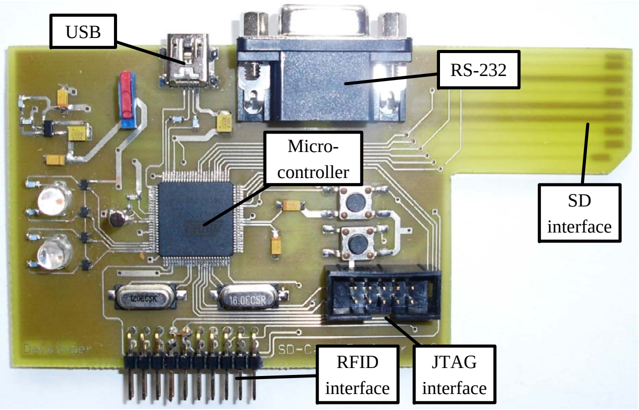


Figure 3. The prototype of the SD-capable RFID reader, shown without the external RFID reader module, which is usually connected to the RFID interface

running at low speed (max. 400 kHz). Subsequently, the card enters the *Data Transfer mode* making the card's content available, e.g., to the OS. Within this data-transfer mode, every time the SD card reader requests data information, a data block of 512 Bytes needs to be transferred from the MMC card, which is emulated by the microcontroller, to the card reader and vice versa. This project uses the FAT16 architecture to implement a particular memory structure. FAT16 is used because of its prevalence in almost all computer systems. Therefore, the microcontroller has to emulate a FAT16 formatted memory architecture containing Master Boot Record (MBR), Volume Boot Record (VBR), and the root directory. The SD card reader, finally, provides the application system, usually a PC, PDA or smartphone, with the emulated card's data, i.e., files and directories. Currently, there are four files (FAT.HEX, ROOTDIR.HEX, VBR.HEX and RFID.TXT) with the latter file containing the information (in this case the UIDs) of the RFID transponders (see Figure 6). The other three files contain the structure of the emulated FAT, VBR, and root directory.

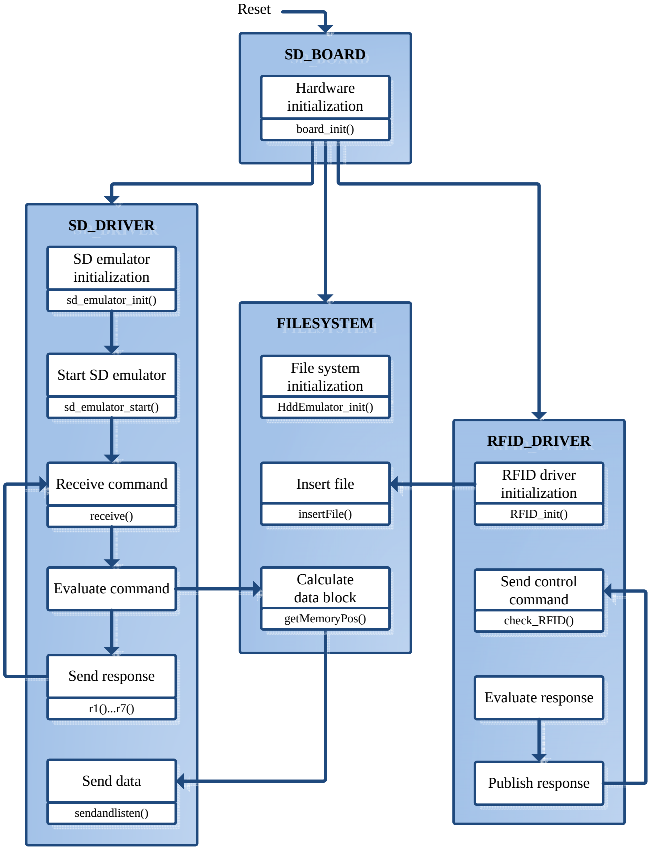


Figure 4. Overview of the SD-capable HF-RFID reader's firmware

2) Realization of the SD Protocol within the Firmware of the Reader: Figure 4 shows a rough overview of the system's firmware structure and the appropriate work flow. After starting the μC by applying power, the system initializes its hardware parts (Reset). Subsequently, three main parts are loaded, running in three different threads within a thread-like environment. In this context, *thread-like* describes an environment, in which threads are treated separately but consecutively in an quasi-OS (operating system) environment as the μC includes a single-core CPU. The *RFID_DRIVER* part (right hand side in Figure 4) has two main functionalities. The first functionality covers the control (initialization and communication) of the RFID module. This includes, e.g., the request for the available transponders, with the resulting and received UIDs, respectively. The second functionality is the preparation of the transponders' data for the adoption to the FAT file system.

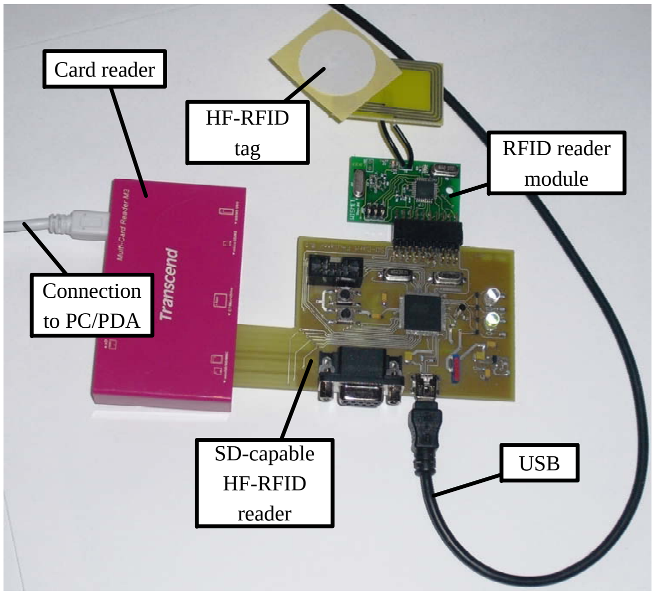


Figure 5. The system setup for verification

Taking a more detailed view at the reader's firmware shows, that there is one major problem regarding the timing issues of the SD protocol. This already above mentioned problem is further discussed in Section IV.

The part *FILESYSTEM* (centered in Figure 4) has three main functionalities. The first one involves the generation of the underlying FAT16 file system. The file system is initialized by creating its administrative datasets: MBR, VBR, and root directory. The communication with the RFID part (*RFID_DRIVER*) and the SD part (*SD_DRIVER*) describes the second functionality of the *FILESYSTEM* part. The third functionality is responsible for creating the appropriate files within the FAT16 file system. Additionally, the content of the files has to be created, too, including the transponders' data (see Figure 6).

The *SD_DRIVER* part (left hand side in Figure 4) has to deal with the SD respectively MMC protocol. This means, all timing issues regarding the SD bus have to be handled by this part of the firmware. The part is μC hardware dependent, and therefore the system's performance is somehow limited at this point. The addressed hardware dependence is directly related to the hardware components, particularly the µC with its rather limited capabilities to cope with fast incoming and outgoing signals. This issue is further discussed in Section IV. However, the realization of this part of the firmware is mainly done by polling several internal flags of the µC. The reason for not choosing interrupts is caused by the overwhelming amount of stack an interrupt-driven approach would need. The main drawback would not be the memory usage but the loss of time needed to return from the sub-functions (e.g., interrupt routines). The main external signal, and therefore the triggered internal flag of the µC, is the SD clock signal generated by the SD card reader. By choosing an interrupt-driven approach, the time for calling the interrupt function and returning from it (to the main function) would not fulfill the needs for an appropriate operation of the system.

By briefly summarizing the firmware architecture, one can say that the firmware controls the external RFID module to get the transponders' data (e.g., the UIDs). This data is evaluated and packed into the emulated FAT16 structure of the system. A connected SD card reader may read the FAT16 structure and forward the data to the overlying operating system, which can display the appropriate data (UID) within a file.

III. Verification

This section describes the verification of the SD-RFID combination. The setup to verify the system is given as in Figure 5. The figure shows the SD card reader, which is connected to a computer via USB (left side of the figure), whereas the SD-capable HF-RFID reader is located in the center. The RFID reader module (with antenna and RFID tag) is connected at the upper side of the reader. A USB debugging connection is shown at the bottom of the picture. The system is connected to the SD card reader (Type: Transcend Multi-Card Reader M3).

To prove the correct operation of the system, the SD card reader and the system were connected to different operating systems, including Windows XP, Windows 7, Linux (Ubuntu) and Mac OS X (Snow Leopard). The results are shown in Figure 6. The screenshot on the left was made in Windows 7 showing additionally the four different files, including the *RFID.txt* containing the UID of the transponder. The UID with the value of '00 00 00 00' stands for an invalid data communication between transponder and reader. The current UID of the transponder used is therefore '3C 50 8B 2A'. The Linux screenshot is located on the bottom of the picture and shows the capacity of the card; in this case it is 128 MByte. The background of the picture shows a Mac OS X screenshot containing the content of the emulated FAT16 system (top left), the content of the *RFID.txt* file (center right), and the information of the SD-capable HF-RFID reader, called *EMULATOR* (right side of the figure).

IV. Performance Issues

The limitations appearing during implementation are mostly generated by various SD card reader implementations of the SD protocol standard. A huge drawback is the minimum clock rate, various card readers manage. The maximum frequency for initializing the SD card (400 kHz) is managed and adhered by every card reader.

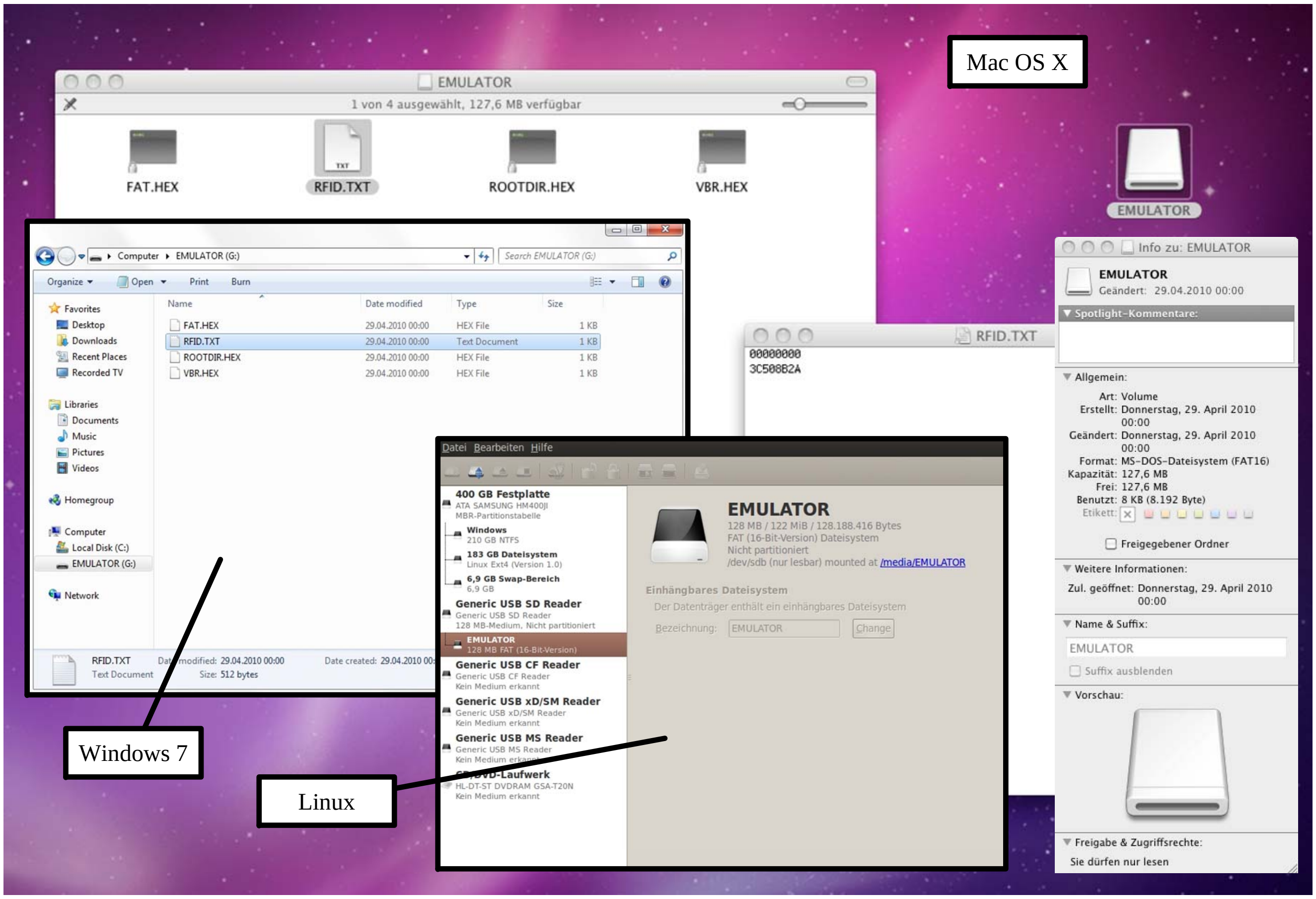
Although the SD card (or MMC card) can return its maximum allowed clock rate to the SD card reader to prevent the reader to read out data too fast, some SD card readers ignore that property and start off with a frequency far too high for the µC to cope with. Other card readers just reset and search for another card. These problems create some kind of drawback for this particular system.

The limitations itself are governed by the maximum clock rate the microcontroller is able to handle. Internal calculations showed that the maximum clock rate is about 1 MHz using the µC at the maximum clock rate of 66 MHz. These timing issues are the reason why a flag-driven polling approach is used instead of choosing an interrupt-driven approach. It can be shown that polling the flags (which are also used by the interrupts) provides a higher data throughput than an interrupt-driven approach. The bottleneck, regarding the timing problem, occurred at the point where SD data blocks of 512 Byte have to be transferred from the card reader to the system and vice versa, as some processes have to work in parallel, which, of course takes time, if only a single-core CPU is available.

V. Conclusion and Future Work

This paper presented an approach to overcome existing obstacles with RFID reader architectures using the SD interface. While standard SD reader approaches generally use the SDIO interface, the work presented in this paper shows a method to use the standard SD interface to work with every possible kind of SD card reader. The system is based on a microcontroller fully emulating a FAT16 file system to be able to transfer RFID-based data, e.g., the UID of a transponder, to a superior system, e.g., a PC, PDA or smartphone, using a standard SD card reader interface. Therefore, the system not only controls a connected RFID reader module, but also the SD or MMC card protocol to communicate with the SD card reader. Successful tests with various operating systems have been carried out (see Figure 6) to prove the principle of this structure.

During the work different kinds of limitations were located. Future implementations should account for these drawbacks by implementing time robust structures. For instance, one option would inherit FPGA structures for the direct interface to the SD card reader. Also, other options



Name	Date modified	Type	Size
FAT.HEX	29.04.2010 00:00	HEX File	1 KB
RFID.TXT	29.04.2010 00:00	Text Document	1 KB
ROOTDIR.HEX	29.04.2010 00:00	HEX File	1 KB
VBR.HEX	29.04.2010 00:00	HEX File	1 KB

RFID.TXT Date modified: 29.04.2010 00:00 Date created: 29.04.2010 00:
Text Document Size: 512 bytes

RFID.TXT

```
00000000
3C508B2A
```

400 GB Festplatte
ATA SAMSUNG HM400JI
MBR-Partitionstabelle

Windows
210 GB NTFS

183 GB Dateisystem
Linux Ext4 (Version 1.0)

6,9 GB Swap-Bereich
6,9 GB

Generic USB SD Reader
Generic USB SD Reader
128 MB-Medium, Nicht partitioniert

EMULATOR
128 MB FAT (16-Bit-Version)

Generic USB CF Reader
Generic USB CF Reader
Kein Medium erkannt

Generic USB xD/SM Reader
Generic USB xD/SM Reader
Kein Medium erkannt

Generic USB MS Reader
Generic USB MS Reader
Kein Medium erkannt

CD/DVD-Laufwerk
HL-DT-ST DVDRAM GSA-T20N
Kein Medium erkannt

EMULATOR
128 MB / 122 MiB / 128.188.416 Bytes
FAT (16-Bit-Version) Dateisystem
Nicht partitioniert
/dev/sdb (nur lesbar) mounted at /media/EMULATOR

Einhängbares Dateisystem
Der Datenträger enthält ein einhängbares Dateisystem
Bezeichnung: EMULATOR Change

EMULATOR
Geändert: 29.04.2010 00:00
▼ Spotlight-Kommentare:
▼ Allgemein:
Art: Volume
Erstellt: Donnerstag, 29. April 2010 00:00
Geändert: Donnerstag, 29. April 2010 00:00
Format: MS-DOS-Dateisystem (FAT16)
Kapazität: 127,6 MB
Frei: 127,6 MB
Benutzt: 8 KB (8.192 Byte)
Etikett: ☒
□ Freigegebener Ordner
▼ Weitere Informationen:
Zul. geöffnet: Donnerstag, 29. April 2010 00:00
▼ Name & Suffix:
EMULATOR
□ Suffix ausblenden
▼ Vorschau:
▼ Freigabe & Zugriffsrechte:
Sie dürfen nur lesen

Figure 6. Verification of the system with different operating systems: Linux (Ubuntu 9.04, [10]), Windows 7 [11] and Mac OS X (Snow Leopard, [12])

like CPLDs (Complex Programmable Logic Device) would come to the fore.

Acknowledgment

The authors would like to thank the Fraunhofer Institute for Integrated Circuits (IIS) in Erlangen for its support on this project. Additionally, we would like to thank Heinrich Dietsch for useful discussions and the anonymous reviewers for providing helpful comments.

References

[1] trading-house.net AG, “Research - RFID Market to Reach $5.35 Billion This Year, Says ABI Research,” *www.ad-hoc-news.de*, Mar 2010. [Online]. Available: http://www.ad-hoc-news.de/research-rfid-market-to-reach-5-35-billion-this--/de/Unternehmensnachrichten/21106631

[2] K. Finkenzeller, *RFID Handbook: Fundamentals and Applications in Contactless Smart Cards, Radio Frequency Identification and Near-Field Communication*. Wiley, 2010.

[3] A. Loeffler, U. Wissendheit, H. Gerhaeuser, M. Hoffmann, A. M. Zadeh, and D. Kuznetsova, “A SOAP capable HF-RFID-Reader,” in *RFID SysTech 2008 , 4th European Workshop on RFID Systems and Technologies*, 2008.

[4] SD Association, “http://www.sdcard.org/home/,” Feb 2010. [Online]. Available: {http://www.sdcard.org/home/}

[5] RFReader Corporation, “http://www.rfreader.com/,” Mar 2010. [Online]. Available: {http://www.rfreader.com/}

[6] Sirit Inc., “http://www.sirit.com/,” Mar 2010. [Online]. Available: {http://www.sirit.com/}

[7] J. Axelson, *USB mass storage: designing and programming devices and embedded hosts*, ser. E-libro. Lakeview Research LLC, 2006.

[8] Atmel Corporation, “http://www.atmel.com/,” Jan 2010. [Online]. Available: {http://www.atmel.com/}

[9] MultiMediaCard Association, “http://www.mmca.org/,” Feb 2010. [Online]. Available: {http://www.mmca.org/}

[10] R. Petersen, *Ubuntu 9.04 Desktop Handbook*. Surfing Turtle Press, 2009.

[11] J. Boyce, *Windows 7 Bible*, ser. Bible Series. John Wiley & Sons, 2009.

[12] D. Pogue, *Mac OS X Snow Leopard: The Missing Manual*, ser. Missing manual. O’Reilly Media, 2009.

LiquidKeyboard: An Ergonomic, Adaptive QWERTY Keyboard for Touchscreens and Surfaces

Christian Sax
School of computing and communications
University of Technology Sydney
UTS
Sydney, Australia
sax@it.uts.edu.au

Hannes Lau
Electrical Engineering and Computer Sciences
Technische Universität Berlin
TU Berlin
Berlin, Germany
hlau@cs.tu-berlin.de

Elaine Lawrence
School of computing and communications
University of Technology Sydney
UTS
Sydney, Australia
elaine@it.uts.edu.au

Abstract—**Virtual touchscreen keyboards provide poor text input performance in comparison to physical keyboards, a fact, which can partly be attributed to the weaker tactile feedback, they offer. Users are unable to feel the keys on which their fingers are resting, and usually hover their hands over the keyboard, hitting each key individually. Consequently, users cannot use all ten fingers for typing, which decreases their input speed and causes hand fatigue. We present a keyboard prototype called LiquidKeyboard, which adapts to the user's natural finger positions on a touchscreen and discuss options to allow for the fingers to rest on the screen while typing. When invoked, the keyboard appears directly under the user's fingertips and is able to follow finger movements. As the positions of the surrounding keys are fixed in relation to each finger, users can find and touch the keys without tactile feedback. Additionally the user controlled positioning of the keys allows the keyboard to adapt to the physical specification of the user's hand, such as the size of each finger.**

Keywords-Keyboard; touchscreen; touch surface; adaptive; touch-type; QWERTY.

I. INTRODUCTION

The way in which users interact with computers and electronic devices has changed in recent years as both users and electronic devices become more mobile and wireless connectivity becomes the norm. Ubiquitous computer systems are now a part of everyday life in business, education and social settings. Driven by the increased need for user friendly input interaction devices and powered by the availability of new hardware technology, the input methods used on these devices have deviated from standard-sized physical, tangible keyboards to methods more suitable for the specific physical limitations of mobile devices. Popular and very successful touch-screen phones such as the Apple's iPhone [3] or the iPad tablet computer [4] set the benchmarks in this development, creating a massive shift from physical key or stylus based text input methods to virtual keys (called softkeys) displayed on a touchscreen.

Touchscreen systems employ displays with smooth surfaces that are capable of sensing finger touches, which can then be related to interface elements displayed on the screen directly underneath them. This technology enables direct user interaction with the display, without the need for additional physical keys. For example, a button can be activated when the user touches an image of it on the screen. It can therefore be placed on the interface surfaces without the need to increase the actual device size. As a result, product designers can reduce the number of physical keys to the bare minimum (such as an 'on' and 'off' switch) while, at the same time, increase the screen real estate. Users interact with the interfaces more directly by pushing buttons on the screen, moving objects around, and performing gestures to trigger functions.

On touchscreens and touch surfaces users cannot feel controls even though they can manipulate them by touch. Although the same visual language, such as a button, is used on the user interface one cannot feel the button or its edges. Mobile user interfaces try to compensate the lack of tactile feedback by relying heavily on visual and auditory cues. There are some approaches to make these screen interactions more tangible, such as a screen that changes the surface structure, but these are not yet available as mainstream consumer devices [1, 17, 21].

A. *Existing touchscreen text input methods*

Textual input with human computer interfaces is measured in two scales: speed and accuracy [15]. Table 1 compares the text input speed in WPM (words per minute) of selected text input systems. These systems can be divided into two groups: firstly, the selective method where users have to select each letter individually and secondly, the predictive method that aim to predict the intended word by analyzing every user input [9]. The mechanical QWERTY keyboard, a selective method system and the most common English keyboard layout, serves as a baseline to compare the text input speeds. WPM figures show the average speed, however, they can vary greatly depending on the user and serve as a guide only.

In comparison to physical keyboards the average input rate on touchscreen keyboards is low (see Table 1). A typical word rate, even with the use of input prediction, is around 15-30 words/min according to [13]. This mainly derives from the fact that devices that are fitted with touchscreens are generally smaller and do not provide sufficient screen space to allow the user to use both hands for typing. But even keyboards on larger devices such the iPad cannot be

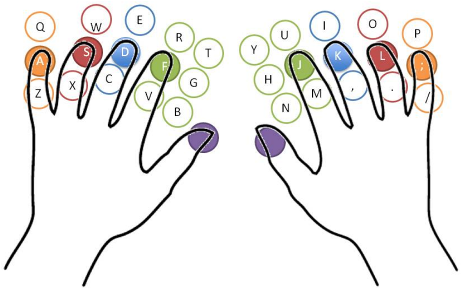


Figure 1 – Showing the keys under the finger contacts

used easily with 10 fingers, as users cannot feel if a finger is on a particular key or not.

On an English QWERTY keyboard layout the fingers are placed on the A-S-D-F and J-K-L-; keys for the left and right hand fingers respectively – these keys are called home keys. Both thumbs rest on the space key. Proficient touch-type writers know where other keys are in relation to the home keys and do not need to look at the keyboard while typing. Yet, on a touchscreen, the tactile guide to keep the fingers on their respective home keys is missing. Without tactile feedback users cannot relate to the key's location and the spatial position of surrounding keys. As a consequence they have to rely on visual orientation, look at the keys to hit the right ones and cannot keep their gaze on the actual task. This increases the eye movements (called saccades), meaning that users cannot perform any action or notice changes on the interface [11] – a lower text input speed and a higher error rate result.

In summary, current touchscreen devices, such as the iPad, do not provide tactile feedback resulting in users needing visual cues to compensate for this lack of tactile feedback. This results in poor typing speeds of around 15-30 WPM. We introduce an alternative to the current available touchscreen keyboards, with which we aim to leverage the ten finger touch-typing and compensate the missing tactile feedback with our adaptive keyboard approach.

After a description of the LiquidKeyboard's concept in Section 2, this paper shows two prototypical implementations, which employ different algorithmic approaches to project the keys under the users fingertips. Section 4 discusses problems that arise when the LiquidKeyboard is implemented on devices that are unable to sense the pressure of a user's touch and presents different approaches to overcome them. Finally, Section 5 discusses plans for future research and implementations followed by the conclusion in Section 6.

II. THE LIQUID KEYBOARD

The design of our new keyboard system, called *LiquidKeyboard*, aims to empower users to utilize all ten fingers on the touchscreen as on a normal physical keyboard. The home keys are displayed under each finger and follow the individual finger position on the screen. Users are free to place their fingers anywhere and do not have to adapt to the straight key rows found on most keyboards. Adjacent keys to the home keys are situated next to them and form groups around fingers (see Figure 1 and Table 2). For instance, the right hand middle finger is on the 'K' home key which forms a group with the 'I' and ',' keys being above and below as on a physical QWERTY keyboard. Groups follow the finger touches on the surface keeping the distance between the home key and the adjacent keys constant. Consequently, the surrounding keys will always be at the same position in relation to the users' finger, no matter where they are on the surface. Users will not have worry about hitting the wrong keys when they shift their fingers.

TABLE I. COMPARISON OF DIFFERENT TEXT INPUT METHODS

Input method	PM	Advantages	Disadvantages	WPM	Applications
Mechanical. QWERTY keyboard	No	-Fast due to physical keys and tactile feedback -All 10 fingers can be used for text input	-Too many keys to be fully supported on mobile systems -Requires physical keys for tactile feedback	64.8 [19]	Desktop computers, phones
Virtual pseudo QWERT keyboard	Yes	-Keyboard layout similar to mechanical QWERTY	-Most of the time very small hence not all keys are supported	18.5 [14]	Tablet computers, handhelds, touchscreen phones
Gesture on pseudo QWERTY virtual keyboard	Yes	-Keys do not need to be hint individually	-User have to learn input method	25 [12]	Touch phones, tablet computers, handheld
Un-constrained handwriting word recognition	No	-Natural handwriting is used as the input method-No vocabulary has to be learned	-Recognition rate can be poor -A stylus is required for most systems	24.1 [11]	Handheld, phones, tablet computers
Single handwritten character recognition	No	-Natural handwriting is used as the input method	-Only one character at a time can be recognised. -Some system require special alphabet	21 [6]	Handheld, Phones, tablet computers

ISBN: 978-1-61208-003-1

The LiquidKeyboard is designed for touch sensitive surface systems where the screen is big enough to accommodate two hands, such as the iPad. On these systems users can write with both hands while using the 10 finger touch-typing.

Microsoft has patented an input concept where they split a QWERTY keyboard in two halves [16]; one for the left and one for the right hand using the touch of the users' palms on the screen as additional orientation cues. The LiquidKeyboard moves beyond the Microsoft concept having three major advantages. Firstly, the keys follow the individual finger touches movements, secondly palm touches are not required as keyboard orientation reference points (see also Prototype 2 with rotation of key groups) and lastly the keyboard layout adapts to each individual finger/hand shape.

Another LiquidKeyboard benefit is that the interface is easy to learn due to its consistent keyboard layout. As the keyboard can be invoked at any place on the surface, the keyboard adapts to the users' hand physiology such as the hand size and finger position. In forthcoming user tests that compare the LiquidKeyboard to traditional softkey keyboards, we expect to see a decrease in hand fatigue as users can rest their fingers on the screen while typing and do not need to hold them in a hovering position.

We believe our system could improve the usability of mobile emergency system such as the 'Portable Medical Monitoring Computer' [20]. The LiquidKeyboards adaptive capabilities can provide a low cost and effective text input system for different types of touchscreen and touch surface systems.

III. Implementation of the Prototypes

The LiquidKeyboard was implemented in two prototypes to allow empirical testing and to prepare for future project phases. To be able to test on differing platforms we used web technologies, namely HTML, CSS, and JavaScript to create web applications that can run in every Gecko or WebKit based web browser. WebKit specific JavaScript API extensions allow us to harness the multi-touch capability of Apple's iPhone and iPad and react to the touch of multiple fingers. In this first phase of the project both prototypes support only the right hand-side of a QWERTY keyboard, i.e. the home keys are 'J-K-L-;'. The following sections will explain the implementation of the softkey activation and the two different prototypes.

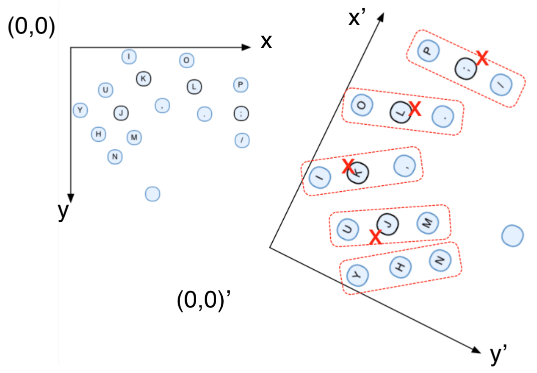


Figure 2 – Original (left) and adapted (right, primed) keyboard layout of the first prototype. The users' initial touches are marked with crosses. The dashed lines depict the key groups that will be moved in unison.

TABLE II. Home keys and their grouping. Bold letters indicate the home keys

Home key	Group
A	Q **A** Z
S	W **S** X
D	E **D** C
F	R T **F** G V B
J	Y U H **J** N M
K	I **K** ,
L	O **L** .
;	P **;** /
SPACE	**SPACE**

A. Key activation

In both prototypes, the keys are defined as points on the plane without spatial extent. After initialization, each touch on the surface is associated to the closest key by a simple nearest neighbour search algorithm [11]. As long as the user's finger remains on the screen the association is maintained and the key is considered pressed. As a consequence users do not have to hit the keys exactly to activate them.

B. Prototype 1 with a keyboard layout transformation

As soon as four touches are sensed on the screen we use these touch positions to place the entire keyboard. To do so, the first prototype has a basic keyboard layout stored in the application, which specifies a position for each key including the home keys. To find an adapted keyboard layout we map the points where the user has touched the screen to the stored positions of the home keys. Then we determine a rotation angle, scale factor and translation vector for a two dimensional Helmert transformation of the stored layout that would bring the home keys from their original positions as close as possible to the positions the user touched (see our current prototype implementations). As the equations to determine the transformation parameters are over determined by the four reference points we use a least-square adjustment as shown in [7] and remain with a rest deviation between the desired coordinates for the home keys and their real positions as images of the transformation applied to the stored layout. The correct association of the four chosen positions to the appropriate home keys is initially unknown, which leads us to consider all possible mappings and chose the one with the lowest remaining deviation.

In a second step, we move each key group identified in Table 2 until the respective home key fits exactly under the user's touch, effectively clearing any deviation from the previous step. This second step, the translation of a key group is also performed each time the user moves a finger resting on a home key. The first step, however, is only used to determine the initial position of the keyboard, once the user touches the screen.

C. Prototype 2 with rotation of key groups

Experiments with the first prototype showed that the keys do not only have to follow the individual finger movements but that the key groups will also need to adapt their orientation based on the hand's position. As the calculation of the transformation parameters proved to be too slow to be carried out every time the user moved a home key we implemented a second prototype with a simpler geometric model, which allows for rotation of the key groups while the user is moving his or her hand on the display.

In this second model, once five touches have been registered (including the thumb), we approximate the user's palm position by calculating a circle with an outline that comes as close as possible to all five touch points in a least-square sense [23]. This allows us to approximate the position and orientation of the user's hand and map the touch positions to the home keys. Ordered on a clockwise circle, the first key after the biggest angular gap is associated with the user's thumb and therefore with the space key while the second touch is mapped to the index finger and 'J' key. All other home keys follow in a clockwise order.

The best-fit circle is used to determine the mapping of the user's fingers the home keys and discarded thereafter. As depicted in Figure 3 we use the apex of an isosceles triangle based on the index and little finger positions to estimate the position of the user's wrist. The angles and side ratios of the triangle are constants and calculated a priori based on the average length (finger tips to wrist = d_1) and breadth (index to little finger = d_2) of the human hand (see (1)) [22].

This method proved to approximate the position of the user's wrist closely enough to align the keys of each group on a ray originating at the wrist position and passing through the group's home key (see also Figure 3). With a measured distance d_1 and the constant c the second LiquidKeyboard prototype is able to update the wrist position and rotate the key groups fast enough to parallel the user's hand movements.

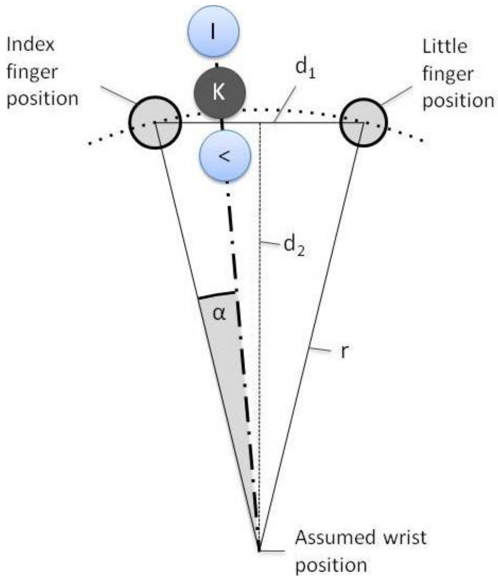


Figure 3 – Shows the rotation of the K home key group with a given index and little finger position

IV. Discussion

A. The keyboard layout

Experiments with 4 different users have shown that both prototypes have a good layout adaption of home keys to the sensed finger touch positions on the screen. The key activation is done by the nearest neighbour search algorithm rather than by sensing touch events within a defining geometric area (such as a rectangle or circle) representing a key. This solution helps with keyboard layouts where keys were spread out because keys are still activated if the sensed touch is close enough to the key but would be out of range of the visual indicated key.

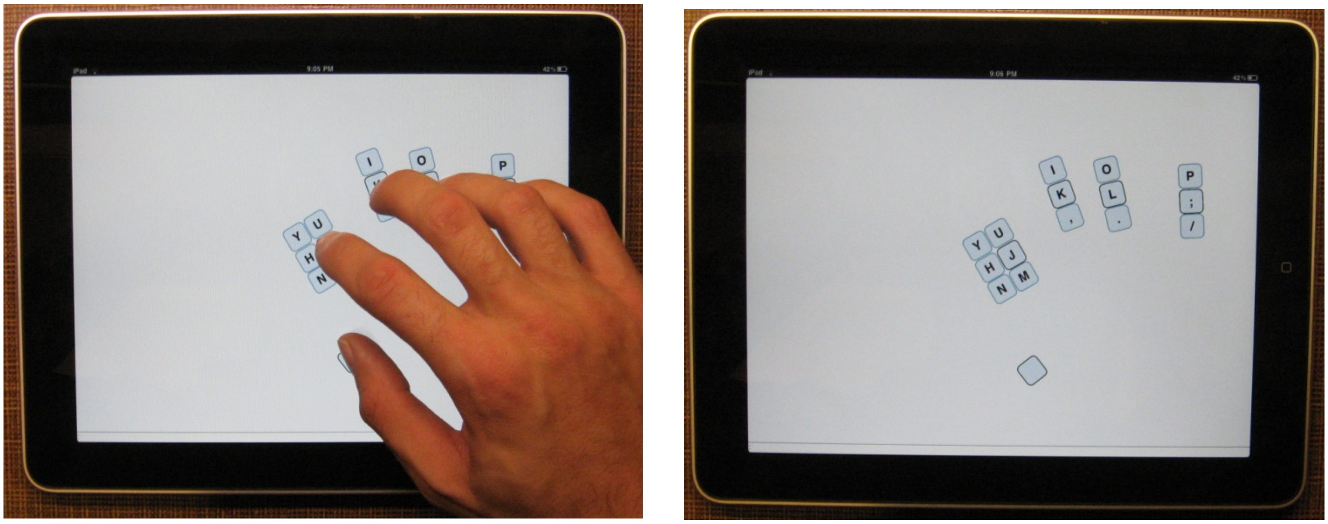


Figure 4 – The left photo shows the second LiquidKeyboard prototype on an Apple iPad. The right picture shows how the keyboard adapts to the shape of the user's hand.

$$c = \frac{d_1}{d_2} \tag{1}$$

$$c \approx 0.47 \tag{2}$$

The findings of the first prototype, which adapted only to the fingers' positions but not to different wrist positions, were taken forward into the next prototype generation. By adding an algorithm that rotates the key groups according to sensed index and little finger position a more user friendly and ergonomic keyboard layout was reached. The assumptions made on the wrist position in the described equation and geometric setup (see (2)) rotate the key groups in an adequate angle that match the actual human anatomy in our trials. Having the adjacent keys pointing towards the user's wrist resulted in an ergonomic type feeling with keys being at the right spot.

The detection of the index and little finger touch necessary for the rotation and the realignment computation of the key groups had no negative impact on the keyboard's responsiveness.

B. *Activation of the home keys*

Experiments with both prototypes have shown that the activation of the home keys is problematic. Fingers are resting on the display, which allows the algorithm to sense the touch positions and adapt the keyboard layout accordingly. This poses two usability challenges for the activation of the home keys: First, while other keys will only be touched with intent to activate them, the home keys will also sense touches of fingers returning to the home position after activating a key in the same key group. These touches however are not meant to activate the home keys. Secondly, it is desirable for the user to be able to activate any home key while his or her finger is resting on it. On a physical keyboard this is possible by changing the finger pressure and depressing the key. However, most current touchscreen systems are unable to determine the finger pressure of a touch.

In a naïve implementation where any touch triggers activation and any activation requires a touch, users would have to lift a finger and place it back in order to activate a home key on which they previously rested their finger. Thus the input recognized by the LiquidKeyboard would be incomplete; instead of the intended word 'kilogram' the input 'iogrm' would be read. Furthermore, if users return their finger to a home key after activating one of the adjacent keys, the home key would unintentionally be activated as well, adding additional letters to the recognized word. With both effects considered a naïve implementation would read the word 'kilogram' as 'i**k**o**l**gr**f**m**j**' (additional letters in bold).

Obviously the output of such naïve implementation would be far from the desired result. In the following sections we propose and discuss multiple solutions for the LiquidKeyboard, which overcome these usability challenges and are close to the well-known input paradigm of a physical keyboard hardware solution.

With touchscreens that are capable of sensing pressure users could increase their finger pressure in order to activate a home key on which their finger was already resting. In the keyboard would sense that the pressure is not high enough to activate the key. There are pressure sensitive touchscreen patents by Apple [2] and technologies by Peratech [18], which have the potential to solve this problem with a hardware solution. However, for touchscreen devices without the capability to measure the finger pressure employed by the user touching the screen, software solutions can be applied.

1) *Sensing pressure with increasing touch surface area.*

In our current prototype, implementations a sensed touch is recognized as two coordinates describing the position of a single point on the surface. The information passed to the keyboard is therefore independent from the actual touch area on the touchscreen, i.e. no matter how big or small the finger is the result will be a single point. The touch area of a finger on a screen increases when the finger is pressed harder against the surface. Hence the resting finger and a finger that is actively pressing against the screen is different; the latter will have larger touch area. We intend to leverage this effect for our home keys to sense whether users are resting their fingers on the screen or are activating one of them.

2) *Alternative keyboard layout solution.*

On devices that are unable to sense the pressure or covered area of a users' touch the keyboard layout could be modified. By moving the home keys in front of the users' fingertips the user would be able to activate them just like any other key by moving his or her fingers to the keys position and touching it. After doing so the user could return his finger to the previous home position without unintentionally activating a key there.

3) *Dictionary solution.*

Alternatively predictive text algorithms can be used to associate the input string recognized by the keyboard. For instance the input "ikolgrfmj" would be mapped to the English word "kilogram". If the mapping is ambiguous and multiple words exist whose input would be recognized as "ikolgrfmj" the input context could be used to determine the word the user intended to type.

We believe that the best user experience will be achieved by sensing the touch pressure, as this more closely resembles the user interaction paradigm closest to a normal mechanical keyboard. However, until the required hardware becomes available, sensing the increase of the finger touch area with increased pressure or the use of a specialized predictive text algorithm seem to be the best solutions with the current technology at hand.

V. FUTURE IMPLEMENTATION

The keyboard should not only adapt to the users' finger and hand positions but also to frequently miss-hit keys. In our current prototypes each touch to the surface is associated with the closest key, thus the user does not need to hit the displayed keys exactly. In order to make the keyboard more adaptive, the key found to be closest to the users touch can be moved partway towards the touch location, making future attempts to find the key at the same location more likely to succeed. With this mechanism the key layout will adapt to the users' writing style as in a similar implementation by [8].

We will extend the current prototypes to provide a complete QWERTY layout and support input for both hands as a basis for forthcoming user testing. This will allow us to compare the performance of the LiquidKeyboard to existing input methods with regard to their input speed and accuracy. Our third prototype will measure the effect of the solutions that we presented for the home key touch differentiation problem on the overall user experience and to identify further possible usability issues related to the proposed interaction methods.

VI. Conclusion

The LiquidKeyboard leverages a widespread and commonly used text input concept (the QWERTY keyboard layout) and makes it usable on touchscreens and touch surface interfaces. Our approach compensates for the lack of tactile feedback on smooth touch sensitive surfaces by making the keyboard adapt to the finger position and by following finger touch points. Users can rest their finger on the screen as on a mechanical keyboard and do not have to worry about the keys positions as they follow the user's fingers.

In the discussion we identified the home key activation problem as a challenge for our touchscreen keyboard and addressed it with solutions in soft- and hardware. Investigating these ideas and testing and comparing them in greater detail will be a field for further research.

References

[1] Apple Computer Inc., *Keystroke tactility arrangement on a smooth touch surface*, United States Patent application number: 20070247429, Accessed on 24/12/2010 from http://www.faqs.org/patents/app/20090315830

[2] Apple Computer Inc., *Force Imaging Input Device and System*, Patent number: 20070229464, Accessed on 24/12/2010 from http://www.patentstorm.us/applications/20070229464/claims.html

[3] Apple Computer Inc., *iPhone*, 2010, Accessed on 10/6/2010 from http://www.apple.com/iphone/

[4] Apple Computer Inc., *iPad*, 2010, Accessed 10/6/2010 from http://www.apple.com/ipad/

[5] E. Clarkson, J. Clawson, K. Lyons, and T. Starner, An empirical study of typing rates on mini-QWERTY keyboards. In *Extended Abstracts on Human Factors in Computing Systems* (CHI '05, Portland, OR, USA, April 02-07 2005), ACM, New York, NY, pp.1288-1291, Accessed on 24/12/2010 from doi: 10.1145/1056808.1056898

[6] M. D. Fleetwood, M. D. Byrne, P. Centgraf, K. Q. Dudziak, B. Lin, and D. Mogilev, An Evaluation of Text-Entry in Palm OS - Graffiti and the Virtual Keyboard. *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting 2002*, pp. 597-601.

[7] C. D. Ghilani. *Adjustment Computations: Spatial Data Analysis*, Wiley, 5th Edition, March 2010, Accessed on 24/12/2010 from doi: 10.1002/9780470586266

[8] J. Himberg, J. Häkkilä, P. Kangas, and J. Mäntyjärvi. On-line personalization of a touch screen based keyboard. In *Proceedings of the 8th international Conference on intelligent User interfaces* (IUI '03, Miami, Florida, USA, January 12-15 2003). ACM, New York, NY, pp.77-84, Accessed on 24/12/2010 from doi: 10.1145/604045.604061

[9] P. Isokoski. *Manual text input: experiments, models, and systems*. Doctoral dissertation, University of Tampere, Finland, 2004. Electronic dissertation, Accessed on 7/12/2010 from http://acta.uta.fi/pdf/951-44-5959-8.pdf

[10] R. J. Jacob, Eye tracking in advanced interface design. In *Virtual Environments and Advanced interface Design*, W. Barfield and T. A. Furness, Eds. Oxford University Press, New York, NY, 1995, pp.258-288.

[11] D. E. Knuth, *The art of computer programming*, Vol. 1 (3rd ed.): fundamental algorithms, Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, 1997.

[12] P. Kristensson and L. C. Denby. Text entry performance of state of the art unconstrained handwriting recognition: a longitudinal user study. In *Proceedings of the 27th international Conference on Human Factors in Computing Systems* (CHI '09, Boston, MA, USA, April 04-09 2009). ACM, New York, NY, pp.567-570, Accessed on 24/12/2010 from doi: 10.1145/1518701.1518788

[13] O Kristensson, Five Challenges for Intelligent Text Entry Methods, *AI Magazine*, Vol. 30, No. 4, 2009. Accessed on 10/6/2010 from http://www.aaai.org/ojs/index.php/aimagazine/article/view/2269/0

[14] M. H. Lopez, and I. S. MacKenzie, and S. Castelluci, *Text Entry with the Apple iPhone and the Nintendo Wii*, 2009. Accessed on 10/9/2010 from www.malchevic.com/papers/iphone_paper.pdf

[15] I. S. MacKenzie and R. W. Soukreff, Text Entry for Mobile Computing- Models and Methods, Theory and Practice. In *Human-Computer Interaction*, Vol. 17, Issue 2&3 September 2002, pp.147-198.

[16] Microsoft Corporation, *Virtual keyboard based activation and dismissal*, Patent number US20090237361, Filed 24/9/2009, Accessed on 10/6/2010 from http://www.patents.com/virtual-keyboard-based-activation-and-dismissal-20090237361.html

[17] Nokia, *Tactile Touch Screen*, Patent number PCT/ET2006/009377, Filed 27/9/2006, Accessed on 10/6/2010 from http://www.unwiredview.com/wp-content/uploads/2008/07/nokia-tactiles.pdf

[18] Peratech Ltd., *Paratech homepage*, 2010, Accessed on 10/6/2010 from http://www.peratech.com/index.php

[19] H. Roeber, J. Bacus, and C. Tomasi, Typing in Thin Air: The Canesta Projection Keyboard - a new Method of Interaction with Electronic Devices. In *CHI '03 Extended Abstracts*, pp. 712-713.

[20] C. Sax and E. Lawrence, Tangible Information: Gestures for a Portable e-Nursing touch screen interface. In *Procceedings of the Healthcom 2009 11th International Conference* (16-18 December 2009), pp.1-8, Accessed on 24/12/2010 from doi: 10.1109/HEALTH.2009.5406211

[21] Staska, *Nokia haptikos tactile touchscreen details emerge,* 2008, Accessed on 10/6/2010 from http://www.unwiredview.com/2008/07/08/nokia-haptikos-tactile-touchscreen-details-emerge/

[22] A. R. Tilley and Henry Dreyfuss Associates, *The Measure of Man and Woman: Human Factors in Design*, John Wiley & Sons, 1993.

[23] D. Umbach and K. N. Jones. A few methods for fitting circles to data. *IEEE Transactions on Instrumentation and Measurement, Vol. 52, Issue* 6, 2003. pp.1881-1885.

Effect of Internet on Arab Societies

Ali A. Sakr
Kafrel-Sheikh University, Egypt
ali_asakr@yahoo.com

***Abstract* - The Internet affects the behavior of the young generations; it causes a mixture of virtual publicity, amentia and idiopathy. Internet addicts suffer some radiation defects due to their long sessions, this may result in blood ionization, leukemia or eye cataract. The paper presents a statistical study on health engineering concerning the effects of Internet on the Arab Society. It presents also some guidelines for a better Internet use.**

Keywords – IPS; ISP; HDI.

I. INTRODUCTION

Internet handles data among hosts in different sites. Information transfer via Internet results in technology evolution and economy developments. Internet handles e-commerce, educational sites, e-mail, chat, data transfer, etc.,; it can handle twitters, virus and spam. The use of Internet doubles annually. Many statistics were developed to evaluate the usability of Internet. The next sections explore statistical studies about Internet users, the harms and benefits of Internet, wireless Internet and their harms, and explore a regional study for Internet users in Egypt.

II. STATISTICAL STUDIES ABOUT INTERNET USERS

In USA, by 2005 [1], it was declared that the yearly access to Internet was about 10.8 millions PC's, 46.1 millions of mobile systems, and 1.7 millions of PDA's. This concludes that the mobile systems are the dominant agents, because ease of use. In USA, by 2005, about 55.3% have Internet access [1]. Statistical study for Internet users is shown in Figure 1.

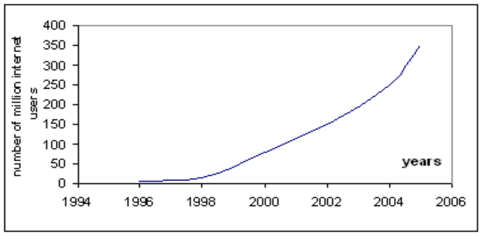


Figure 1. Growth in Internet over the world [1]

The studies by 2006 [2] indicated that 35.8% of the world population have computers, and only 15.7% have Internet access. This percent increases in the Asia by a rate higher than in Europe. Conrad [2], expected Internet users by 2010 to be 60%. The world's economic crises reduced this ratio [3]. The percentage of Internet access depends on the user's age; as shown in fig2. The dominant user ages are between 18 and 30; these form 52% of the users. They visit adult, news, and job sites. The percentage of using Internet decreases for ages higher than 50 years. Internet performance is improved annually by about 40% [3]. Performance of Internet is estimated via set of metrics like: admissibility, sustainability, trip delays, and loss ratio. These metrics affect the users' temper and Internet returns.

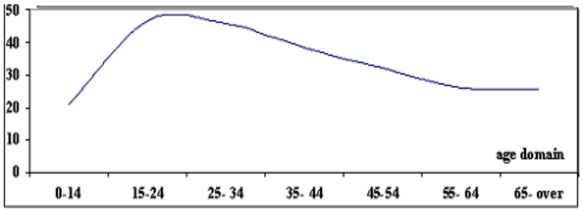


Figure 2. Statistical study for using Internet among ages in USA [1]

III. INTERNET USAGE

Middle east, Asia, and Africa still use Internet with rates about 6%, but with a high growth rate. The highest Internet use is accomplished in Japan and North America, as shown in Figure 3.

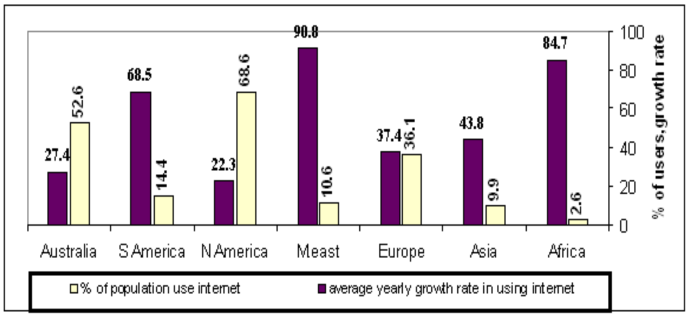


Figure 3.a- Average use for Internet and average growth rate [2]

Most users use Internet via home ADSL, Internet centers and telephone lines as shown in Fig 4[4]. In France, by 2005, a study shows that more than 55% of people use Internet, 26% of them use modem dialup, 53% use ADSL, other 21% use wireless cells.

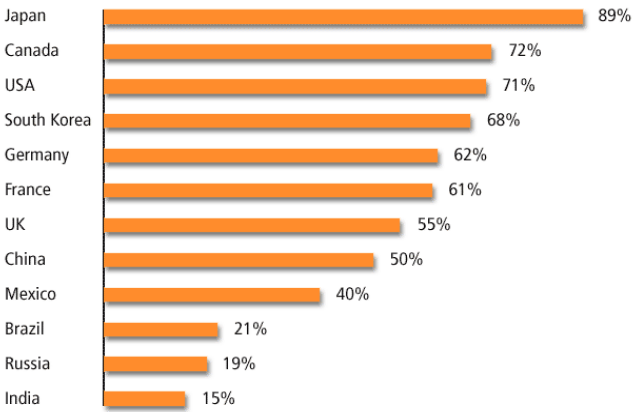


Figure 3b. Percentage of Internet use in a set of monitored countries [4]

The rate of using wireless net increases while the dialup rates reduce. In Arab countries, Arab Emerit has the highest ratio for Internet users (about 30%). Saudi Arabia users are about 10% , while in Egypt users are about 12% of population, as shown in Figure 4.

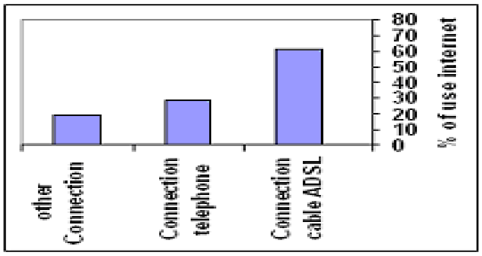


Figure 4a. Rates of using Internet media [2]

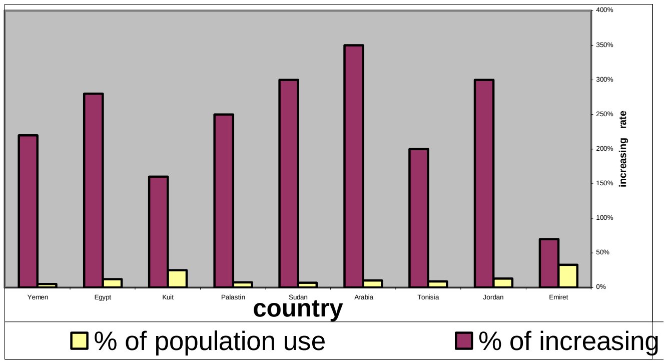


Fig4b. Internet in Arab countries [21]

IV. INTERNET ECONOMY

ISPs develop advertisements, adult videos, news, free software, free email, messengers, IDM accelerators, on line games, search engines, etc., to lure users to spend the longest time on Internet. The average money spend for Internet per month over the world exceeds $5 billions [6]. The peak periods of using Internet are about 10AM, and 9PM. This time varies due to the geographic timing among the countries. Therefore, the Internet load is almost regular. The access to the Internet depends on levels of life, and culture as shown in Figure 5. It was noticed that people with high income, use internet with more ratios. The accessed sites varies depending on the ages and cultures of users. It was noticed in US and Europe [6] that 87% of the families that have colleagues, have access to Internet, while less than 70% of the families that have no colleagues, access the Internet. The e-commerce are the most objectives, while the educational sites have low access rates and form about2% of the accessed sites [5]. Figure 5, indicates relations between HDI (Human Development Index), and the use of Internet.

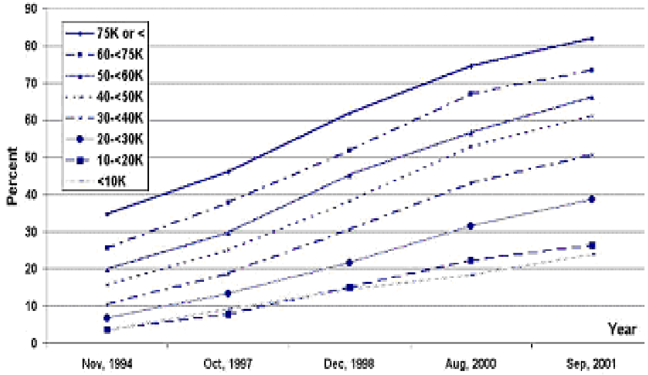


Figure 5a. Percentage of Internet users regarding Iicome in US, age> 18 years [6]

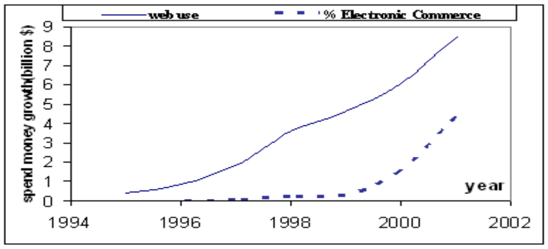


Figure 5b. Average money/year for Internet and e-Commerce, in US [6]

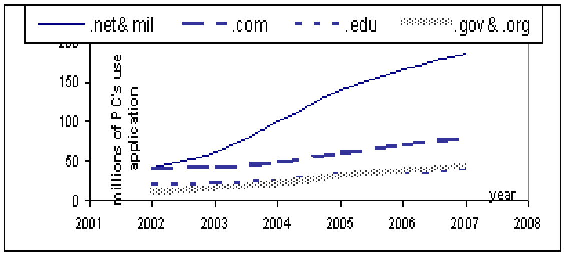


Figure 5c. Internet accessed sites, 2005 [4]

TABLE I. THE STATISTICAL TOPICS OF INTERNET APPLICATIONS IN THE WORLD [3][7]

Adult sites, Spams, e-commerce, email, interactive games, declarations, etc.	28%
File transfers, scientific downloads and uploads, e-learning, training programs, software activities, etc.	21%
Youtube videos, clips, audios	14%
Video conferencing, chats, forum	18%
Job seeking, news, web searches, etc.	19%

Spammers and Intruders destroy information and emails. Spammers catch the emails of victims by offering free services, forum, online games, chat rooms, messengers, videos, pornographic sites, anonymous connections, fake websites, lottery scams, etc. USA, China, Korea, and Russia deliver more than 93% of the world's spasm [8]. Regarding the languages, the percentage of Chinese emerged sites is 14%, Arabic sites is 12.8%, while the English sites are emerged by a rate more than 70% [8] as shown in Figure 6. Anti spam programs can detect only less than 80% of the spam. Spam filters may cause negative detections (reject the needed messages) or positives detections (reject un-needed messages), according to the used keywords. Curious and Intruders form more than 40% of spammers. Spam data consume bandwidth of Internet [8]. This results in less available bandwidth and more paid fees for the same data. A statistical e-mail spam was declared by ITU [7] as shown in fig6. The plot indicates that the advertisements trade form about 50%, while adult dates, pharmacy and chat form more than 30%.

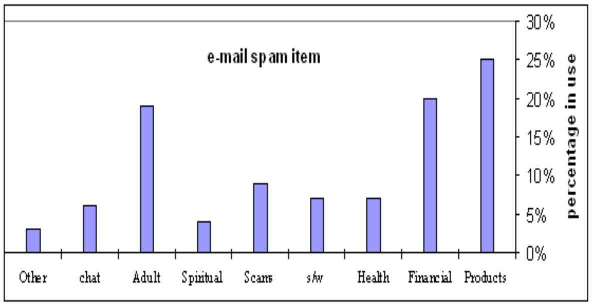


Figure 6-a. Percentage of Internet spam [7]

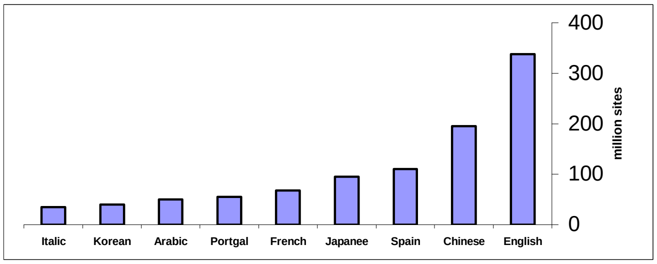
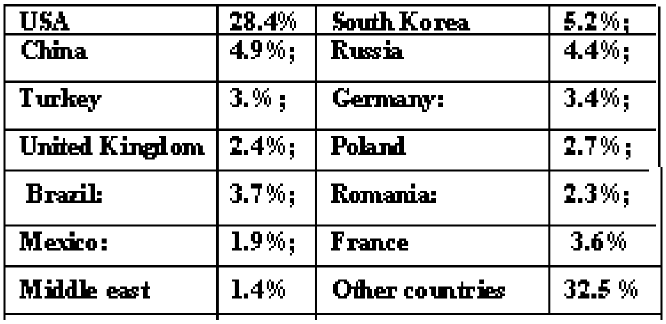


Figure 6b. Percentage of Internet sites versus language [21]

In 2006, ITU recorded an average spam with $55 billion. In 2007, this rate was increased to more than $100 billion [8]. In 2008, the estimated cost of Internet spam in US was about $17billions, and formed more than 92% of the received e-mail [9]. Filtering spam improves the bandwidth, admissibility, sustainability and serviceability of the Internet. About 80% of spammers in America and Europe are malicious. The other 20% entities use Internet mails for marketing in low cost. E-mail spam grew exponentially, it comprises more than 85% of the e-mail among the world, which results that about 85% of Internet utility waste for annoyments [10]. This decreases the returnee for the Internet vendors. Table 2 indicates the countries that spread spam around the world.

TABLE II. COUNTRIES THAT SPREAD SPAM [9]



USA	28.4%	South Korea	5.2%;
China	4.9%;	Russia	4.4%;
Turkey	3.%;	Germany:	3.4%;
United Kingdom	2.4%;	Poland	2.7%;
Brazil:	3.7%;	Romania:	2.3%;
Mexico:	1.9%;	France	3.6%
Middle east	1.4%	Other countries	32.5 %

EUIMC estimated the cost of "junk email" to be €10 billion [10] in 2005, while the spam cost in USA exceeded $13billion in 2007 [8]. In 2008[9], about $7billions were recorded for adult pharmaceutical spam**.** Anti-spam programs were developed to avoid attacks, protect and filter the junk e-mail, block spam, and terminate spammers' accounts. Spam Avoidance Routines **(**SAR) are saved in firewalls and routers to block the spam spreading. SAR catch the duplicated messages and delete them. Many SAR use statistical filters like: Mozilla thunderbird, mailwasher, and spam assassination software. SAR list spammers, deny their access, use learning techniques to analyze the content of e-mails, and filter spams [10].

V. INTERNET HARMS FOR HUMAN HEALTH

Taylor [11] concluded a relation between computer exposure and early cataract. In 2001 [12], a study investigated a form of eye cancer called "melanoma" caused for computer addicts. The research sounds that there is a threefold increase in eye cancers among people who regularly addict the use of Internet. Long exposure periods, cause brain headache and idiocy. Long periods of hearing clips, using headphone may deteriorate eardrum. The radiation penetrates the organic cells of chronics, and affects the regularity of endocrine and pineal gland[15]. Cataract, results in loss of lens' transparency, and may lead to blindness. Radiation suppresses the body's immune system [13,14]. HCI has limited the exposure power to be less than 10 mw. The electromagnet radiations affect the skin and may lead to melanoma. Dermatologists in American Cancer Society (**ACS**) declared that 20% of USA Internet chronics has melanoma that may lead to skin cancer [15]. The wireless Internet has the main effective ratio. ITU [3] had stated that the level of service sustainability affect the user temper and may led him nervous or gloomy.

Computers, mobiles and wireless LANs, radiate harmonics that cause blood ionizing. The harmonics cause cell heating and results in disturbing cell functions. Although the power of the harmonics is low, but the long sessions increase the doses that damage the cells[16]. Radiations interact and generate disturbed genes that perform malfunctions. When the structure of the cells is altered, this may result in cancers. An acute radiation dose is defined to be 10 radon, during a period of 3 hours or more. This causes eye inflammation. If the dose is greater than 100 radon, This may damage the lymphatic tissues and may cause convulsions [13]. The radiations harms depend on the power of the electromagnet radiations that is absorbed by the human body. When the power of a channel becomes weak, a hand over is assigned to another channel with a higher power. The high power signals are more harmful. Electromagnet radiations, thermal radiations, and ionizing radiations, cause unbalanced cellular performance which may lead to cancer. Bluetooth radiations, and radiations produced at private WLANs, GSM, may lead also to cancer [15]. The mobile Internet use 10 watts for upload and MSC use 100 watts for download. This power is exponentially decreases with the distance from MSC. Electromagnet radiations penetrate the human tissues to a depth depends on their frequency, and power. The high frequencies and high power penetrate more distances and result in more harms[16]. These radiations are absorbed causing heat, which disturbs the functions of human cells, and may cause fever or cancer. Edwards[17] recorded a relationship between Internet addicts and the cataracts. The study showed that the prolonged exposure to radiation, may lead to cell damages and these damages accumulate and may be unhealed. Cataracts result in loss the transparency of lens. If cataracts are untreated, this may lead to blindness. Radiation explodes the eyes' veins, causes high pressure in the eyes, and deteriorates the vision. Wireless radiations also reduce the efficiency of the immune system. Posluns [19] reported that a cell phone call lasting more than two minutes affects the activity of a child's brain for up to an hour afterwards. This brain electricity leads to a lack of concentration, and aggressive behaviors [19]. Neurologists reported a fourfold increase in the risk of Alzheimer's disease for persons who have jobs with EMF exposure [15].

About 65% of the sites are covered with wired links, about 30% are covered with wireless and satellite links, the other regions are unmonitored [18]. These areas are rare populations on south America, middle Africa, Arctic and Azerbaijan regions. Cables are more radiation immune. WHO and ITU recommended replacing the satellite links with fiber links [16]. The satellite links are used at the countries that suffer from poor telecommunication infrastructures [18]. Fig 7, indicates that the number of monitored sites doubles every three years.

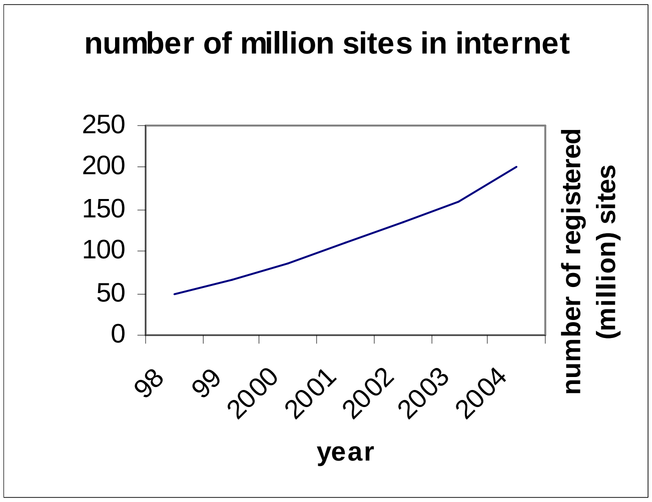


Figure 7. The growth of Internet registered sites

Interrupted or out of service sessions, degrade QoS, and lead user annoyments. The long delays reduce user satisfaction. This leads to anxiety, violence, or numbness. Internet' vendors developed anti-spam to filter traffic and omit threats. Table 4 indicates the percentages of Internet activities.

TABLE III. ACTIVITIES OF INTERNET [3,18]

Action	%	action	%
Business & video	11.9	search	11.4
E- learning	4.5	e-commerce, GIS	14.0
Content explore	8.3	Social deal	14.4
Mobile	2.8	Wild deal	2.0
Chat video	6.3	Photo sharing	1.4
messaging	4.9	Consulting services, forums	10.4
Hosted services	6.1	Software Down loads	1.6

Figure 8 indicates the percentage of the e-commerce ratios in Arab countries.

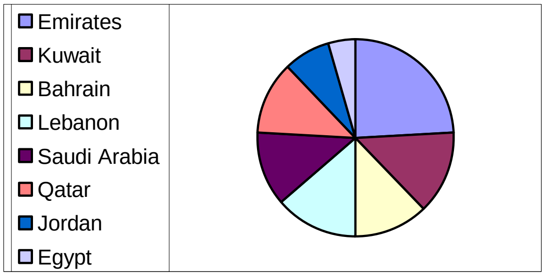
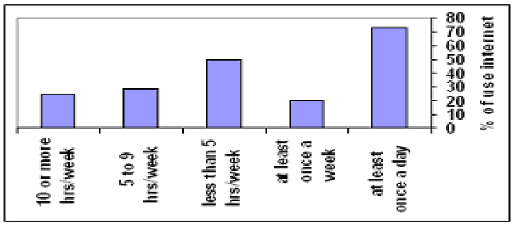


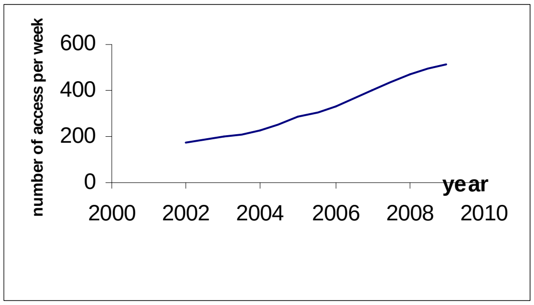
Figure 8. The e-Commerce ratios in Arab Countries [21]

VI. STATISTICAL EXPERIMENTS IN EGYPTIAN UNIVERSITIES

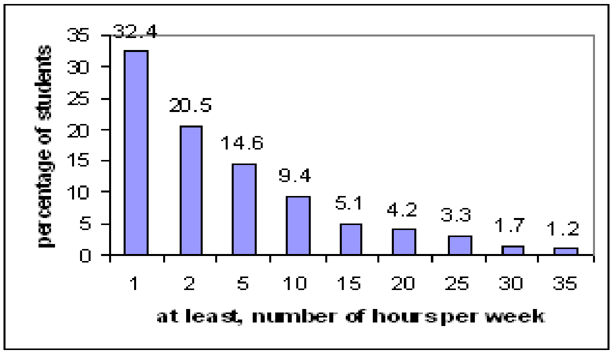
ITU [3] used the average growth rate for using Internet as a HDI. A study was done for Electrical Engineering students in Tanta, Zaqaziq, and Kafrel-sheikh universities. The study assured that about 60% of the students, visit Internet sites at least two hours per week, about 7.6% use internet for less than an hour per week, as shown in Figure 9. The average time of weekly Internet use in these campuses, is doubled every couple of years. The rates of Internet access are reduced at vacations by a factor between 10-40%. Users in age between 18-54 years form more than 92% of internet users. Internet addicts may daily access Internet for more than ten hours. About 83% of the accessed sites are by users whom age less than 30years. These users concern on chat, clips, news, forums, and movies. This concludes that the percentage of guided Internet use is less than 20% of the total use, more than 80% of internet economy waste for twitters and more than 80% of internet outlay has a negative returnee for time and money.



a. Rate of access Internet, 2007



b- Number of students access Internet / week



c– Percentage of Internet use

Figure 9. Sample of Electric Engineering dep.'s at Tanta, Zaqaziq and Kafrel-Sheikh Universities,
Sample size = 600 Students

Table 5, was collected in 20 Cyber-nets in Tanta, Zaqaziq and Kafrel- sheikh cities (Egypt) during summer 2009, it indicates the most benefits and problems arises by internet usage

A study was done in 20 Cybers in Tanta, Zaqaziq and Kafrel- sheikh cities (Egypt) during 2009, the sample size was 200 healthy persons, 100 of them are Internet addicts, with ages between 40 and 60years. This study indicated during the year, that addicts suffer from convulsion, 20% more than these that are not addicts. Their eyes suffer iris inflammation, and cataracts,15% more than these that are not addicts. Radiation for long sessions causes blood ionization, and high pressure, 15% of the Internet addicts suffer from high brain electricity, heart attack, and leukemia more than these that are not addicts. About 14% suffer hype, obesity, tingling, backaches, anal fissure, piles and muscles weakness. The family divorce rate was increased with 12% due to the virtual relations among the Internet. About 6% has paranoia, schizophrenia due to chat sessions.

VII. CONCLUSION AND REMARKS

The Internet consumes our time, money and health. The paper discussed the harms due to the ill-guided use of Internet, concerning the economy and health. The paper presented a statistical study for the Egyptian society concerning the problems of Internet addicts, and effect of wireless networking. The paper concludes the next guides for Internet users to reduce the harms. It is preferred to use LCD screens with desktop systems, connected to fiber land cables to reduce the effect of wireless networks. User must reduce the use of Laptop to avoid the electromagnet radiations. The use of Internet, must be not more than 2 hours, then take a break and resume the session later. This helps the body to discharge the accumulated absorbed radiations. People using glasses must use frameless eyeglass, with no wires, because the wires act as antennas that focus the radiation directly onto brain and eyes [11]. Internet addicts must take anti-oxidants and zinc doses, to reduce the effect of radiations on blood ionization. Zinc protects the eye from oxidative damage [13]. Long periods of exposure cause headaches, dizziness, lack in concentration, eye irritation, depression, anxiety, fatigue, muscle spasms, and numbness. Radiations disturb the pineal gland that regulates sleep, blood pressure, cholesterol levels, and immune system [17]. Therefore, the guided use of Internet, will save at least 80% of our time against Internet, save effort, health, and result in a perfect returnee.

REFERENCES

[1] Ran Atkinson, Almeida, V. Almeida, and D. Yates. Measuring the Performance of World-Wide Web Servers. Technical Report, Boston University, USA: pp. 139-147, 2005.

[2] David Conrad, General Manager, IANA, Bangkok, Statistics of Internet Zones. IANA annual report, Nov 2006.

[3] C.1, General Telecommunication Statistics, statistical year book, ITU recommendations, 2008

[4] Edwin K. W. Cheung, Institute of Vocational Education (TM), Hong Kong. A Statistical Analysis of Students' Perceptions In Internet Learning, Systems, Cybernetics and Informatics, Volume 1 - Number 2, 2007

[5] Cheung, E. and Cheung, W., Empirical Study of E-Commerce, Proceedings of The 9th International Conference on Computers in E-commerce (ICCE) , Net2002, Vol. 2, pp 719-724, 2002

[6] Mark Pollard, Morin A., UK National Statistic Institute, Internet Access, annual economic report, Aug, 2008.

[7] ITU-T Recommendation, D.600R, Spam Cost Methodology for the Regional Tariff Group, by ITU-T, the World Telecommunication Standardization Assembly group, 2005.

[8] Jesse Varsalone, Spam Statistics, Facts, Security and Administrative Guide, Elsevier, 2008

[9] Coase R.H. , Libbey M. Robert P., and Ferris R. MIT Spam conference, Cambridge MA, The Problems Of Spam Economic Analysis, Journal of Organizational Computing and Electronic Commerce, v18, no 4, 10-28, 2009

[10] European Union for Internal Marketing Commission, Geneva, Internet Economy Report, 2009.

[11] H R Taylor, S West, B. Muñoz, F S Rosenthal, S B Bressler, Eugene Sobel, The Long-Term Effects of Visible Light on the Eye, Report of Department of Ophthalmology and Visual Sciences, Washington University, School of Medicine, USA. 2005

[12] J J Broerse, A Snijders-Keilholz, , Department of Clinical Oncology, Leiden University Medical Center, Report on Effect of Radiation For Human Eye, The Netherlands, 2006.

[13] P Chen, W Hao, L Rife, X P Wang, D Shen, J Chen, T Ogden, G B Van Boemel, Department of Ophthalmology, University of Southern California, Los Angeles, California, USA. Report on Safe Radiation, trends in wireless health care technology , 2008

[14] E. Motta , Jennifer Turney, and S. Wiedenbeck , vendome group,Internet Usability Evolution, Journal of HCI report , 282-306, 2009.

[15] National Research Council of National Academies, www.national-academies.org, Washington, DC. Encyclopedia of Health Effects, EMF Effect on Human Cells, 2009

[16] Laura Ost. And Wonil Hwang, NIST(National Institute of Standards and Technologies), Boulder, USA. www.nist.gov/index.html, Annual Report For Safe Radiation, 2005

[17] Gordon Edwards, Health Risks From Exposure to Low Levels of Ionizing Radiation, Canadian environmental advisory, 2008

[18] Cheung, E. and Cheung, W, "Performance of Web Tools", proceeding of SCI2001, July 22-25, 2001.

[19] Jeffery Posluns, Mobile Community, ISBN 1-932266-86-0, Elsevier pub. San Diego, 2008.

[20] J. T. Jansen, Anitesh Barua, and C. Klein, Center for Research on Electronic Commerce, Texas, USA. Measuring the Internet Economy, E-Commerce& E-Government: Benefits, Awareness and Services, The Arab Telecom and Internet Forum (2003). 5th ITU Annual Meeting for Telecom Development in the Arab Region, in conjunction of 5th Arab Telecom and Internet Economy Conference, Beirut, May 28-31, 2003

[21] Technical report. Workshop on Digital Library Services, The Arab Telecom and Internet Forum (2008): 6th ITU Annual Meeting for Telecom Development in the Arab Region, in conjunction of NIT: National Institute of Telecommunications, Beirut; Lebanon, May 28-31, 2008

TABLE IV. THE MAIN TOPCIS OF INTERNET REGARDING THE AGES AND PERCENTAGE OF USERS IN EGYPT SOCIETY

age	percentage	popular interest w.r.t. age	problems	benefits
<16	24%	games, chats, clips, movies,	violence, idiopathy, dumpiness, violence, crimes, curiosity, crimes, abnormality,	build intelligence curiosity, prune emotions, grow relations,
16<age<30	39%	job seeking news, forum, adult sites, email,		
30<age<40	14%	sales, exchange, conferences, software applications, e-learning,	divorce, abnormality, mal-parenthood,	software training, applied software, e-commerce, IT, news update, distance learning,
40<age<50 50<age< 60	10% 8%	advertisement, web search, electronic government, e-commerce, , commercial trades,	high brain electricity, rheumatism, high pressure , heart attacks, hypertension, blood ionization, iris inflammation	
age>60	5%	news, forum,		entertainments

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-116-8

Automated Organizational Network Analysis for Enterprise 2.0

Hady Abi-Nader
School of Systems Engineering
University of Reading
Reading, UK
h.abinader@reading.ac.uk

Abstract— **Social Networking Sites have recently become a mainstream communications technology for many people around the world. Major IT vendors are releasing social software designed for use in a business/commercial context. These Enterprise 2.0 technologies have impressive collaboration and information sharing functionality, but so far they do not have any organizational network analysis (ONA) features that reveal any patterns of connectivity within business units. This paper shows the impact of organizational network analysis techniques and social networks on organizational performance, we also give an overview on current enterprise social software, and most importantly, we highlight how Enterprise 2.0 can help automate an organizational network analysis.**

Keywords - enterprise social software; enterprise 2.0; social network analysis; web 2.0; organizational network analysis; ona.

I. Introduction

Social capital and innovation are nowadays regarded as one of the main competitive advantages of an organization, a factor that helps succeed in a highly competitive market and overcome tough economic conditions. Social capital depends on social interaction. In recent years, social network sites like Facebook and MySpace have been booming with popularity. These Web 2.0 technologies are now coming in packages specially designed for enterprises as major vendors (IBM, Oracle, Microsoft) are releasing social business software. This phenomenon is described as Enterprise 2.0

In today’s knowledge-intensive organizations, most work of importance is heavily reliant on informal networks of employees within organizations. However, most organizations do not know how to effectively analyze this informal structure in ways that can have a positive impact on organizational performance.[1] Apart from implementing a collaborative technology in the workplace, many organizations are not taking any concrete actions to support these networks. [2]

A Social Network is a joint combination of actors and relations. [3] Social Network Analysis displays relationships as graphs, with nodes representing individuals, and edges representing interactions types. The degree and type of interactions may be represented by the lengths and widths of nodes and edges. Even though there are many tools to help perform a social network analysis for an organization, the information used to perform the analysis is almost always gathered manually through different ways such as observation, recording of activities, and/or long questionnaires to employees, interviews and diaries. Incorrect reporting might happen when participants record their own activities, and different interactions with other employees may not be remembered equally as well. [4]

This process can become extremely complex, time-consuming, and costly when applied on large organizations, moreover reliability of the results can be questioned because of the complex nature of this process.

The aim of this paper is to show an existing gap in order to help to bridge the wealth of information in organizations that has recently become accessible by enterprise social software with the social network analysis techniques in use. In this field, some work has been done on automated social network analysis technologies on interactions by email messages, files, instant text messages, network activities, etc. [5] but it seems that no work yet focuses on automating the process of organizational network analysis based on the data in enterprise social software, which is the focus of this paper.

The next section will show an overview on current Enterprise Social Software on the market and their different features. Section III will introduce Organizational Network Analysis and what it involves. Finally, Section IV will show the further research intended on this topic.

II. Enterprise Social Software

Small and large enterprises are always looking for tools to bring their employees closer together, even for workers located in the same place, they are often part of multiple projects which require attention and time management, thus the need for web-based collaboration. [11]

Social Network Sites offer many features from forming communities and creating blogs, to sharing photos and videos and organizing events. These technologies are now coming in packages specially designed for large enterprises. [12] As compared to Facebook that “helps you connect and share with the people in your life”, the following tools keep the employees in touch with their colleagues and the projects to get work done.

The most popular Enterprise Social Software on the market is Jive’s Social Business Software previously known as ClearSpace. In addition to *Profiles*, which is the main social networking feature, Jive has an *Analytics* Module which tracks user activities and offers some statistics. Information can be easily exported to other formats and tools such as Microsoft Excel and Microsoft Access. An *Insight*

Module reports on user engagement and sentiment extracted from user-generated content. [15] [16]

IBM Lotus Connections is the closest product to Jive's [14]. Lotus Connections wraps many social networking technologies into one package. It has discussion forums, blogs, bookmarks, and a *homepage* allowing a customized overview of the user's social network: A list of colleagues' profiles, new entries in the Wikis, latest blogs, popular bookmarks, and new activities. [17]

Socialtext is a platform that allows organizations of all sizes to collaborate using social networking features. It has *collaborative weblogs* to facilitate internal communication and a social messaging service for micro-sharing among a group of colleagues to share brief messages like on Twitter, messaging is in near real-time and kept to short messages of 140 characters to encourage briefness. [18]

Microsoft is also on the Enterprise 2.0 market through SharePoint. In SharePoint, Microsoft focuses on function over user-friendliness. SharePoint has some collaborative tools such as blogs, wikis, and message boards along with other features such as share calendars, task lists, etc. Components are not very well connected to others. [11] Microsoft added the social networking functions as features of its already-existing product rather than make a separate package. [12] *Team Sites* are a collaboration tool in SharePoint that offers groups some capabilities such as document libraries, lists, group calendaring, tasks, contacts, and announcements. Other features include the common Blogs, Wikis, and Discussions forums. [19]

Oracle has introduced new features to its Oracle Beehive enterprise-collaboration platform. The product now includes team workspaces and instant messaging, Web and voice conferencing, as well as the standard email, calendar, and other features. [13] Other major Enterprise Social Software include SuiteTwo and Salesforce.

A. Summary of functionalities

With the increased usage of Enterprise 2.0 software in organizations, it is expected that in the future employees might start to rely more and more on these collaborative technologies to collaborate more than email.

The Enterprise 2.0 tools previously mentioned encourage collaboration, save companies a considerable amount of time, and have many revolutionary features that capture the essence of Web 2.0 in all its ways. However, they do not offer a network insight or reveal any patterns of connectivity in a network, thus not benefiting from the dynamic and immense amount of information available in them.

The limitations of all existing enterprise software are that none of them contains a social network analysis module integrated in them.

Many of these have available source code, and support plug-ins, while others offer the option to export data to Excel or Access such as Jive SBS, or have an API such as IBM Connections, will be useful in the future for adding a tool for network analysis.

III. Organizational Network Analysis

Based on the incredible amount of media coverage, many people might believe that social networks are a recent discovery – a phenomena resulting from consumer participation in web sites such as MySpace, LinkedIn and Facebook.[7] However, social network analysis has been an interdisciplinary field/multidisciplinary method from the very beginning. [3] Network analysis can be traced to three or four disciplines: Psychology, Anthropology, Sociology, and Mathematics.

Social network analysis is the mapping and measuring of relationships and flows between people, groups, organizations, computers, URLs, and other connected information/knowledge entities. [8] The two main elements of any social network are actors and relations. Their combination jointly constitutes a social network. [3]

According to Knoke (2008), an actor may be an individual person or collectivities such as informal groups and formal organizations. Common examples of individual actors include employees in a corporate work team, or high school students attending a graduation. Collective actors might be firms competing in an industry, or political parties holding seats in a parliament. [3]

A relation is generally defined as a specific kind of contact, connection, or tie between a pair of actors, or dyad. Relations may be either directed, where one actor initiates and the second actor receives (e.g., advising), or non-directed, where mutuality occurs (e.g., conversion). A relation is not an attribute of one actor but a joint property that exists only as long as both actors maintain their association.

Organizational network analysis (ONA) can provide a deep inner view into the workings of an organization, a powerful means of making invisible patterns of information flow and collaboration in strategically important groups visible.[21] Even in small, contained groups, executives are often surprised by patterns of collaboration that are quite different from their beliefs and from the formal organization chart. [2]

There are a lot of research and work done in this area and still carried out. Cross and Parker are among the researchers leading in the field of Organizational Network Analysis are Cross and Parker. Rob Cross has been researching the area of applying social network analysis ideas to business issues and has worked with over 200 leading organizations on a variety of solutions including innovation, revenue growth, cost containment and talent management. Their work describes a full methodology to conduct an organizational social analysis, the highlight the process that transforms a formal organizational chart (Figure 1) into an information one revealed by SNA (Figure 2).

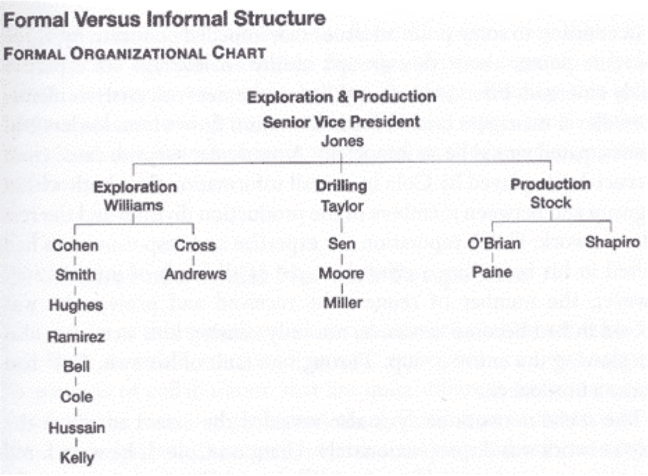


Figure 1. Formal Organizational Chart [20]

Conducting a Social Network Analysis manually involves the following steps of identifying a strategically important group, assessing meaningful and actionable relationships such as relationships that reveal collaboration, rigidity, or supportiveness in a network, then conducting a survey, and then analyze the results.

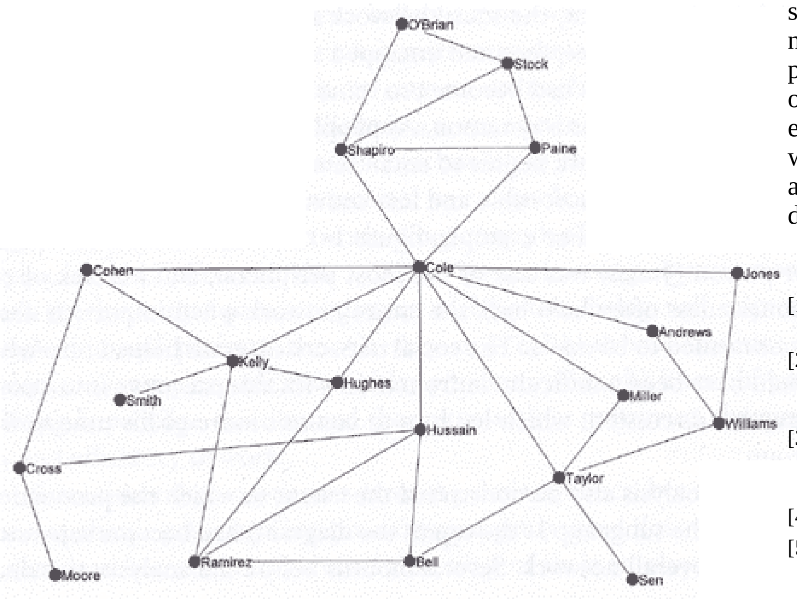


Figure 2. Informal Organizational Chart as revealed by Social Network Analysis [20]

Among the many revelations that were a result of the analysis in the example chart obtained in Figure 2 is that many of the senior people had become too-removed from the groups day-to-day operations, and this helped turn what could have been a difficult confrontation with a particular executive into constructive discussion that led the person to commit more time to the group. [20] Another point emerged was the role that Cole played, being the point of contact between different groups on the network and central for the information flow. The social network analysis revealed the extent the entire network was reliant on Cole, if Cole leaves his job, this would have a significant impact on the organization [20].

Organizational Network Analysis may be suitable in supporting mergers, partnerships, and large-scale change in companies where it may highlight the information flow and knowledge transfer. Similarly, it may be useful for developing communities of practice within an organization as the process may identify the key members of the community and assess the strength of connectivity within a community, along with other situations such as improving strategic decision making and promoting innovation [2].

IV. FURTHER RESEARCH

Network analysis can be very helpful in revealing patterns of connectivity in specific functions, divisions, or business units. [2] By going through the features of the main Enterprise 2.0 packages on the market, we saw the lack of the social analysis component in these enterprise software, hence the need to more research in this area to benefit from the abundant data that becomes available as a result of using these packages in organizations.

In the further research in this area, a tool will be developed to get social FOAF data from an organization's social software, and analyze it applying organizational network analysis techniques to automatically discover patterns of connectivity and the informal structure of the organizational chart as inferred from the interaction of employees on the social software. The automated analysis will remove the high cost and the complexity of conducting an ONA manually through interviews, surveys, without the data in social software.

REFERENCES

[1] Cross, R., A. Parker, and L. Sasson, Networks in the Knowledge Economy. 1st Edition ed. 2003: Oxford University Press, USA. 368.

[2] Cross, R. and A. Parker, The Hidden Power of Social Networks: Understanding How Work Really Gets Done in Organizations. 2004: Harvard Business School Press.

[3] Knoke, D. and S. Yang, Social Network Analysis. 2nd Edition ed. Quantitative Applications in the Social Sciences. 2008: SAGE Publications. 144.

[4] Mayfield, A., What is social media? 2007, iCrossing.

[5] Korba, L., et al. Automated Social Network Analysis for Collaborative Work. in Proceedings of the Third International Conference on Cooperative Design, Visualization and Engineering (CDVE). 2006. Palma de Mallorca, Spain.

[6] Gibson, O., Internet means end for media barons, says Murdoch, in The Guardian. 2006.

[7] Gotta, M. Analysis Of Social Networks: Telling Old Stories In New Ways. 2008 3 April 2008. Available from: http://mikeg.typepad.com/perceptions/2008/04/analysis-of-soc.html (accessed 25 September 2010).

[8] Krebs, V. Social Network Analysis, A Brief Introduction. 2008; Available from: http://www.orgnet.com/sna.html (accessed 25 September 2010).

[9] Rowe, R., et al. Automated social hierarchy detection through email network analysis. in International Conference on Knowledge Discovery and Data Mining. 2007. San Jose, California: ACM.

[10] Creamer, G., et al. Segmentation and Automated Social Hierarchy Detection through Email Network Analysis. in Advances in Web Mining and Web Usage Analysis: 9th International Workshop on Knowledge Discovery on the Web, WebKDD 2007, and 1st International Workshop on Social Networks Analysis, SNA-KDD 2007. Revised papers. 2009: Springer-Verlag.

[11] Harris, J. Jive's Clearspace Brings Collaboration and Social Tools To Enterprises. 7 April 2008; Available from: http://webworkerdaily.com/2008/04/07/jives-clearspace-brings-collaboration-and-social-tools-to-enterprises/ (accessed 26 September 2010).

[12] Hamm, S., IBM's Social Networking Push, in BusinessWeek. 2007.

[13] Kolakowski, N. Oracle Beehive Platform Offering Social Networking, Collaboration Tools. 4 May 2009; Available from: http://www.eweek.com/c/a/Messaging-and-Collaboration/Oracle-Beehive-Platform-Offering-Social-Networking-Collaboration-Tools-387441/ (accessed 26 September 2010) .

[14] Hersh, D. Another Clearspace Competitor: Lotus Connections. 24 January 2007; Available from: http://www.jivesoftware.com/jivespace/community/jivetalks/blog/2007/01/24/another-clearspace-competitor-lotus-connections (accessed 22 August 2010) .

[15] Jive SBS. Available from: http://www.jivesoftware.com/products (accessed 26 September 2010).

[16] Jive introduces SBS 3.0. 10 March 2009 Available from: http://www.kmworld.com/Articles/News/Industry-Watch/Jive-introduces-SBS-3.0--52984.aspx (accessed 25 September 2010).

[17] Lotus Connections. Available from: http://www-01.ibm.com/software/lotus/products/connections/(accessed 22 August 2010).

[18] Socialtext: Products & Services. Available from: http://www.socialtext.com/products/(accessed 22 August 2010).

[19] Microsoft SharePoint - Social Capabilities.Available from: http://sharepoint.microsoft.com/en-us/product/capabilities/Pages/default.aspx (accessed 30 September 2010).

[20] Cross, R., A. Parker, Prusak L., and Borgatti S, Knowing What We Know: Supporting Knowledge Creation and Sharing in Social Networks. 2001: Organizational Dynamics, Vol. 30, No. 2, pp. 100–120.

[21] Cross, R. Introduction to Organizational Network Analysis; Available from: http://www.robcross.org/network_ona.htm (accessed 22 August 2010).

Informal Virtual Organizations: A Perfect Home for SUBJECTs as Building Blocks

Wajeeha Khalil
University of Vienna
Department of Knowledge and Business Engineering
Vienna, Austria
Email: wk_rehman1@yahoo.com

Erich Schikuta
University of Vienna
Department of Knowledge and Business Engineering
Vienna, Austria
Email: erich.schikuta@univie.ac.at

***Abstract*—The present era has witnessed a rapid technological advancement, which has shaped our social connection into a new dimension. Online social networks are gaining acceptance and popularity among masses. PCs and mobile devices are used to connect with each other. Availability of online social networks on remote devices, such as cell phones, has made it even more pervasively. A large population of internet users today is actively participating in online social networks (e.g., Facebook, Myspace, Twitter, Blogger). The percentage of members is growing higher each year. Online social networks are a specific type of Virtual Organizations, called informal VOs. This paper focuses on users and resources gathered and provided by online social networks. It presents online social networks as informal VOs and develops a generic pattern of user and its special type, which we call *SUBJECT*. It identifies the roles served by users as Subjects in online social networks and it reveals how users can be ”consumed” as a resource in an informal VO.**

***Keywords*-Social Networks; Virtual Organization; Subject; Logical Resources.**

I. Introduction

A Virtual Organization (VO) is an orchestration of globally dispersed resources in a specific domain. VOs represent a combination of entities which are logically associated to achieve a goal. The building blocks of a VO are *Goal*, *User community*, *Tools* and *Resources*. These concepts are detailed in [1]. VOs provide resources which are utilized by users to solve their problems. In the early collaborations, these resources were restricted to storage, computational cycles, print facilities, high performance devices, parallel servers, simulation software, application programs and licensed software. The computing paradigm of Grid computing played an important role in providing such facilities, thereby, establishing a platform for VOs. Some examples of Grid based VOs are [2], [3].

VOs are of different types ranging from dynamic to fixed, temporary to long lived and formal to informal [3]. No matter what the type is, VOs offer resources to the users in the said domain to help them in problem solving activity. Specifically the Users in VOs play an important role. VOs creation, existence, evolution and deletion depends on user requirements. Users range from laymen to experts, beginners to professionals, learners to scholars from every field of life. Informal VOs are part of our lives in the form of social networks (e.g., Facebook, Myspace, MyExperiment) [3]. These user driven networks are typical examples of informal VOs where every user has its own goal for consumption and contribution to the resources pool. Today, online social networks are becoming essential part of life of humans who have access to Internet. People find it easier to connect to each other using social networks. These social networks can be visualized as a collection of small scale informal virtual organizations. Each user is given the right to access a number of resources offered by an online social network by creating a profile. These platforms give a sense of authority to the members by allowing them to initiate different activities. On the other hand, members can participate in the activities initiated by other members. Online social networks are an interesting area to study roles played by members. This paper identifies the resources available in a virtual organization in general. It reveals the role of users as a resource in an online social network. In the context of this paper, online social networks are presented as a special case of VOs.

In all types of VOs, users are classified into following four categories: consumer, contributor, developer, and administrator. Resources offered by VOs are utilized by users in these four capacities. VOs offer globally distributed resources to its users. With the technological shift, resources offered by a VO are also changed. The relation between user and resource is partially overlapping. This situation motivated us to review the users roles and resources offered in a VO. Thus we introduced a new term for a special type of resource in VO in [1], [4], [5], which we called *Subject*.

VOs exist around us in many shapes, as first choice of research community for collaborative computing. Still, standard pattern/methods for building VOs are missing. This situation is the motivation of our research to design a Reference Architecture for VOs. We believe that this endeavor will not only help the research community to find a starting point for building a VO, but also will provide them with a blueprint to extend existing pilots towards a fully-fledged environment for their needs (i.e., identification of existing/missing components). In our vision users and resources are main building blocks of a VO. Therefore we introduce the newly defined notion of *Subject* as a generic building block of a VO, as our first step towards building a

ISBN: 978-1-61208-116-8

Reference Architecture for Virtual Organization, which we call RAVO. A Reference Architecture for Virtual Organization can be defined as ”an open source template that does not only depict the architectural patterns and terminology, but also defines the boundaries where heterogeneous resources from different domains merge collaboratively into a common framework”.

A Subject is defined as a component of a VO, which can consume the resources, offered by a VO and also can act like a resource to be consumed in the VO environment. Therefore a Subject resembled the generic block of a VO, which results into a new definition for VOs. In the view of above discussion we propose a VO as *a set of cooperating building blocks, called Subjects*.

The rest of the paper is as follows: Section 2 defines the resource hierarchy in VOs. Section 3 explains the user acting as a resource and details a generalized pattern for users in VOs. Section 4 contains examples of users focusing their role as *Subject* in informal VOs. Section 4 concludes the paper with future aims.

II. The Resource Hierarchy

A virtual organization is a nonphysical communication model which aims to achieve a common goal. It consists typically of a heterogeneous collection of people and organizations with respect to geographical limits and nature. The existence of a VO is typically identifiable by many individuals, ad-hoc groups, research teams, and national and international organizations deploying a wide range of resources [5]. Initially, resources were meant to be hardware such as storage, high performance devices (measuring earth quake, weather forecast, printers, etc), and software (applications, utilities, simulation facilities) [3]. The extensive use of computer technology for problem solving changed the nature of resources [6]. Now resources are distributed as *logical* and *physical* resources. Defining a resource in a VO environment is dependent on the participating entities and domains in which the VO operates. A categorization of resources is presented in Figure 1.

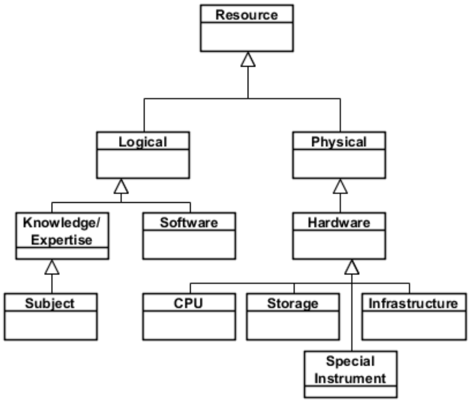


Figure 1. Resource Hierarchy in VO

In our research endeavor, a complex but interesting relationship was discovered between user roles and resources [1]. During resource consumption and contribution, at a certain point, user roles and resources are interchangeable. Some may find these concepts overlapping. Previously, resources are purely considered to be something, which is being consumed by the user, as shown in Equation 1.

$$USER \overset{Consumes}{\leftarrow} RESOURCE \tag{1}$$

However, resources are also contributed by users in a problem solving activity. This situation is defined by Equation 2.

$$USER \overset{Contributes}{\rightarrow} RESOURCE \tag{2}$$

The situation becomes even more complex, when a resource itself is a user. For example, in an exploration scenario, a meteorologist wants to know the reason that causes an unexpected storm. By chance she is the member of LEAD VO [7], [8]. So she searches for available data sources, and utilizes the tools offered by this VO for the analysis. In case of non-satisfactory results, she consults an expert for guidance and performs an analytical activity with changed data sets. In this scenario, the expert opinion is used as a resource, while experts also utilize VO resources for gaining knowledge [8]. In online social networks, the same situation can be easily identified in group discussions. A member asks a question and other members share their experiences, which can provide potential solutions to the problem, and vice versa. So the equation takes the shape as shown below

$$RESOURCE/USER \rightleftharpoons RESOURCE/USER \tag{3}$$

Even more according to our definition of Subject, the equations above can be generalized to Equation 4.

$$SUBJECT \Longleftrightarrow SUBJECT \tag{4}$$

Here, the user is consuming the knowledge of an expert, who acts both as user and resource. Subject is the notion given to a user who itself can be used as resource. There are two reasons for choosing this term. First, a Subject (user) initiates an activity in the VO environment and secondly, a Subject (resource) is under consideration to be useful in a problem solving activity. Figure 2 shows the Subject, resource and user relationship in different types of VOs.

III. Subject as a Resource

The previous section established the human expertise as a resource in a VO. Now we will present a template for a Subject class with its functions and relations to other user types in a VO. User types are a must for a system, because it helps

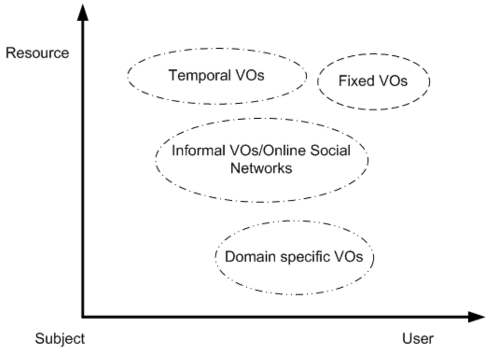


Figure 2. Subject, Resource and User in different VOs

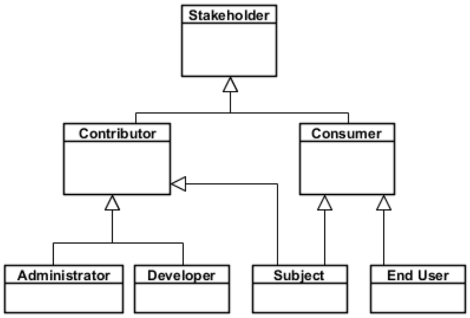


Figure 3. User Roles and their dependencies

in many ways, e.g., by defining trust, building a business model / negotiation model, setting security, authorization at different levels, managing the incoming and outgoing traffic (in dynamic environments), consumption and contribution of resources, and many more. In VOs a user is given a role according to a business model or negotiation pattern for collaboration. Roles may vary as the target domain changes. There are few characteristics and activities that can be generalized. Figure 3 presents the classification model of a *user* in a VO environment. This model covers both formal and informal types of VOs. A class diagram is created using UML 2.0 to present the pattern.

A User is the superclass with attributes defined in the context such as

- Id: string (any combination used for authentication)
- Role: string (assigned role in the said VO)
- Status: string (active, passive)

A User class is further specialized into two broad categories of **Contributor** and **Consumer**.

- The **Contributor** class presents the instances of a user, who contributes to the VO. The main method is **contribute**(). This class is further divided into **Developer** and **Administrator**, performing pure contribution and no utilization. Here, the **Subject** is also a subclass of **Contributor**. It realizes the role of a contributor while it can also act as a consumer in a VO environment. It inherits the attributes of the main **User** class and provides contributing methods (functions). The type of contribution can be added according to the role of the user and the domain in which it is active.
  - **Administrator:** Another potential subclass is Administrator, who monitors the VO platform for smooth use and in case detects and manages hardware and/or software crashes. Again the Administrator class can be represented by a group of paid experts, who are specialized in their respective jobs assigned. Participating organizations can hire such professionals to monitor the services they are offering to the VO.
  - The **Developer** class includes the professionals and application developers from participating organizations. There can be professionals who contribute open source software to the improvement of IT support in specific domains. However, any person can contribute knowledge in form of applications in a specific domain, even if they are not member of a participant organization.
- The **Consumer** class represents the class of users, who just utilize the resources by performing pure consumption only. This class has two subclasses called **End User** and **Subject**. It contains a method **consume**(), which shows that an instance of this class will be able to consume the resources offered by the VO.
  - The **End User** class represents a set of users who only consume the resources provided by the VO.
  - The **Subject** class represents the category of users who utilize the resources and also contribute to the VO environment. Currently two generalized methods are associated with the Subject class namely, **contribute**() and **consume**(). An instance of the Subject class is capable both to utilize the resources of the VO and to contribute to the VO at the same time. Instances of the Subject class can act as a Consumer or Contributor (as a resource), who share partial characteristics of their superclass.

A business model can also be developed on the basis of this categorization. Users belonging to the **Subject** class, can be given a high priority. This priority can entitle them to benefits such as money, free memberships to different participating organizations, utilizations of resources (test beds access, and access to reference material, etc).

The Contributors can have the 2nd highest priority, because they are the paid members of the VO. They develop tools for the maintenance of VO and monitor it. Such users are employed by the system. A possible subcategory could be developers contributing open source applications for the improvement of IT environment. They can be given priorities according to their contribution to the system, e.g., free resource consumption.

IV. Informal VO and Subject

The above described user classification can be observed in different domains. We presented the elaborated roles in VOs for E-learning [4] and Computational Intelligence (CI) [5]. Both are examples of formal virtual organization. *Ambient intelligence* is taking social networking to a new level of awareness [9]. This awareness is propagated from relatively constant contact with one's friends and colleagues via social networking platforms on the Internet. Informal VOs realize the concept of ambient awareness. Social network fall in the category of informal Virtual Organization. Informal VO are characterized by absence of a specific goal,rather they are user driven [3].

Online social networks are user driven, with no specific goals. However, they can be joined to meet several goals (e.g., making friends, playing games, joining research, religious, social, health, sports groups, to communicate with distant relatives or friends, promote different causes, advertise, participate in discussion forums etc). Goals can be anything supported by the platform. Here, it can be clearly observed that every user is a resource of this informal virtual organization. It exists only due to the relationship between the users and improves with the feedback they provide. Popular social networking websites are Facebook [10], Myspace [11], Twitter [12], and Blogger [13], etc.

To justify the patterns developed in the previous section, we choose Facebook as an example of an informal VO. Facebook [10] is a popular online social network launched in February 2004. It is selected as an example to identify the roles and resource dependencies in informal VOs. The activities performed by Users are

- Create a profile, update and set privacy settings, delete and add applications
- Add people as friends (send, reject and accept requests)
- Send and receive private and public message
- Notify of updating to friends
- Define status settings
- Chat with online friends
- Make lists depending upon privacy settings
- Add photos
- Add videos
- Create notes
- Join networks organized by workplace, school, or college
- Like fan pages
- Join and start groups, networks
- Send a virtual ”poke” to each other (a notification in turn tells a user that they have been poked)
- Send gifts
- Visit marketplace
- Play games

In social networks every need or goal is dependent on another user. If a user wants friends, so this user is looking for a resource (friend). She plays games, which are provided by other users (in most cases). She joins a cause, which is initiated by another user. In any of the above listed actions a user needs other users and their expertise or shared information to fulfill her needs.

On the other hand, information, expertise, material, pictures, videos provided by her can act as a resource for other users. She can initiate fan clubs, discussion groups and any cause, to invite people and grow her community. A use case was developed to understand the activities performed in this informal VO, shown in Figure 4. The user roles and their interaction with existing resources is detailed below.

- **Subject :** The role of a user as a resource is more profound in an informal VO than in formal ones. This is illustrated clearly in our current example of Facebook. A user creates an id and is given right to perform several activates, as listed above. Here the user is a contributor and a consumer herself. For example, a member uploads a video or photo or creates a note, which is being watched by other users and vice versa. Sending and receiving friend requests, messages (open and private), initiating groups, causes and campaigns, joining groups, reading and writing notes, sending and receiving gifts, communicating with friends through wall, and chat and status updates are the activities performed as Subject.
  A Subject also gains information from news feeds. An interesting facet are business promotions, which play the role of End user. Many products are introduced to E-communities using social networks by their manufacturers. Facebook is also used by different manufacturers to reach their customers. News channels, media, health, education, research communities, etc., all use social networks according to their requirements and goals.
- **Developer :** Members also play games, utilize applications developed and contributed by developers to the platform.
- **Administrator :** Group of specialized person(s) maintains the platform for performance, backup and routine maintenance.

V. Conclusion

This paper presented the concept of resources and users in both formal VOs and informal VOs. A resource hierarchy is defined and the role of a user as a resource was observed and discussed in different environments. The understanding of user roles is necessary for building a trust model for VOs. This approach was extended by a generic pattern for users in VOs and was justified using online social networks (e.g., Facebook). The concepts are elaborated with examples to understand when a user changes her role from a consumer to a resource and starts contributing to the environment. Hence the term *Subject* was justified.

Online social networks provide resources to its members. Every member contributes to the community silently. The impression of a member as a consumer is fading by growing needs of ”give and take” collaborations. This new concept of *Subject* fits well into the nature of online social communities. It will help in the future research on VOs to understand the concept of a Subject as a fixpoint where users and resources become the same. It will also set the bases of user roles in designing a Reference Architecture for VO (RAVO) as our future direction.

References

[1] W. Khalil and E. Schikuta, “Towards a virtual organisation for computational intelligence,” in *Proceedings of the 2010 Fourth International Conference on Digital Society*, ser. ICDS ’10. Washington, DC, USA: IEEE Computer Society, 2010, pp. 144–149.

[2] L. Hluchy, O. Habala, V. Tran, B. Simo, J. Astalos, and M. Dobrucky, “Infrastructure for grid-based virtual organizations,” in *Computational Science - ICCS 2004*, ser. Lecture Notes in Computer Science, M. Bubak, G. D. v. Albada, P. M. A. Sloot, and J. J. Dongarra, Eds. Springer Berlin / Heidelberg, 2004, vol. 3036, pp. 124–131.

[3] C. Kesselman, I. Foster, J. Cummings, K. A. Lawrence, and T. Finholt, “Beyond being there: A blueprint for advancing the design, development, and evaluation of virtual organizations,” NSF Workshop, Tech. Rep., May 2008.

[4] K. Wajeeha, M. Juergen, and S. Erich, “Veloci: A virtual e-learning organization for computational intelligence,” in *World Conference on Educational Multimedia, Hypermedia; Telecommunications ED-MEDIA 2010*, Toronto, Canada, 6 2010.

[5] K. Wajeeha, , juergen Mangler, and S. Erich, “Virtual organization for computational intelligence (voci):architecture and realization,” in *International Joint Conference on Neural Networks 2010 (WCCI2010)*, Barcelona, Spain, 2 2010.

[6] U. Farooq and W. Khalil, “Grid as human’s assistant: A logical solution provider for physical problems,” in *CTS ’06: Proceedings of the International Symposium on Collaborative Technologies and Systems*. Washington, DC, USA: IEEE Computer Society, 2006, pp. 312–317.

[7] [Retrieved: December 6, 2010]. [Online]. Available: http://portal.leadproject.org/gridsphere/gridsphere

[8] D. Gannon. (2008, January) Building virtual organizations around super computing grids and clouds. Indiana University and Tera Grid Infrastructure Group.

[9] E. Aarts and R. Wichert, “Ambient intelligence,” in *Technology Guide*, H.-J. Bullinger, Ed. Springer Berlin Heidelberg, 2009, pp. 244–249. [Online]. Available: http://dx.doi.org/10.1007/978-3-540-88546-7_47

[10] [Retrieved: December 6, 2010]. [Online]. Available: www.facebook.com

[11] [Retrieved: December 6, 2010]. [Online]. Available: http://www.myspace.com/

[12] [Retrieved: December 6, 2010]. [Online]. Available: www.twitter.com

[13] [Retrieved: December 6, 2010]. [Online]. Available: https://www.blogger.com/start

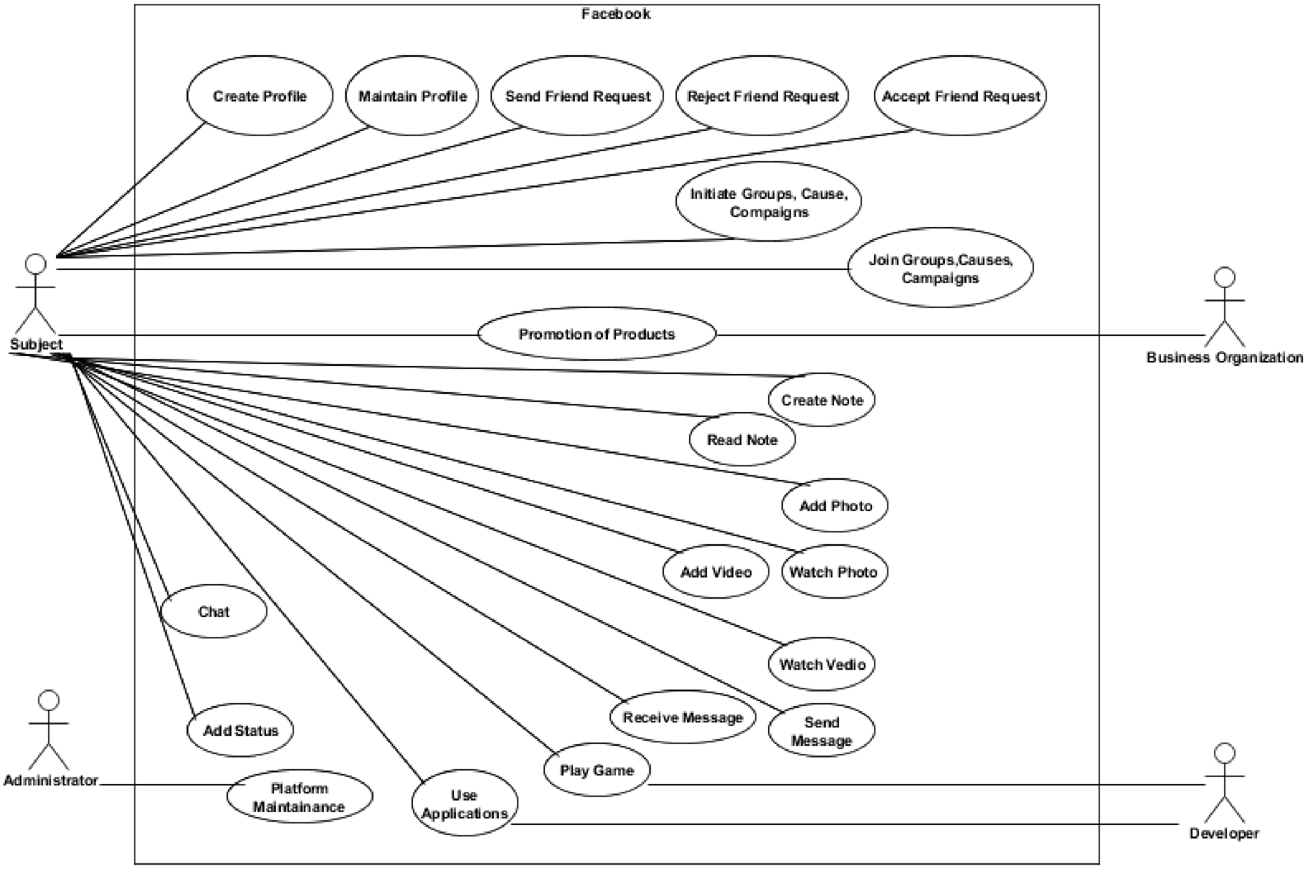


Figure 4. Facebook: User Roles

Trust in Peer-to-Peer Digital Society: An Economics Perspective

Dershing Luo
Department of Information Management
National Taiwan University
Department of Information Management,
China University of Science and Technology
Taipei, Taiwan (R.O.C.)
dsluo@cc.cust.edu.tw

Jen Wel Chen
Department of Information Management
National Taiwan University
Department of Business Administration
Chinese Culture University
Taipei, Taiwan
jwchen@sce.pccu.edu.tw

Ching Cha Hsieh
Department of Information Management
National Taiwan University
Taipei, Taiwan (R.O.C.)
cchsieh@im.ntu.edu.tw

Abstract—**Peer-to-Peer (P2P) technique manifests file sharing, distributed computing, and communication and collaboration between peers in contrast to client/ server architectures. P2P digital society is emerging in our life. It changes our way of living, work, and play. Meanwhile, the trust is still a core issue in P2P digital society. The objective of this paper is to survey trust in P2P digital society from the point of economics view. Through employing the framework previously proposed, some implications on trust in P2P applications are given from the economics perspective. In this paper, we analyze trust emerging in P2P digital society through the economics perspective from the following levels: technology, product, business process, market, and the macroeconomic.**

***Keywords* - trust; peer-to-peer (P2P); economics.**

1. Introduction

As a result of the quick transformation of the information society, human relationships and activities exhibit different styles than before. However, the concept of trust has remained fundamental for all social activities throughout human society. Trust is essential, especially in distributive information technology (IT) architectures as Peer-to-Peer (P2P) architectures. Research, both theoretical and empirical addresses the nature and role of trust, the moderators of trust, the antecedents of trust, as well as empirical methods in examining trust [5]. Whereas the economic analysis offers the researcher the opportunity to sort out explanations for changing market structure, to identify critical factors that result in firm success and failure, to interpret the speed and breadth of adoption of the new Internet technologies, and much more [9].

Kauffman and Walden [9] noted that many of the developments on the Internet seem new to most managers. But economic analysis provides a longstanding and well-developed theoretical vantage point from which to observe and interpret these developments in the context of continuing technological innovation in the business economy. Wang, Hori, and Sakurai [22] also noted the economic and social characteristics of trust system in P2P applications. Through the way, they may have an accurate characterization of the structural properties of the network. This can be of fundamental importance to understand the dynamics of the system. Only we have a clear picture of the economics of trust that the significant improvements can be expected. The objective of this paper is to survey trust in P2P digital society through this systematic economics framework.

Kauffman and Walden [9] proposed a comprehensive framework for electronic commerce (EC) research from the economics perspective. Though they aimed to survey EC researches, we employ their multiple levels of analysis from an economics perspective which is different from the usual technology-first perspective. Their level of analysis and relevant economic theories are noted in Table 1. In this paper, we will address the perspective of economics of P2P digital society.

The remainder of this paper is organized as follows. Section 2 introduces the trust and P2P applications. Section 3 describes the economics of the trust in P2P applications. Finally, future issues and applications are given in Section 4.

2. Preliminaries for Trust and P2P APPLICATIONS

We firstly state the concept of trust and then describe the issues of trust in P2P Applications.

2.1. Trust

The concept of trust has been studied in diverse contexts; therefore there are manifold definitions of trust. Trust has been defined in various terms, ranging from “the willingness to be vulnerable to the actions of another party” to “the probability one attaches to cooperative behavior by other parties” [18]. Trust has also been defined as the belief by one party that another party will behave in a predictable manner [13]. From the viewpoint of processes, trust may be also classified into two cases: one is general trust (developed between customers and

ISBN: 978-1-61208-116-8

companies over time and after experiences), the other is initial trust (developed after a customer's first experience with the company's web site).

Actually, trust is a psychological status of involved parties willing to pursue further interactions to achieve a planned goal. However, the key concepts are twofold. Firstly, for the trusting party, there must exist uncertainty about a potential or existing relationship—business, social, or otherwise—that leads to a certain perception of risk or vulnerability. Secondly, this perception of risk is generally based upon the beliefs regarding the ability, integrity, and benevolence of the trustee [11].

TABLE 1 LEVEL OF ANALYSIS AND RELEVANT ECONOMIC THEORIES COMPOSED BY KAUFFMAN AND WALDEN [8]

Level of Analysis	Relevant Economic Theories
Technology	Public goods theory, efficient pricing, game theory, negotiation, network externalities and standards
Product	Pricing theory, versioning, information goods, switching costs, network externalities and standards, economies of scale and scope
Business Process	Adoption theory, economics of design, cost-benefit analysis, allocation of benefits, IT value, negotiation, economies of scale and scope, game theory
Market	Auction theory, industrial organization, transaction costs, market microstructure, intermediation, adoption and diffusion, perfect competition, returns to scale, optimal market structure
Macroeconomy	Monetary economics, taxation theory, labor economics, regulatory economics, public goods

2.2. *Trust in P2P Applications*

Casadesus-Masanell, Hervás, and Mitchell [1] gave a vivid description of trust in P2P environments. They mentioned that as of 2006, the term 'pirate' was used more frequently to describe a person downloading music, movies and software than a buccaneer robbing ships on the high seas. Eye patches, swords and talking parrots had been replaced with software such as Napster [23], Kazaa [24], Limewire [25], eDonkey [26] and BitTorrent [27]. The battle between proprietary systems and P2P file sharing seems to be forever prolonged.

Trust also mediates the social relationships between websites and consumer characteristics and behavioral intent related to websites [10, 18, 21]. Reputation determines trust in an electronic store, which affects the attitude, risk perception, and in turn, the willingness to buy in an electronic store. The P2P reputation systems are closely related to the efforts of online communities, such as eBay, to develop incentive-compatible systems for rating the performance of a distributed set of users. Trust (and reputation) has related much with the incentives mentioned above. Kollock [10] proposed that the key design issues in evaluation are: evaluators, evaluations, history of evaluations, summary measures, and modifying evaluations. However, the design of reputation systems for P2P networks is complicated by two factors: one is that the distributed and intermediated nature of P2P network interactions makes it easy for users to conceal or change their identity, and the other is that in some fully distributed applications, the administration of the rating system must also be distributed throughout the network, making it vulnerable to coordinated gaming strategies.

Tan and Thoen [20] proposed a generic trust model for the successful performance of the transactions. They argued that the agent's trust in transaction is a combination of both external and internal factors. The external factors are the agent's trust in the other party (both objective / subjective trust reason) and its trust in the control mechanisms (both objective / subjective trust reason). The internal factors are potential gain and risk, along with its attitude. In [16] argued that trust in the vendor is defined as a multi-dimensional construct with two inter-related components—trust beliefs (perceptions of the competence, benevolence, and integrity of the vendor), and trusting intentions—willingness to depend (i.e., a decision to make oneself vulnerable to the vendor). Three factors are proposed for building consumer trust in the vendor: structural assurance (i.e., consumer perceptions of the safety of the web environment), perceived web vendor reputation, and perceived website quality.

The factors that determine trust are usually objective. Trust is built from experience (e.g., calculus), personal knowledge or bias (e.g., perception), and contextual elements (e.g., social characteristics) [13]. The evaluation of trust in a public key certification authority is already mentioned by [2]. They identified two major categories of trust factors, the first category being requirements for quality, quality of services offered, commitment, personnel responsibility and authority, and the second category being the conformance to qualified policy [13]. Krishnan, Smith, and Telang [12] noted the following approaches that provide quality of service differentiation to users based on a distributed mechanism that tracks user reputations: the evolutionary prisoner's dilemma (EPD), "stamp trading" mechanisms, "tit for tat" treatment, admission control system.

Significant questions addressed in this section include the various costs of services or topologies, the role of incentives in improving network performance, the motivations of users who consume and provide resources in P2P networks, the application of trust and recommendation mechanisms to the unique environments present in P2P networks, and the balance among copyright holders, entrepreneurs, and consumers. All these above advantages and trends are highly related to a great level of trust in the peers. Without trust, there can be hardly any success in P2P environments. This leaves a necessity of understanding trust, in particular, from the perspective of economics. When we stand on the foundation of economics, we can find the equilibrium of technology and the ecology.

3. ECONOMICS ON TRUST IN P2P DIGITAL SOCIETY

In this paper, we employ the framework proposed by [9] to survey trust in P2P environments from the perspective of economics. Moreover, we will identify important areas of research integrating an economic perspective into the analysis of P2P environment. In their framework, Kauffman and Walden [9] noted that there are two prerequisite parties. They are firms and consumers. Firms use related technology to create new kinds of products (especially information goods (e.g., MP3 music recordings and digital newspapers) and to recreate services (especially information-intensive operational services). Consumers use technology in a variety of ways as individuals, in homogeneous groups and overall in society at large so as to take advantage of the leveraging benefits that are offered (e.g., for improved product search, to acquire more information, to enable infomediation and for enhanced personalization, etc.)

Both Firms and consumers use technology to form the whole economic hierarchy. Besides the two players they mentioned, we think there is still one party which plays an important role. That is the third party, such as governments, and stand-alone organizations. They promote and maintain the stability of the whole economic hierarchy. Following these parties, we proceed to each analysis level from technology, to product, business process, and market, and moreover, to the macroeconomic. Through Lu's work [13], the existing P2P systems may be classified into three categories, these being file sharing, distributed computing, and communication and collaboration. Our framework is depicted in Table 2. The issues in each cell are illustrated in the following paragraphs.

3.1. *The analysis level of technology*

The technological innovations have impacts at a number of levels of analysis [8]. With regard to the economic characteristics, the services provided on P2P networks are different, hence there is a need for new theoretical models as well as empirical and experimental analyses in order to understand the behaviors of P2P users.

The design of reputation systems for P2P reputation systems, such as eBay, is complicated by two factors. First, the distributed and intermediated nature of P2P network interactions makes it easy for users to conceal or change their identity. Secondly, in many some fully distributed applications, the administration of the rating system must also be distributed throughout the network, making it vulnerable to coordinated gaming strategies [12].

Based on an economic model, [22] proposed a VCG (Vickrey-Clarke-Grove)-like reputation remuneration mechanism in order to stimulate rational peers not only to provide reputation feedback, but also to offer feedback truthfully. Considering that trust and reputation is subjective, they divided the trust into functional trust and referral trust, and extended the referral trust to include two factors: similarity and truthfulness, both which can efficiently reduce the trust inference error.

P2P is merely a network architecture that could be deployed to a number of other applications such as distributed computing, instant messaging, voice telephony, spam filtering, and other commercial activities. In the category of file sharing, there are several cost/ benefits we have to consider. They are free rider, latency (benefit) cost, service cost, routing cost, topology maintenance cost, etc. Among these, the free rider matters trust most. For the category of distributed computing, we have to take quality of service into consideration in order to ensure the trustworthiness of computing. According to the category of communication and collaboration, both security and reputation mechanism are employed to guarantee the trust among peers.

TABLE 2 THE ISSUES ON TRUST IN PEER-TO-PEER APPLICATIONS FROM THE ECONOMICS PERSPECTIVE

Level of Analysis	**file sharing**	**distributed computing**	**communication & collaboration**
Technology	● free rider	● quality of service	● reputation mechanism ● Security
Product	● intellectual property rights ● extent ● balance ● free-riding ● privacy	● specialization ● collaborative filtering ● semantic web ● privacy ● intellectual property rights ● storage capacity	● contents ● recognition schemes ● payment methods ● privacy ● quality of service
Business Process	● effect ● presence management ● network management ● free-riding ● simplicity ● robustness ● liability ● motivation free-riding	● efficiency ● collaborative filtering ● mobile agents ● frequency of requests ● semantic web assistance ● collaborative filtering ● mobile agents	● type of services ● contents ● quality of service ● interaction ● altruistic behaviour ● regulation ● rewards ● recommendation agent
Market	● user behaviour,	● a third party ● trust provider ● social interaction	● valuable information ● the trustworthiness of the participants
Macro-economy	● culture ● intellectual property debate	● social welfare	● labour market

3.2. *The analysis level of product*

In a traditional economy, we have a classification of goods as depicted in Table 3. However, information goods share the properties of both public goods and quasi-public goods (club goods). Due to P2P characteristics, the rivalry is hard to maintain. This causes a dilemma in trust. In P2P environments, the nonexcludability is accomplished on account of network resources and are typically provided to all members of the network equally. Given that a downloading user will share the content they download, the non-rivalry is accomplished because the net number of opportunities to download does not decrease for other users on the network.

However, this non-rivalry property will not hold if some network users consume network resources but do not share their content in return. This phenomenon is known as free-riding, which is a common characteristic of P2P networks. It is also a fundamental tension between individual rationality and collective welfare. This causes both over-consumption and under-provision of community resources. It seems to extrapolate to P2P networks is the inability of individually rational behavior to bring about socially optimal outcomes.

TABLE 3 KINDS OF GOODS IN DIFFERENT CHARACTERISTICS OF SUPPLY AND DEMAND

Demand Supply	Rivalry	Non-Rivalry
Excludability	Private goods	Quasi-public goods
Non-Excludability	Quasi-Private goods	Public goods

In typical economic models, individual economic actors will only consider their private utility when making decisions about consumption and provision. They will not consider the impact (a.k.a. the externality) this decision will impose on other community members. Because of this, in the absence of external incentives, the self-interested consumption of public goods may deplete the overall public utility. This is known as the "tragedy of the commons". Common examples of such "tragedies" include over-grazing by farmers using public lands and over-fishing of public waters [12]. Krishnan et al. [12] detailed the differences between resources provided over P2P networks and other typical public and club goods. Consequently, in P2P environments, the network externality has deftly solved the above problem.

Information goods have characteristics that are different from traditional goods, and they prompt us to ask new questions related to their design, their pricing and their support in the marketplace [9]. In a network economy, the supply is unlimited on account the network externalities. Besides this, the consumption of the good by one user does not reduce the utility of other users. The Internet is exposing the impossibility of sustaining a transaction-based economy. As the net drives the cost of certain goods and services toward zero, it strips profit from transactions. We have to find out some advanced P2P business models, e.g., pricing strategy, the combination of brand loyalty and network effects [1,6,17].

The significant open questions addressed here is the balance between the rights of copyright holders, entrepreneurs, and consumers. It is increasingly important for coming researches. In the category of file sharing, the P2P technique impacts the intellectual property. The appropriate context the intellectual property should protect is always needed to be carefully defined. For the category of distributed computing, we still have to deal the following issues: specialization, collaborative filtering, semantic web, privacy, and storage capacity. As the category of communication and collaboration, contents recognition schemes, payment methods, privacy and quality of service will be core issues.

3.3. *The analysis level of business process*

The firms are also able to take advantage of P2P capabilities, and make them available in the marketplace in the form of assistance for consumers' purchase of goods, and in the form of information services, as well [9]. This high-end computing power allows for real-time pricing of a variety of products (e.g., news and information access, excess inventory -- and not just stock prices), the technologies of electronic auctions (e.g., FirstAuction [28], MobShop [29] and Priceline.com [30]), comparison engines (e.g., MySimon [31], one-to-one marketing based on collaborative filtering (e.g., Net Perceptions [32]), and entirely new business models such as Yahoo [33] and eBay [34].

When used appropriately, the new technologies allow firms to streamline their business processes to achieve lower operating costs and increase sales revenue, as well as to improve channel coordination. The technologies also can beneficially impact the overall costs associated with doing business, e.g., creating a presence in the marketplace, replacing the physical infrastructure of a selling organization with a virtual infrastructure, and improving the immediacy and responsiveness of the firm while broadening its coverage in the marketplace.

More and more firms take advantage of the P2P technique to transform their business process. However, there are some emerging trust issues. In the category of file sharing, there are effect presence management, network management, free-riding, simplicity, robustness, and liability. For the category of distributed computing, it still includes collaborative filtering, mobile agents, frequency of requests, semantic web assistance, collaborative filtering, and mobile agents. As the category of communication and collaboration, in order to have a closer relationship with provider, customers, and peers, they have to deal contents, quality of service, behavior interaction, altruistic behavior, regulation, rewards, and recommendation agents.

3.4. *The analysis level of market*

The new business models for the firm and the transformed business processes associated with them influence, and, in turn, are influenced by the market context in which they exist [9]. The rapid technological changes create opportunities for disintermediation, as we have seen with the new electronic intermediaries in the travel industry, such as Microsoft's Expedia.com [35], Internet Travel Network / GetThere.com [36] and Preview Travel [37]. Meanwhile, the airlines are increasingly interested in providing producer direct electronic booking solutions, even to the point of cooperating with one another to attract booking.

Tan, Yang and Veliyath [19] mentioned guanxi, a type of particularistic trust observed in Confucian societies has mostly been viewed as a static phenomenon. Based on neo-institutionalist trust perspectives, they argued that the role of guanxi also arose from the paucity of market system trust created by the absence of well-established market institutions during China's transition from a centrally planned to a market economy. Johanson [8] noted that a fundamental element of planned economies is the idea that exchange between units and firms should be planned by authorities remote from the performance of the exchange. The institutions, where plan governance is the main mechanism, are therefore assumed to affect how trust is developed and maintained in an economy. However, when the planned economy undergoes institutional changes and plan governance erodes, trust is likely to play a different role. He suggests that both inter-unit exchanges and inter-firm exchanges contain three dimensions, which are influenced by the degree of plan governance: knowledge use, interaction, and dynamics. His study shows that plan governance in various degrees influenced the exchanges, which, in turn, gave a different level of importance to the role of trust in the economy.

The future of marketing is trust-based [6, 15]. The rapid technological changes create de-intermediation and re-intermediation in the market. The significant questions addressed here are as follows: the motivations of various actors, culture, and planned economy. User behavior is important for future researches in the category of file sharing. A third party trust provider for social interaction is urgent for the category of distributed computing. Valuable information about the trustworthiness of the participants is demanding in the category of communication and collaboration.

3.5. *The analysis level of macroeconomic*

Kauffman and Walden [9] stated most observers recognize that the impacts are occurring at the microeconomic and macroeconomic levels. With the rapid adoption and diffusion of technologies and the Internet, along with the high levels of new business capital that are being made available for firms, the macro-level effects are likely to be substantial, even if they are difficult to track and accurately measure with today's tools and approaches.

Business is not constrained by physical borders on account of information technology (IT). Globalization brought about by IT is a problem not traditionally faced by firms. Never before has a firm’s first digital storefront been accessible by customers on a global basis on its first day of operation. Much research remains to be done to understand how this pressure will shape firms in the future [8]. This creates issues that must be solved involving how to tax and regulate business across state lines and across national boarders. Meanwhile, the use of P2P technology is transforming the search characteristics of the labor market. Glaessner, Kellermann, and McNevin [7] noted developing public policy to improve or establish electronic security needs to be built. Macroeconomy level shifts occur in the demographics of employment and firm growth and value, as well as changes in the issues that government regulators must track to ensure that social welfare is maintained [9]. Digital money is made possible through encryption technologies. Current obstacles to electronic money involve adoption inertia and the characteristics it must have to be a substitute for traditional money.

In [3] demonstrate the impact of Napster phenomenon on the ethic of digital delivery. It noted the connections of P2P techniques and culture, legacy, and industry. On the moving from print economy to digital economy of information, the economies of publishing have to find the balance points. The system quality and culture significantly affect trust in the IT artifact and point to rich possibilities for future research in these areas. As we can see, there are still many questions remain to be explored. In the category of file sharing, the culture and intellectual property debate will continued. For the category of distributed computing, the social welfare is ultimately concerned. As the category of communication and collaboration, the following issues are addressed: labor market, public policy, social welfare, digital money and monetary policy, tax and regulation across national boarders.

4. Conclusion and future work

The environment under the peer-to-peer (P2P) architecture is increasingly important. In this paper, through the survey of P2P, trust and their economics, we may see that environment employed P2P computing model is undergoing. In particular, we focused on the economics of the trust in P2P environment. We also identified the related variables as shown on Table 2 for the future studies. This paper is an opening effort to explore the economic issues on trust in P2P environment.

As Table 2 shown, we analysis the economic issues on trust in P2P digital society through a comprehensive level. For the technology and product level, the issues are being researched relatively more than other levels. However, for the business process, market, and macroeconomic level, there are still many left to be studied both empirically and theoretically. In Table 2, the terms underlined mean already noticed now, while those not underlined means more changeable. This implies we have distinguished the factors which may influence the trust in P2P environments. So we may have a clear scheme for the future P2P environment. In the changeable variables, we may adjust the levels of trust. While in those not easily changed variables, showing inherent properties for P2P environment, it is most likely that we may hold the levels of trust constantly.

Current trust mechanisms mostly are based on social control philosophy, that is, two peers collaborate each other in sharing information on services. However, this kind of mechanism lacks of service semantics. A Security-By-Contract (SxC) mechanism based on security behavior and quality of service may be complementary [4]. Through SxC approach, the peers choose and trust according to service characteristics, as well as discover services according to interface and semantic description. The combination of the above two kinds of mechanism mean that trust between peers is based on both semantic and social approach [4]. In this way, we may model the evolution experiences in different collaboration stages.

From the economics perspective, we also provide some implications on the future of P2P environment such as the influence factors, killer-applications, evaluations, etc. These are stated as follows: (1) The existence of extensive free-riding on these networks has social and economic implication. New theoretical models as well as experimental and empirical data to understand user behavior are necessary. Thus we may build new models to understand the relationships between the characteristics of information goods and trust. Besides, the trust alteration between peers will be more dynamic than ever. A new approach based on a stronger (semantics-based) notion of trust is needed [4]. (2) When specifying the important costs, their influence (or even collapsing) factors and the workable incentive mechanisms, both pricing and non-pricing models should be taken into considerations. Thus we may draw a picture of what the killer applications for P2P systems. Besides, we may testify the relationships between the cost and trust. Owing to different essences than usual contracts, we depict that there should be some modifications for transaction cost theory and agent theory. (3) Developing a formal representation of trust evolution between two peers in order to validate that whether the trust will regress in function on the peers' experience as well as the third party between two trust established peers. Other user contexts such as behavior, motivation, characteristics, culture, and public key infrastructure (PKI) are undergoing areas of research. All the stakeholders have to design new distribution mechanisms for information goods. The implementation with novel approaches on the effects of trust for the above contexts is worth further understanding.

REFERENCES

[1] Casadesus-Masanell, Ramon, Andrés Hervás, and Jordan Mitchell (2006). Peer-to-Peer File Sharing and the Market for Digital Information Goods. Harvard Business School, N2-706-479.

[2] Chadwick, David W. and Andrew Basden (2001). Evaluating trust in a public key certification authority. Computer and Security. 20 (7). pp.592-611.

[3] DeVoss, D`anielle Nicole, and James E. Porter (2006). Why Napster matters to writing: Filesharing as a new ethic of digital delivery. Computers and Composition 23 (2006) 178−210.

[4] Dragoni, Nicola (2010). A Survey on Trust-Based Web Service Provision Approaches. Dependability (DEPEND), 2010 Third International Conference on , vol., no., pp.83-91, 18-25 July 2010.

[5] Gefen, David, Izak Benbasat, and Paul A. Pavlou (2008). A Research Agenda for Trust in Online Environments Journal of Management Information Systems / Spring 2008, Vol. 24, No. 4, pp. 275−286.

[6] Ghosemajumder, Shuman (2002). Advanced Peer-Based Technology Business Models A new economic framework for the digital distribution of music, film, and other intellectual property works. Massachusetts Institute of Technology Sloan School of Management.

[7] Glaessner, Thomas, Tom Kellermann, and Valerie McNevin (2002). Electronic Security: Risk Mitigation in Financial Transactions Public Policy Issues. The World Bank Financial Sector Strategy and Policy Department, POLICY RESEARCH WORKING PAPER 2870.

[8] Johanson, Martin (2008). Institutions, exchange and trust: A study of the Russian transition to a market economy. Journal of International Management 14 (2008) 46−64.

[9] Kauffman, Robert J. and Eric A. Walden (2001). Economics and Electronic Commerce: Survey and Directions for Research. International Journal of Electronic Commerce, 5 (4). pp. 5-116.

[10] Kollock, Peter (1999). The Production of Trust in Online Markets. In Advances in Group Processes (Vol. 16), edited by E. J. Lawler, M. Macy, S. Thyne, and H. A. Walker. Greenwich, CT: JAI Press. 1999.

[11] Koufaris, Marios and William Hampton-Sosa (2004). The development of initial trust in an online company by new customers. Information & Management. 41. pp. 377-397.

[12] Krishnan, Ramayya, Michael D. Smith, and Rahul Telang (2003). The Economics of Peer-to-Peer networks. JITTA : Journal of Information Technology Theory and Application. Hong Kong. Vol.5, Iss. 3, ; pg. 31, 14 pgs.

[13] Lekkas, Dimitrios (2003). Establishing and managing trust within the public key infrastructure. Computer Communications. 26, p.1815-1825.

[14] Lu, Yiling (2003). Roadmap for Tool Support for Collaborative Ontology Engineering. Master thesis of the Department of Computer Science, University of Victoria.

[15] Lyon, Fergus (2000) Trust, Networks and Norms: The Creation of Social Capital in Agricultural Economies in Ghana. World Development Vol. 28, No. 4, pp. 663-681.

[16] McKnight, D. H. V. Choudhury, and C. Kacmar (2002). The impact of initial consumer trust on intentions to transact with a web site: a trust building model. Journal of Strategic Information Systems. 11. pp.297-323.

[17] Stalnaker, Stan (2009). The Next Evolution in Economics: Rethinking Growth. Harvard Business Review. http://blogs.hbr.org/hbr/hbr-now/2009/08/a-new-approach-to-economics.html

[18] Sultan, Fareena, Glen L. Urban, Venkatesh Shankar, and I. Yakov Bart (2002). Determinants and Role of Trust in E-Business: A Large Scale Empirical Study. eBusiness Research Center Working Paper 4282-02. December 13, 2002.

[19] Tan, Justin, Jun Yang, and Rajaram Veliyath (2008). Particularistic and system trust among small and medium enterprises: A comparative study in China's transition economy. Journal of Business Venturing 24 (6), pp. 544-557.

[20] Tan, Yao-Hua and Walter Thoen (2002). Formal aspects of a generic model of trust for electronic commerce. Decision Support System. 33. pp.233-246.

[21] Tang, Zhulei, Yu (Jeferey) Hu, and Michael D. Smith (2008). Gaining Trust Through Online Privacy Protection: Self-Regulation, Mandatory Standards, or Caveat Emptor. Journal of Management Information Systems / Spring 2008, Vol. 24, No. 4, pp. 153−173.

[22] Wang, Yu-Feng, Yoshiaki Hori, and Kouichi Sakurai (2008). Characterizing economic and social properties of trust and reputation systems in P2P environment. Journal of Computer Science and Technology, 23(1), pp. 129-140.

[23] http://www.napster.co.uk/ Retrieved on 2011/01/08.

[24] http://www.kazaa.com Retrieved on 2011/01/08.

[25] http://www.limewire.com/en Retrieved on 2011/01/08.

[26] http://en.wikipedia.org/wiki/EDonkey_network Accessed on 2011/ 01/ 08.

[27] http://www.bittorrent.com/ Retrieved on 2011/01/08.

[28] http://www.firstauction.com Retrieved on 2011/01/08.

[29] http://www.mobshop.com Retrieved on 2011/01/08.

[30] http://www.priceline.com Retrieved on 2011/01/08.

[31] http://www.mysimon.com Retrieved on 2011/01/08.

[32] http://www.netperceptions.com Retrieved on 2011/01/08.

[33] http://www.yahoo.com Retrieved on 2011/01/08.

[34] http://www.ebay.com Retrieved on 2011/01/08.

[35] http://www.expedia.com Retrieved on 2011/01/08.

[36] http://www.getthere.com Retrieved on 2011/01/08.

[37] http://www.previewtravel.com Retrieved on 2011/01/08.

A Formal Methodology for Procedural Security Assessment

Komminist Weldemariam and Adolfo Villafiorita
Center For Information Technology
Fondazione Bruno Kessler
Trento 38100, Italy
Email: (sisai,adolfo)@fbk.eu

***Abstract*—Formal analysis techniques can deliver important support during ICT-based innovation (or redesign) efforts in e-government services. This paper discusses a formal methodology for assessing the procedural security of an organization. We do so by explicitly reasoning on critical information flow named assets flows. With this it is possible to understand how critical assets are modified in unlawful manner, which can trigger security and privacy violations, thereby (automatically) detecting security weaknesses within an organization under evaluation.**

***Keywords*-procedures; security assessment; modeling and analysis; formal methods.**

I. Introduction

Currently, several organizations and enterprises across many countries are evaluating and introducing ICT-based solutions with the aim of improving and delivering quality (public) services. For instance, very recently the Italian Government launched a *certified email* (in Italian "Posta Certificat@") service for its citizens [1]. This service enables citizens to legally communicate with the public administration or institutions through a certified e-mail system, with the aim of achieving a paperless bureaucracy, thereby reducing time, energy and money waste for institutions and citizens. In this setting, a significant portion of asset that can contain information and data, much of which is sensitive (e.g., the certified email account), is managed and controlled by introducing organizational regulations and procedures in order to enhance the security and privacy of (non-) digital assets. Such sensitive assets can also be used in business exchanges among (in the above scenario, e.g., citizens with PA) inter-business collaborations and (virtual) organizations with a certain understanding on the different roles the participants play; and, at the same time by including assumptions on their correct and incorrect behaviors, and their rights, duties, and obligations in order to avoid misunderstanding and ambiguities in such business relationships. Not to mention, these assets and their interrelations can also contain inherent weaknesses or vulnerabilities [2], [3], [4] and which are of two types.

The first, although out of scope, is technical vulnerabilities (for which a number of techniques exist), a hardware or software weakness, or design deficiency, that leaves a system open to attack, thereby resulting in unacceptable risk of information compromise, information alteration, or service denial [5]. The second one is procedural vulnerabilities, weaknesses within an organization due to the lack of proper implementation of security policies related to managerial or procedural deficiency, resulting in compromising the security and privacy of the organization as well as individuals within the organization [3], [6]. However, techniques that can help to model and assess such vulnerabilities are absent or very unsatisfactory, and thus procedural security analysis.

This paper complements our previous work [7] by showing how formal techniques can be used for the modeling and analysis of procedures in an organization under evaluation. We do so by presenting a formal framework for representing organization system as assets-flows. The concepts of our framework (roles and actors, actions and processes, responsibilities and constraints) can allow (business or security) analysts to capture organization model in a way that is both intuitive and mathematically formal. The use of a formal technique can allow us to determine whether a given information stipulates certain (procedural) security properties —e.g., that the responsibilities assigned to roles are fulfilled and that the constraints are maintained. With this it is possible to understand how critical assets are modified in unlawful manner within an organization. Thus, we believe that, this is important for both developed and developing nations where the development and deployment of ICT-based solutions in several areas of security-critical e-government services or applications are in progress.

The next section briefly describes the background material for procedural security analysis. Its formal model is presented in Section III. Section IV discusses the mapping of such model into executable specifications. Finally, conclusion and future work are discussed in Section V.

II. Procedural Security Analysis

A typical approach for inherent information flow within an organization is access control (see, e.g., in [8]). That is the accesses of objects by subjects, restricted by specific access permissions —e.g., by assigning *read* and *write* permissions to some sensitive assets. Moreover, approaches such as based on formal techniques and methodologies have been used to model (system) processes [9], [10], [11], [12]. These works mainly concentrate on constructing business process

Copyright (c) IARIA, 2011. ISBN: 978-1-61208-116-8

models with correct creation and termination of artifacts during their lifecycle, by providing some supports to perform automated analysis. Accordingly, some of these approaches hint integrations with existing formal methods' tools. However, they hardly concentrate on security analysis especially on analyzing the security of organizational or procedural weaknesses. As noted in [13], [14], risks and attacks not only depend upon the security levels the new systems offer, but also occur by circumventing on the procedures and controls regulating the way in which the systems are operated. For example, what happen if one can get a fake certified email by circumventing the procedures required for request and delivery of such services. Obviously, this could lead to maliciously communicate with the PA thereby accessing public services accordingly. Therefore, it is important to analyze the security of such procedures that grant accesses to sensitive data and services, and thus procedural security analysis.

To be able to conduct a security analysis, at least enough information must be present to deal with assets and global threats, i.e. at an abstract level, subjects and objects in the procedures and system must be identifiable. The starting point of the procedural security analysis methodology is an initial model, describing a coarse procedure or system process without security-related aspects. This model describe the procedure or procedures to be analyzed in a systematic way. Secondly, we extend this model with attack information, meaning that we generate an extended model from the model defined in the previous step. In the extended model, thus, not only assets are modified according to what the procedures define but they can also be transformed by the (random) execution of one or more threat actions. Thirdly, the encoding of the asset-flows in terms of executable specifications is performed using formal language. Fourth, we specify security properties for formal analysis. More specifically, we specify the (un-)desired (procedural) security properties —namely, the security goals that have to be satisfied (unsatisfied), are then encoded using mathematical formula, which in turn together with the model are given as input to the analysis tool. Thus, we perform security verification and assess the results. Security verification is the verification that the global security requirements are fulfilled with respect to the threat scenario. If the result of the security verification is that a particular security requirement is violated, there is a corresponding attack on the procedures and consequently on the system. Otherwise, the procedure is secure given the assumptions included in the model. This is obviously via the model checker, i.e., if a property is proved to be false, the analysis tool generates a counterexample which opens up further discussion.

III. A Formal Model of Procedural Security

Figure 1 shows a high-level representation of the information and the behavioral (i.e., the lifecycle) models of assets. The perspective shown in the figure offers three complementary views: workflow, assets class, and state machine diagram views. In the *workflow diagram* view, workflow activity sequences are defined. The *state machine* view describes the behavior of an asset in terms of a transition system in which transitions are enabled due to explicit execution of workflow activities. The activities in the workflow are transformation functions that influence the behaviors of the assets. A finite state transition diagram for each feature of an asset constitutes the global state machine for that asset.

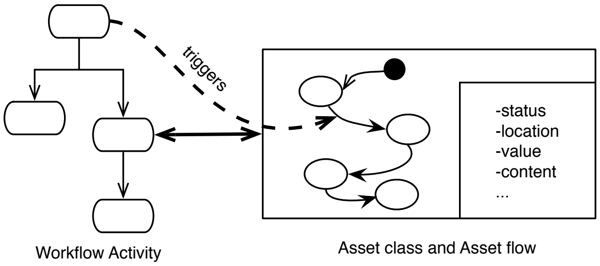


Figure 1. An asset-flow view of a business process model

For instance, Figure 2 shows a simple example of asset-flow model for asset instance $\underline{\mathbf{A}}$ with three states $[s_1]$, $[s_2]$, and $[s_3]$. The corresponding finite state machine, therefore, will possibly have three sequential states each of which corresponds to $\underline{\mathbf{A}}$'s current features values.

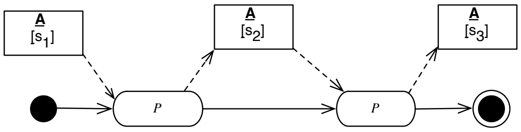


Figure 2. Example of a single instance Asset-flow model in three states.

A. Formalization of the Model Elements

We assume the following notations and their definitions:

- $\mathcal{T}_p$ be a set of primitive types, such as bounded integer and boolean;
- $\mathcal{C}$ be a set of asset classes (*names*);
- $\mathcal{A}$ be a set of attributes (*names*);
- $\mathcal{ID}_C$ be a set of identifiers that describes the identifiers for each asset class $C \in \mathcal{C}$;
- $\mathcal{S}$ is a set of assets states, where each $s \in \mathcal{S}$ is a sort of truth assignment over the variables values.

Note that all the above sets are finite, which is essential for the formal verification process such as by using model checkers.

A *type* $\mathcal{T}$ is an element of the primitive types $\mathcal{T}_p$ and the class identifiers C; namely, $\mathcal{T} = \mathcal{T}_p \cup \mathcal{C}$ (we assume that $\mathcal{T}_p$ and $\mathcal{C}$ are disjoint).

Definition *3.1:* An asset class signature is a triple $\langle C, A, \psi \rangle$ where $C \in \mathcal{C}$, $A \in \mathcal{A}$ is a a set of attributes for the asset class C, and $\psi : \mathcal{A} \rightarrow \mathcal{T}$ is a total function that maps each attribute of an asset into its corresponding type.

Without loss of generality, we assume a fixed interpretation domain associated to each $\mathcal{T}$ type. That is the *domain* of each type $t \in \mathcal{T}$, denoted $\mathcal{D}^t$, is defined in the following way: if $t \in \mathcal{T}_p$ is a primitive type, then the domain $\mathcal{D}^t$ is some known set of $values$ of type (e.g., integer or boolean); if $t \in \mathcal{C}$ an identifier type, then $\mathcal{D}^t$ defines existing instances of an asset class identifier for t (i.e., $\mathcal{D}^t = \mathcal{ID}_t$). We require all variables must have their corresponding *values* all along their life. For undefined location and unassigned content of an asset, we use an *undefined* and a *null* constant values respectively. The interpretation is that the location is not known and the content value is not either assigned yet or reset to contain null.

Definition *3.2:* An asset instance is a triple $\langle ID_C, C, \phi \rangle$, where $ID_C \in \mathcal{ID}$ is a class identifier and ϕ a partial function, given an instance of a class C, that assigns each variable $a \in A_C$ of type $t \in \mathcal{T}$ a value in D^t (i.e., $\phi(a) = D^t(\psi(a))$).

An asset can have multiple instances. We denote the set of asset instances by $\vec{\mathcal{O}}_{C, C \in \mathcal{C}}$ and $\vec{\mathcal{O}}$ for all instances over Σ. However, in this work, we mainly focus on how a single asset instance $I \in \vec{\mathcal{O}}_C$ can evolve from some initial state through other states. The set of *variable-value* pairs for I defines the *state* of a given asset. The *state* of an asset is, therefore, the current situation called *"snapshot"* of an asset instance I and its value is the truth assignment over the variables. An asset is *initial*, if all the variables are in their initial state and ϕ is undefined for some attributes and *final*, if all the variables values do not change anymore.

The information provided previously is the basis for the formalization, namely how we represent the assets structure and workflows to arrive at what we call executable models. The formalization allows the model to be more amenable to formal analysis, hence it shifts the focus to dynamic aspects of the assets.

1) Defining Workflow formally: As noted earlier, the initial values for the variables of an asset can be assigned at the time of the instance creation or otherwise assigned by an analyst. However, only due to the execution of a workflow activity over these variables can possibly change the initial configuration of the asset.

Roughly speaking, a workflow activity is described by input assets, preconditions, and effects of the activity over the assets (a similar interpretation can be found in [9], [11], [15]). The effect of a workflow activity is regarded as a change in state of the input assets. Not all assets change their states thought, since it is not always the case that an execution of a workflow activity enables state transition to all the input assets (e.g., reading the content of a password does not change its state).

For each executable workflow activity we specify which actors participate in the workflow with predefined privileges or responsibilities or both. These information not only allow to describe who does what during the execution of an activity, but, more importantly in the context of organizational security modeling and analysis, who manages what data and with what privileges. Such information are static, namely they are known before executing a workflow (e.g., as described in a legal document or contractual agreement between two entities) and are encoded in our model to describe a workflow scenario. We, therefore, use these information along with the activities to describe a workflow model as a deterministic finite state machine in which the states are constructed by a set of activities, and the transitions are described by the current state and a matching condition over the accessory information.

Formally, we define the workflow model as follows.

Definition *3.3 ($\mathcal{W}$):* A workflow model is a quadruple $\langle P, s_0, s_f, C, \Delta \rangle$ where

- P is a set of activities or processes (*names*);
- $s_0, s_f \in P$ are initial and final activities of the workflow respectively;
- C is guard expression over accessory information, and;
- $\Delta \subseteq P \times C \times P$ is a transition relation between a current activity and its successor activities in which a transition is labelled with a condition over accessory information.

The above definition is meant to express the fact that there exist a set of activities within a particular workflow, that describes a procedure under analysis, in which by knowing the current state of the workflow, and if a condition is met, it should be obvious to determine the next state of the workflow. We call an instance of a workflow model, a program counter *"pc"* that contains the value of the current state (i.e., the active activity) in the workflow. There is one program counter *"pc"* for each workflow model at run time. In actual business process or workflow specification, in fact, it is possible to have multiple activities that can run in synchronous or asynchronous mode. We focus on sequential execution of a workflow in this work.

2) Defining Asset-flows formally: The state of an asset is specified by the assignments of values to variables (or simply valuations), which allows to describe the evolution of an asset. The evolution is expressed by the sequence of states through which an asset undergoes during the execution of a process. Since the state of an asset is described by the valuations over its variables, therefore, it makes sense to encode the state of each variable as a finite state machine. The workflow instances, along with some matching conditions, define transitions for modeling the lifecycle of the assets. Thus, an asset-flow can easily be modeled using a transition system that facilitate formal analysis.

Definition *3.4:* An asset-flow model (AFM) is a 5-tuple $\langle AS, I, \mathcal{W}_\pi, C_\pi, \Delta_\pi \rangle$ where

- $AS \in \mathcal{S}$ is a finite set of assets' (instances) states;
- $I \subseteq AS$ is an initial states of the assets;
- $\mathcal{W}_\pi$ is a set of workflow instances;
- C_π is a set of conditions constructed over the attributes representing the matching construct as a guard, that specify the condition must meet for the state to be changed, along with the current activity;
- $\Delta_\pi \subseteq AS \times \mathcal{W}_\pi \times C_\pi \times AS$ is a transition relation between a current state of an asset and its successor states in which a transition is labelled with an activity and a condition.

A collection of individual AFM constitutes *assets-flow* models, and we represent it by $\mathcal{M}$. Therefore, $\mathcal{M}$ is regarded as the global configuration of the domain of interest, namely the procedures under analysis. The semantic of the global configuration $\mathcal{M}$ can be interpreted in the following way. Each $m \in \mathcal{M}$ is regarded as an abstract state machine, which has three major components: a workflow activity sequence (possibly maintained in a queue), a workflow activity dispatcher, and an activity processor. Workflow activities are added to the end of the activity queue. The activity dispatcher chooses, dequeues, and provides the next *"pc"* (i.e., an activity) to the activity processor. Each *"pc"* is then used as a *transformation* function that can possibly change the state of an asset by modifying or changing one or more variables values of the asset. One state machine per feature variable encodes the lifecycle of that state variable. A set of such state machines constitutes the global state machine for the corresponding asset instance. By defining a semantic for the state machines corresponding to each feature of an asset and linking it with $m \in \mathcal{M}$, therefore, we have implicitly defined how $\mathcal{M}$ behaves.

B. Model Extension

In order to analyze what are the possible attacks of a given (set of) procedures, we need to encode asset threats in the nominal model and generate the extended model for $\mathcal{M}$. Structurally, in fact, there is no difference between $\mathcal{M}$ and the extended model. However, the main difference lies on the assets state *set* and on the transitions specification. This means that, the extended model possibly will have more states than the other due to the execution of threat-actions that can change the state of an asset into an undesired one. On the transitions side, on the other hand, the definition did specify the fact that transitions are triggered only by nominal workflow activities. We need to incorporate in the extended model the fact that an asset could be in any possible states and that such states can also be changed by the execution of malicious processes.

However, it is pretty straightforward from the definition we gave and by extending the definition of the workflow model to include all the malicious processes that an adversary might execute. Thus, in extended model, assets are not only manipulated according to what should happen in the nominal case, but can also be transformed by the execution of one or more assets threat-actions.

IV. Encoding using Formal Language

Assets-flow models $\mathcal{M}$ can become executable specification to allow formal analysis through verification tools on their evolution, including their malicious evolution due to threat-actions. Our aim here is to represent the model $\mathcal{M}$ into executable specification using NuSMV input language [16]. As noted in [16], the NuSMV semantic is based on a state-based formalism in which the behavior is defined by Kripke transition systems. However, the definition we gave for $\mathcal{M}$ is an action-based formalism in which the behavior is defined by (a sort of) labelled transition systems. Thus, we need to rearrange the previous definition to align with the semantic of Kripke structure so that the encoding of NuSMV specifications can be tackled.

Definition *4.1:* Let APs are set of atomic propositions ranged over some boolean expressions on the valuations of the variables. An asset flow model (AFM) is a Kripke structure over a set of atomic propositions AP defined by a quadruple $\langle AS_K, I_K, \Delta_K, \mathcal{L}_K \rangle$ where

- AS_K is a finite set of assets (instances) states;
- $I_K \subseteq AS_K$ is set of initial states;
- $\Delta_K \subseteq AS_K \times AS_K$ is a transition relation between a current state of an asset and its successor states;
- $\mathcal{L}_K : AS_K \rightarrow 2^{AP}$ is the labeling function which returns the set of atomic propositions which hold in a state.

The encoding of $\mathcal{M}$ in the NuSMV input language can be treated as a problem of defining a mapping between the two structures, i.e., between the structure specifying the model $\mathcal{M}$ and the Kripke structure. More specifically, the following encoding rules are defined to map $\mathcal{M}$ into the NuSMV counterpart.

Rule 1: The workflow model is encoded in NuSMV as a special module, and each workflow activity $p_i \in P$ for $i = 1, \ldots, n$ representing the domain activities (i.e., processes) in $\mathcal{W}$ are encoded in the NuSMV input language as a scalar variable program counter (pc) in which p_i are its symbolic values.

In order to determine the state transition of the program counter, we introduce some predicates (see Table I). They are mainly associated with the accessary information, such as actor-role and actor-activity assignments. The table also shows the corresponding state variables in NuSMV input language.

Rule 2: The accessary information are encoded in the NuSMV input language within the $Workflow$ module in the following way (see also Table I):

- For each actor-role assignment, we introduce a variable assign_a_r. assign_a_r is true iff the predicate AssignR(a,r) is true for an actor $a \in Actor$ and a role $r \in Role$;

Table I
ACCESSORY INFORMATION AS PREDICATES.

Predicate	Meaning	NuSMV variable
AssignR(a,r)	assignment of actor $a \in Actor$ to role $r \in Role$	assign_a_r
AssignA(a,p)	assignment of actor $a \in Actor$ to an activity $p \in P$	assign_a_p
r_Active_for_a	role $r \in Role$ is active for actor $a \in A$	activefor_a_r
ExecA(a,p)	actor $a \in A$ executes an activity $p \in P$	exec_a_p

- For each actor-process assignment, we introduce a variable assign_a_p. assign_a_p is true iff the predicate AssignA(a,r) is true for an actor $a \in Actor$ and an activity $p \in P$;
- For each role activation r_Active_for_a, we define a state variable Activefor_a_r;
- Similarly, we define a variable Exec_a_p for every actor performing an activity, i.e., iff ExecA(a,p) is true.

Rule 2 defines accessary information for the transition relation of `pc` state variable. Notice that activities can only be executed if the activity instance in question is assigned to an actor —i.e., ExecA(a,p) $\Rightarrow$ AssignA(a,p). Moreover, a group of actors can perform the same activity, as discussed in the previous chapter.

Rule 3: For each asset instance in $\vec{\mathcal{O}}$, a NuSMV module is defined:

```
MODULE ASSET_NAME  (...)
```

Rule 4: An asset with no content in $\mathcal{M}_1$ is mapped to a symbolic value *"null"* in NuSMV. Similarly, an asset whose current location is not known or unspecified in $\mathcal{M}_1$ is mapped to a symbolic value *"unspecified"* in NuSMV.

Rule 5: The $location$, representing all the possible places of an asset, is encoded in the NuSMV input language as scalar variables `loc` in which loc_i for $i = 1, \ldots, n$ and "$undefined''$ are its symbolic values. The $content$, representing all the contents of an asset at a particular point of time, is encoded in the NuSMV input language as `content` in which $content_i$ for $i = 1, \ldots, n$ and "$null''$ are its symbolic values. The $value$, representing all security risk values for an asset, is encoded in NuSMV input language as `value` in which $noValue$, low, $high$ and $critical$ are its symbolic values. Finally, each domain specific `property` of an asset in an asset-flow model is encoded as a boolean value in NuSMV.

Rule 3 states that a module is defined for each asset (instance) in $\mathcal{M}$. In Rule 5, whereas each feature of the asset is defined as a state variable within the asset module specification. An *unknown* `location` and a *null* `value` are both encoded by symbolic values as defined by Rule 4.

Rule 6: The transition specification for each state variable is encoded by the current value of the program counter and some boolean expressions over the current state of the asset.

The above rule (i.e., Rule 6) encodes the transition specifications. The transition from one asset state to the next is determined by the current value of the `pc` and some condition over the current state of the asset instance.

Since all the above rules are related to the encoding of $\mathcal{M}$, we need to provide additional rule for encoding the extended model. The model extension corresponds to proving an extension in the NuSMV model with one or more applicable attack-actions. That is, a specification of how the assets can be in undesired states. This can be done by associating threat-actions with variables defined inside the module per asset instance. Moreover, the `Workflow` module should also need to be extended in order to include the malicious process executions. In particular, the model extension can be done by using the following strategies:

- by defining a scalar state variable to encode all the possible malicious process within the `Workflow` module. Therefore, the program counter does not only have values from nominal workflow activities but also from possible set of malicious workflow activities;
- by defining a transition specification for each activation of a threat-action on asset instance under the corresponding asset module in NuSMV, where the malicious activity is in place for enabling the transition;
- by defining boolean variable to monitor the execution of the corresponding threat-action. This variable will be true iff when the corresponding threat-action takes place;
- by introducing a scalar value "garbage" for the content state variable related to the introduction of malicious asset and a boolean variable. This variable will be true iff a predicate associated with the action (e.g., MAsset(t)) is true by a threat-action, say t.

The above strategies facilitate the task of model extension, by adding a number of boolean appendage variables that are needed to capture the malicious asset flows and the execution of threat actions to form the extended model specification in NuSMV input language. In this way, therefore, the model extension is performed for each applicable threat-action against the normal flow of assets. At this point, we have the model representing the assets-flows, which is ready to supply for the analysis tool. Before the analysis, however, we need to describe the security properties we intend to check against the model using standard LTL/CTL logic. Once all the properties of interest are specified with respect to the analysis goals, the next activities are formal verification and analysis of the results. Security verification is the verification that the global security requirements are fulfilled with respect to the threat scenario. If the result of the security verification is that a particular security requirement is violated, there is a corresponding attack on the procedures and consequently on the system. Otherwise, the procedure is secure given the assumptions included in the model. This is obviously via the model checker, i.e., if a property is proved to be false, the analysis tool generates a counterexample which opens up further discussion.

V. Conclusion and Future Work

We have described a framework where procedurally rich systems in model-based assessment drive the construction of models by extending the usage scenario of procedural security analysis. The presented approach is aimed at basic security aspects relevant for organizational-oriented processes, providing guidelines to analyst performing procedural security analysis based on explicit reasoning on assets-flows. The approach can be used to analyze and evaluate the impact of threats, and consequently to come out with a set of (security) procedural requirements. Thus, an organization or enterprise can apply the approach to assess their procedural security posture prior to introduce ICT-based solutions. Here, assurance is not implied by the trust in the model but follows from the formal analysis of the model. The analysis is based on a set of formal security requirements and provides formal proofs for use as countermeasures (i.e., evidence).

The work described here is still in progress, and we are currently completing the theoretical framework of the approach. We admit that this work clearly lacks a working case study, illustrating a proof of concept of the presented approach. Moreover, the implementation of the approach in terms of a tool is not discussed. However, we are currently defining a model-based verification approach using UML. More specifically, we reuse existing formal semantics for UML activity diagrams specifying workflow models that correspond to the asset-flows semantics discussed in this paper. The semantics will be translated to NuSMV model based on meta-model transformations. To translate a UML activity diagram model into the NuSMV counterpart, we use an intermediate model called activity hyper-edge model which abstracts the activity diagram, specifically according to the semantics of the asset-flow model. The definition of a set of generic library of attack models corresponding to threat-actions is part of the future work.

References

[1] Italian Ministry of Public Affairs and Innovation, "It: Launch of a certified email system to communicate with the public administration," http://www.innovazionepa.gov.it/lazione-del-ministro/iniziative-e-sperimentazioni/sperimentazione-pec/pec-primo-piano.aspx, October 2010.

[2] Federal Agency for Security in Information Technology, "IT Baseline Protection Manual," http://www.iwar.org.uk/comsec/resources/standards/germany/itbpm.pdf., October 2000. Last accessed on Feb. 2011.

[3] Common Criteria, "Common Criteria for Information Technology Security Evaluation," http://www.commoncriteriaportal.org/, 2007. Last accessed on Nov. 2010.

[4] S. E. Parkin, A. van Moorsel, and R. Coles, "An Information Security Ontology Incorporating Human-Behavioural Implications," in *SIN '09*. ACM, 2009, pp. 46–55.

[5] M. Bishop, *Computer Security Art and Science*. Addison-Wesley Longman Publishing Co., Inc., 2002.

[6] B. S. Institution, "BS ISO/IEC 27002:2005 — Information Technology —Security Techniques —Code of Practice for Information Security Management," http://www.iso27001security.com/html/27002.html, 2005. Last accessed on Feb. 2010.

[7] K. Weldemariam and A. Villafiorita, "Formal Procedural Security Modeling and Analysis," in *Proceedings the International Conference on Risks and Security of Internet and Systems*, ser. CRiSiS '08. IEEE, 2008, pp. 249–254.

[8] Ravi S. Sandhu and Edward J. Coyne and Hal L. Feinstein and Charles E. Youman, "Role-Based Access Control Models," *IEEE Computer*, vol. 29, no. 2, pp. 38–47, 1996.

[9] M. Koubarakis and D. Plexousakis, "A Formal Model for Business Process Modeling and Design," in *CAiSE*, ser. LNCS, Benkt Wangler and Lars Bergman, Ed. Springer, 2000, pp. 142–156.

[10] R. Eshuis, "Symbolic Model Checking of UML Activity Diagrams," *ACM Trans. Softw. Eng. Methodol.*, vol. 15, no. 1, pp. 1–38, 2006.

[11] K. Bhattacharya, C. E. Gerede, R. Hull, R. Liu, and J. Su, "Towards Formal Analysis of Artifact-Centric Business Process Models," in *Proceedings of the 5th international conference on Business process management*, ser. BPM '07, vol. 4714. Springer, 2007, pp. 288–304.

[12] A. Deutsch, R. Hull, F. Patrizi, and V. Vianu, "Automatic Verification of Data-Centric Business Processes," in *ICDT '09*. ACM, 2009, pp. 252–267.

[13] A. Xenakis and A. Macintosh, "Procedural Security Analysis of Electronic Voting," in *Proceedings of the 6th international conference on Electronic commerce*, ser. ICEC '04. ACM Press, 2004, pp. 541–546.

[14] K. Weldemariam, "Using Formal Methods for Building More Secure and Reliable e-voting Systems," Ph.D. dissertation, University of Trento, Via Sommarive 14, March 2010.

[15] C. E. Gerede and J. Su, "Specification and Verification of Artifact Behaviors in Business Process Models," in *ICSOC*, ser. LNCS, vol. 4749. Springer, 2007, pp. 181–192.

[16] A. Cimatti, E. Clarke, E. Giunchiglia, F. Giunchiglia, M. Pistore, M. Roveri, R. Sebastiani, and A. Tacchella, "NuSMV 2: An Open Source Tool for Symbolic Model Checking," in *CAV'02*, ser. LNCS. Springer, January 2002, pp. 241–268.

A Framework for Semantic Model Ontologies Generation for E-government Applications

Jean Vincent Fonou Dombeu[1,2]
[1]Department of Software Studies
Vaal University of Technology
Vanderbijlpark, South Africa
Email: fonoudombeu@gmail.com

Magda Huisman[2]
[2]School of Computer, Statistical
and Mathematical Sciences
North-West University
Potchefstroom, South Africa
Email: Magda.Huisman@nwu.ac.za

Zygmunt Szpak[3]
[3]School of Computer Science
The University of Adelaide
Adelaide, Australia
Email: zygmunt.szpak@adelaide.edu.au

***Abstract*—The Web Ontology Language (OWL) standard is increasingly being used to build e-government service ontologies that are integrable and interoperable in e-government environments. However, current works employing OWL ontologies in e-government are more directed to the Semantic Web audience than to the broader e-government community. Furthermore, only a few of these works provide detailed guidelines for constructing OWL ontologies from a business domain. This paper presents a framework for generating semantic model ontologies in OWL syntax from a government service domain. Firstly, the government service domain is analyzed and a domain ontology is constructed to capture its semantic content. Thereafter, a semi-formal representation of the domain ontology is created with the ontology knowledge-base editor Protégé. Finally, the OWL ontology model is imported. This study aims at providing e-government developers, particularly those from the developing world, with an easy to use framework for practicing semantic knowledge representation in e-government processes; thus facilitating the design of e-government systems that can be easily integrated and maintained.**

Keywords - E-government; Interoperability; Ontology; OWL; Protégé; Software Engineering.

I. Introduction

In recent years, many countries worldwide have adopted e-governance, resulting in several applications being developed in various government departments and agencies. The increasing number of autonomous e-government applications has raised several software engineering issues as reusability, maintenance, integration, and interoperability of these applications [1][2][3][4], in the context of one-stop e-government which requires e-government applications to be accessed at a single point and function as a whole for better efficiency [1][5]. In an attempt to address the above issues, semantic model ontologies using the OWL Web Service Standard are frequently used. OWL ontologies allow the composition [7][8], searching, matching, mapping and merging [9][10] of e-services and facilitate their integration [5][8][9], maintenance [8][9] and interoperability [3][7][10][11].

Many works describe ontology modelling and implementation activities in e-government [6][3][7][10][11]. These works demonstrate that OWL is a common language employed for semantic knowledge representation in e-government. However, in this research, we argue that the above works are more directed to the Semantic Web audience than to the broader e-government community. Furthermore, only a few of these works provide detailed guidelines for constructing OWL ontologies from an e-government service domain. This paper presents a framework for generating semantic model ontologies in OWL syntax, from a government service domain. Firstly, the government service domain is analyzed and its domain ontology is constructed. Thereafter, a semi-formal representation of the domain ontology is created and implemented with the ontology knowledge base editor Protégé. Finally, the OWL ontology model is imported. The study aims at providing e-government developers, particularly those from the developing world, with an easy to use framework for practicing semantic knowledge representation in e-government processes; which allow building e-government systems that can be easily integrated and maintained.

The rest of the paper is organized as follows. Section 2 defines ontology and gives its roles in the software engineering field. Ontology modelling and implementation activities in e-government are reviewed in Section 3. The languages and software tools for representing and editing ontologies are presented in Section 4. Section 5 presents the framework for OWL ontology generation. A case study application of the framework is conducted in Section 6. Section 7 carries out a discussion and a conclusion is drawn in the last section.

II. Definition and Roles of Ontology

There are several definitions of ontology in the literature [12]; the most commonly used definition is taken from Gruber [13]. He defined an ontology as an explicit specification of a conceptualization. A conceptualization refers to an abstract and simplified view of a domain of knowledge one wishes to represent for a certain purpose. The domain could be explicitly and formally represented using existing objects, concepts, entities and the relationships that exist between them [13]. The

ISBN: 978-1-61208-116-8

domain could refer to a domain such as medicine, geographic information system, or e-government; it could also refer to an area of problem solving or a knowledge representation language [14]. Ontologies are widely used in disciplines such as software engineering, databases, artificial intelligence, and many more [15]. In these fields, developers use ontologies to represent knowledge in a manner that can be automatically processed by a machine. In [16] and [17] the authors argued that because an ontology represents the concepts of a domain of knowledge and the relationships between them, it provides a shared and common understanding of the structure of information among people and software agents. It also facilitates software development and improves processes in the corresponding domain. Aside from the semantic representation of concepts of a domain of knowledge, an ontology also provides a data type description which specifies the data component of applications [19]. Ontologies are application independent, which allow domain knowledge reuse and easy software maintenance, and contribute to the semantic interoperability of applications [13]. Due to the complexity of government processes various government departments need ontologies to streamline, re-organize government services and to facilitate the integration, maintenance and interoperability of their e-government systems [19][20]. Some works illustrating the current practice of using ontologies in e-government systems are provided in the next section.

III. Use of Ontology in E-government

Salhofer et al. [6] presented an ontological approach for service integration in e-government. A semantic objective and service discovery technique was used to illustrate how e-services could be derived from citizens' needs expressed in the form of simple phrases. The derived e-service ontologies were represented in OWL and the Web Service Modelling Language (WSML). Another ontological approach for semantic interoperability in e-government was proposed by Muthaiyah and Kershberg in [3]. They used a shared hierarchal ontology in which knowledge is organized at different levels with local ontologies. A semantic bridging process methodology was described for the mapping, merging and integration of local ontologies represented in an OWL syntax. In [7], an intelligent platform to host e-government services in the form of a customer-oriented e-government Web portal was put forward. To facilitate services and related public administrations interoperability they introduced the concept of an intelligent document and a Life Event service both of which are semantically modelled with OWL ontology. These allow automatic services composition, advanced searching mechanisms and better usability from the user's point of view. In [8] and [9] the authors presented a software engineering platform for the development and management of e-government services namely ONTOGOV. The ONTOGOV platform uses Semantic Web technologies including OWL-S and Web Service Modelling Ontology (WSMO) to construct eight types of ontologies characterizing the e-government domain; they include: legal ontology, organizational ontology, life-cycle ontology, domain ontology, service ontology, life-event ontology, profile ontology, web service orchestration ontology. These ontologies aim at describing and composing services provided by public administrators. In particular, the life-cycle ontology is used to carry out the maintenance of e-services and the web service orchestration ontology is used for software components and service ontology integration [9]. A multilevel abstraction of life-events for e-government services integration was presented in [10]. In their work, a life-event is defined as a collection of actions needed to deliver a public service satisfying the needs of a citizen in a real-life situation and is modelled using three kinds of ontologies: e-government ontology, regulatory ontology and service ontology. The ontologies are represented in OWL to enable dynamic services integration through semantic searching and matching of concepts [10]. Xiao et al. [11] present yet another ontology-based approach for semantic interoperability in e-government. They describe the business process of e-government services using an E-government Business Ontology (EG-BOnt). Each business process is described in terms of its input, output, resource constraints and logical relations with other relevant businesses. Thereafter, each class of the EG-BOnt is defined using the OWL language for its strong semantic and logic relation expressiveness [11]. Finally, an architecture describing a semantic interoperability framework between different government systems based on the proposed EG-BOnt was presented.

IV. Ontology Representation Languages

The Semantic Web domain provides various languages for representing ontologies including XML, RDF, DAML, and OWL [21]. OWL is the most widely used of these languages because of its high expressive power and the fact that it is the W3C standard ontology language for the Semantic Web [24][26]. Several software tools are used for ontology edition including WebODE, OntoEdit, KAONI, and Protégé [18]. Ontology developers prefer Protégé for its ease of use and its abstraction capabilities; it has a graphical user interface which enables ontology developers to concentrate on conceptual modelling without any knowledge of the syntax of the output language [24]. Furthermore, Protégé is open-source software which is downloadable from the Stanford Medical Informatics website. This paper gives a step-by-step guideline on how e-government developers can design and generate OWL ontologies using Protégé. The next section presents the proposed framework for constructing OWL ontologies from an e-government service domain using Protégé.

V. Framework for OWL Ontologies Generation

The framework starts with an e-government service domain as an input. Domain experts and different information sources are consulted to describe the business process of the domain. A domain ontology is then built to capture the relevant concepts, activities, tasks, regulations and relationships between all the constituents of the e-government service domain. Thereafter, a semi-formal representation of the domain ontology is constructed in the form of a class diagram in UML syntax; this

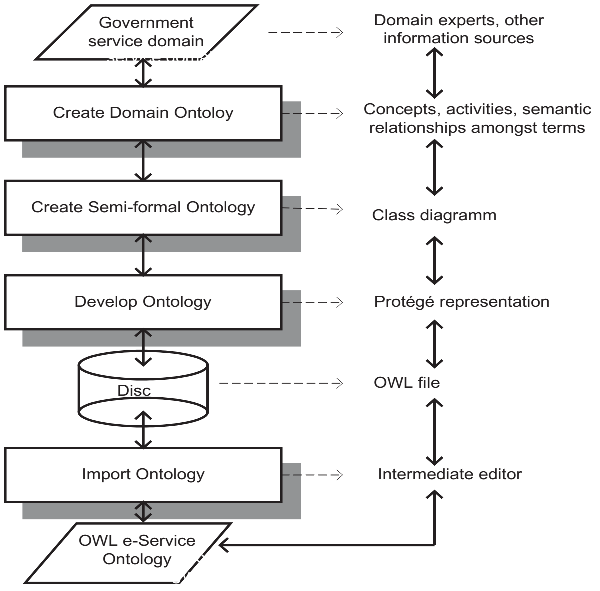

Government service domain
Domain experts, other information sources
Create Domain Ontoloy
Concepts, activities, semantic relationships amongst terms
Create Semi-formal Ontology
Class diagramm
Develop Ontology
Protégé representation
Disc
OWL file
Import Ontology
Intermediate editor
OWL e-Service Ontology

Fig. 1. Framework for OWL Ontology Generation

is done by identifying entities and instances in the domain ontology and categorizing relationships between entities (association, composition, inheritance). The semi-formal version of the ontology is created with Protégé and saved onto the disc. Finally, appropriate software is used to import the OWL version of the ontology from the file. To fulfil the aim of this paper which is to provide e-government developers with a step-by-step guideline for generating semantic model OWL ontologies from e-government service domains, a real-life case study illustrating the steps of the framework provided in Figure 1 is conducted in the next section; each subsection corresponds to a step of the framework in Figure 1 from top to bottom.

VI. Case Study

A. E-government Service Domain

The case study used in this paper was motivated by the fact that, in developing countries and in Sub Saharan Africa (SSA) in particular, almost every government department is somehow involved in the implementation of a programme aiming at improving the welfare of people. These programmes are commonly called development projects and include infrastructure development, water supply and sanitation, education, rural development, health care, ICT infrastructure development and so forth. Thus, we thought that an e-government web application that could interface all the activities related to development projects implementation in a SSA country could bring tremendous advantages; particularly, such a web application would improve the monitoring and evaluation of projects and provide transparency, efficiency and better delivery to populations. In [22], we have proposed an ontology support model for such a web-based e-government application. We evaluated case studies of development projects implementation, consulted domain experts including municipalities and non-governmental organizations employees and academic members, and reviewed publications in related fields including project management, project monitoring and evaluation, and capacity building [22]. Thus, a conceptual/domain ontology of development projects monitoring (OntoDPM) in a developing country was developed [22]. The next section presents the OntoDPM.

B. Create Domain Ontology

The ontology engineering field has established various kinds of ontologies; an exhaustive list of these ontologies could be found in [18]. One of the most commonly used of these ontologies is the conceptual/domain ontology. A domain ontology characterizes domains such as medicine, geology, e-government, and so on; it provides vocabularies about the objects and concepts within a domain and their relationships, the activities that take place in that domain, and theories and elementary principles governing the domain [12].

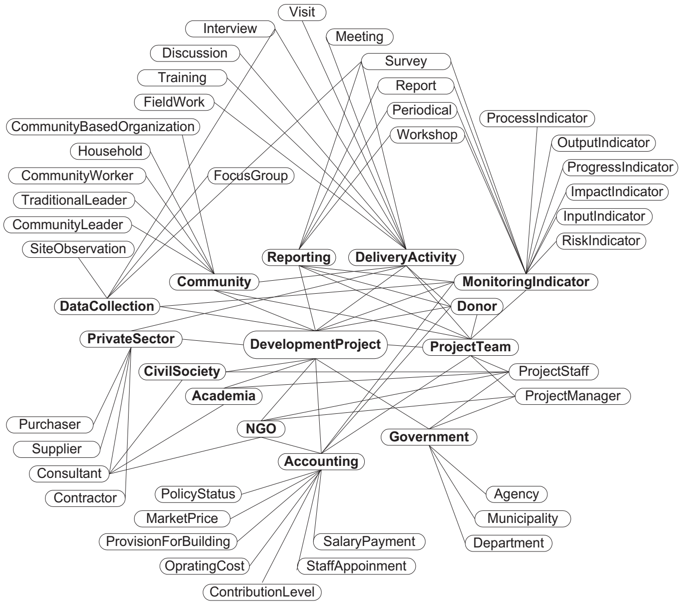


Fig. 2. OntoDPM Domain Ontology

In [22], we used a five step framework adopted from the Uschold and King [14] ontology modelling approach to represent the OntoDPM domain ontology as in Figure 2. The OntoDPM shows the key concepts of the domain (people, stakeholder, financier, monitoring indicator, reporting technique, etc.), the activities carried out in the domain (training, discussion, fieldwork, visit, meeting, etc.) and the relationships between the constituents of the domain. The semi-formal representation of the OntoDPM is provided in the next section.

C. Create the Semi-formal Ontology

Based on the OntoDPM in Figure 2 we designed the class diagram of the ontology. The classes, inheritance structure and the class instances are provided in Table 1. The classes in Table 1 were constructed by identifying entities and instances in the OntoDPM and categorizing relationships between entities

TABLE I
DESIGN DETAILS OF THE ONTODPM CLASS DIAGRAM

Classes	
Development project, monitoring indicator, delivery activity, reporting technique, person, stakeholder, financier, community leader, traditional leader, community worker, project staff, consultant, academic institution, community based organization, civil society, private company, government, donor, non-governmental organization, agency , municipality, department, accounting activity, data collection technique	
Inheritance Structure	
Super Class	**Sub Classes**
Person	project staff, community leader, community worker, traditional leader
Financier	government, donor, non-governmental organization
Stakeholder	academic institution, civil society, private company, community based organization
Government	department, municipality, agency
Class Instances	
Class	**Instances/Individuals**
Monitoring indicator	input indicator, output indicator, impact indicator, risk indicator, process indicator, progress indicator
Delivery activity	survey, meeting, visit, discussion, training, fieldwork, interview
Reporting technique	workshop, written report, periodical, survey
Accounting activity	operating cost, salary payment, contribution level, provision for building, policy status, staff appointment, market price
Data collection technique	site observation, focus group, interview, survey

(composition, association, inheritance). Further, we followed the UML syntax for knowledge representation [16] to represent the semi-formal version of the OntoDPM in UML as depicted in Figure 3. We have chosen the UML knowledge representation formalism because it allows modelling ontologies with instances/individuals, slots and classes, which are also used in Protégé [23].

D. Develop Ontology

We have used the ontology knowledge base editor Protégé [23] to implement the UML class diagram of the OntoDPM in Figure 3. We saved the Protégé file as an OWL file onto the disc; Figure 4 depicts the location and the OWL file icon onto the disc. The Protégé version of the OntoDPM with some hidden components is shown in Figure 5. From the saved OWL file, the OWL ontology will be imported using an appropriate editor.

E. Export the OWL Ontology

Many editors were tested to import/open the OWL file; we found that programming editors including Microsoft Visual Studio, JCreator, and JGrasp could import the OWL file sucessfully. Figure 6 and Figure 7 show the imported OWL ontology in JCreator and Microsot Visual Studio respectively.

VII. DISCUSSION

A detailed discussion on the use of the generated OWL ontology is out of the scope of this paper and will be the focus of our future work. Nevertheless, generating an OWL ontology from a e-government business domain as demonstrated in this

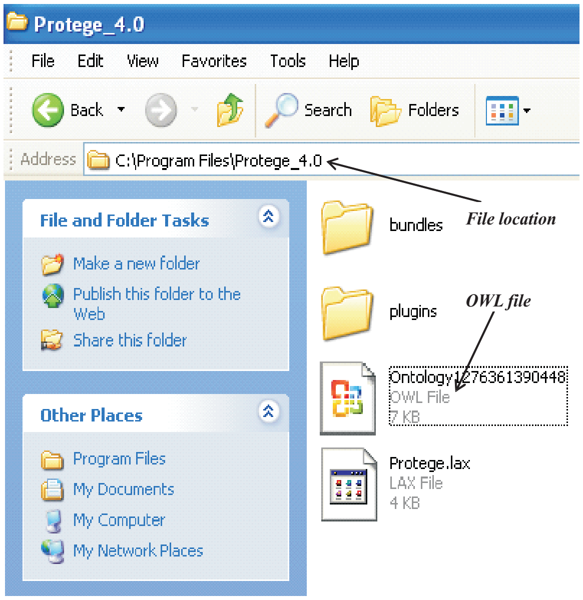
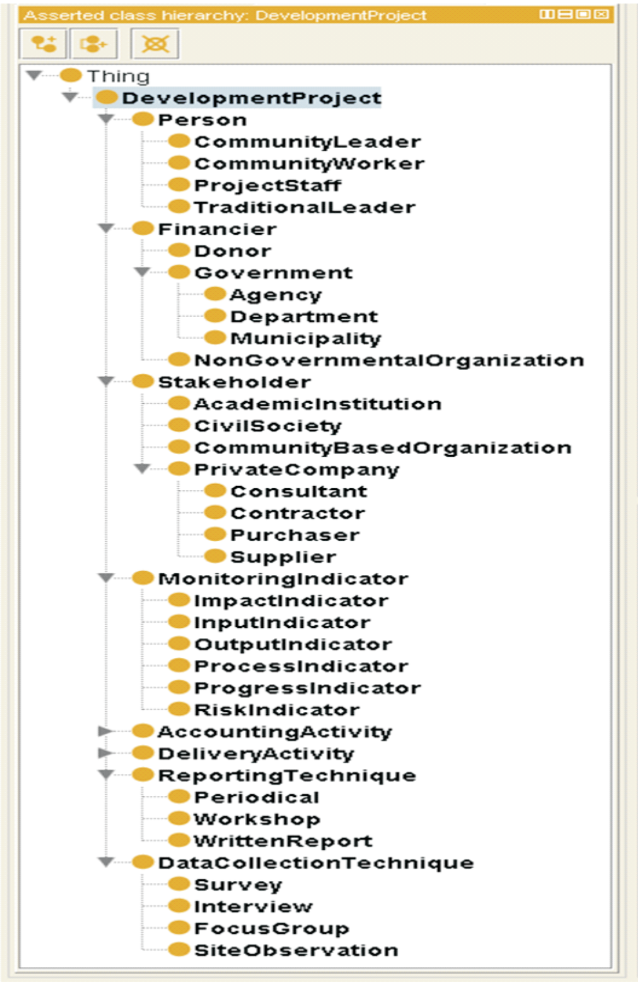
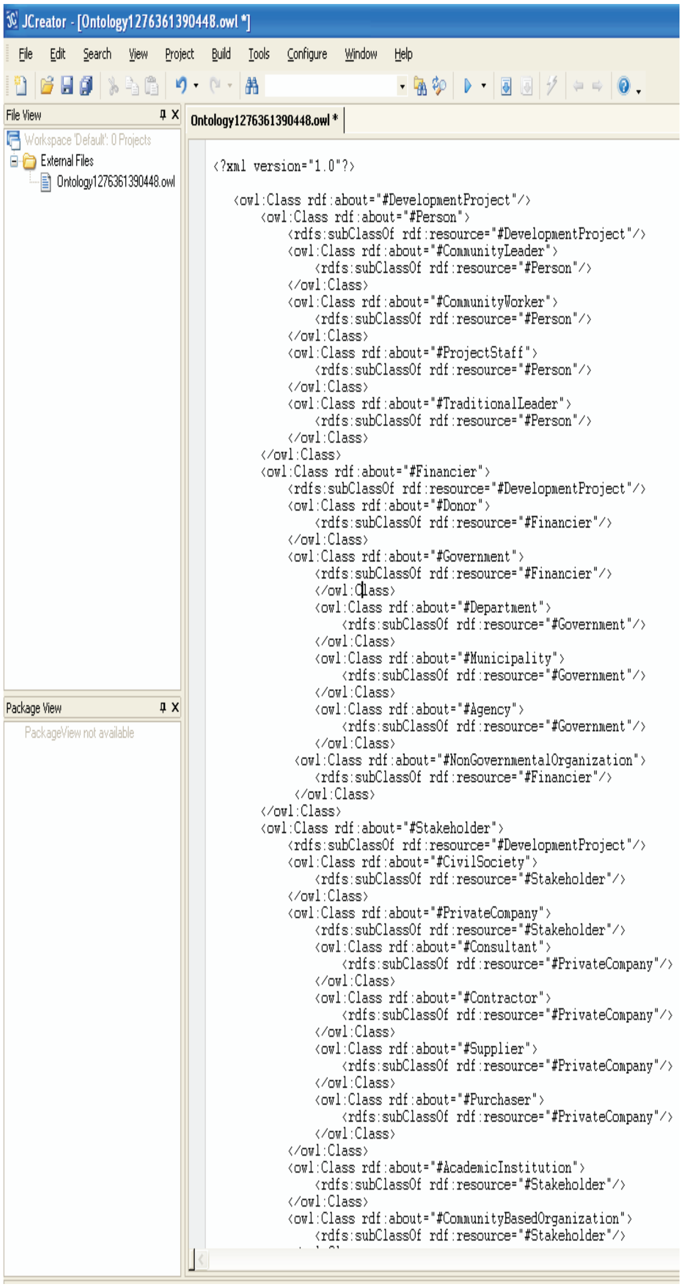


Fig. 4. OWL File and Location onto the Disc



```
Thing
  DevelopmentProject
    Person
      CommunityLeader
      CommunityWorker
      ProjectStaff
      TraditionalLeader
    Financier
      Donor
      Government
        Agency
        Department
        Municipality
      NonGovernmentalOrganization
    Stakeholder
      AcademicInstitution
      CivilSociety
      CommunityBasedOrganization
      PrivateCompany
        Consultant
        Contractor
        Purchaser
        Supplier
    MonitoringIndicator
      ImpactIndicator
      InputIndicator
      OutputIndicator
      ProcessIndicator
      ProgressIndicator
      RiskIndicator
    AccountingActivity
    DeliveryActivity
    ReportingTechnique
      Periodical
      Workshop
      WrittenReport
    DataCollectionTechnique
      Survey
      Interview
      FocusGroup
      SiteObservation
```

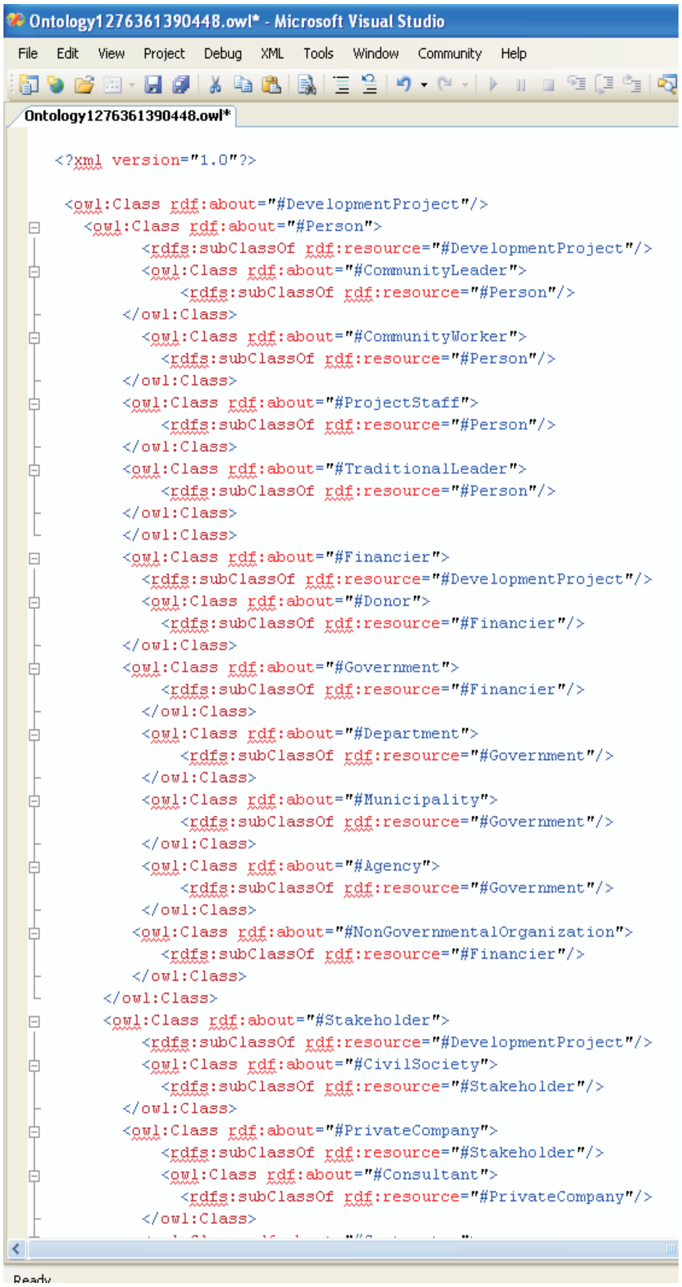
Fig. 5. Protégé Version of the OntoDPM


JCreator - [Ontology1276361390448.owl *]
File Edit Search View Project Build Tools Configure Window Help
File View
Workspace 'Default': 0 Projects
External Files
Ontology1276361390448.owl
Package View
PackageView not available
Ontology1276361390448.owl *

```
<?xml version="1.0"?>

   <owl:Class rdf:about="#DevelopmentProject"/>
       <owl:Class rdf:about="#Person">
           <rdfs:subClassOf rdf:resource="#DevelopmentProject"/>
           <owl:Class rdf:about="#CommunityLeader">
               <rdfs:subClassOf rdf:resource="#Person"/>
           </owl:Class>
           <owl:Class rdf:about="#CommunityWorker">
               <rdfs:subClassOf rdf:resource="#Person"/>
           </owl:Class>
           <owl:Class rdf:about="#ProjectStaff">
               <rdfs:subClassOf rdf:resource="#Person"/>
           </owl:Class>
           <owl:Class rdf:about="#TraditionalLeader">
               <rdfs:subClassOf rdf:resource="#Person"/>
           </owl:Class>
       </owl:Class>
       <owl:Class rdf:about="#Financier">
           <rdfs:subClassOf rdf:resource="#DevelopmentProject"/>
           <owl:Class rdf:about="#Donor">
               <rdfs:subClassOf rdf:resource="#Financier"/>
           </owl:Class>
           <owl:Class rdf:about="#Government">
               <rdfs:subClassOf rdf:resource="#Financier"/>
               </owl:Class>
               <owl:Class rdf:about="#Department">
                   <rdfs:subClassOf rdf:resource="#Government"/>
               </owl:Class>
               <owl:Class rdf:about="#Municipality">
                   <rdfs:subClassOf rdf:resource="#Government"/>
               </owl:Class>
               <owl:Class rdf:about="#Agency">
                   <rdfs:subClassOf rdf:resource="#Government"/>
               </owl:Class>
            <owl:Class rdf:about="#NonGovernmentalOrganization">
               <rdfs:subClassOf rdf:resource="#Financier"/>
            </owl:Class>
       </owl:Class>
       <owl:Class rdf:about="#Stakeholder">
           <rdfs:subClassOf rdf:resource="#DevelopmentProject"/>
           <owl:Class rdf:about="#CivilSociety">
               <rdfs:subClassOf rdf:resource="#Stakeholder"/>
           </owl:Class>
           <owl:Class rdf:about="#PrivateCompany">
               <rdfs:subClassOf rdf:resource="#Stakeholder"/>
               <owl:Class rdf:about="#Consultant">
                   <rdfs:subClassOf rdf:resource="#PrivateCompany"/>
               </owl:Class>
               <owl:Class rdf:about="#Contractor">
                   <rdfs:subClassOf rdf:resource="#PrivateCompany"/>
               </owl:Class>
               <owl:Class rdf:about="#Supplier">
                   <rdfs:subClassOf rdf:resource="#PrivateCompany"/>
               </owl:Class>
               <owl:Class rdf:about="#Purchaser">
                   <rdfs:subClassOf rdf:resource="#PrivateCompany"/>
               </owl:Class>
           </owl:Class>
           <owl:Class rdf:about="#AcademicInstitution">
               <rdfs:subClassOf rdf:resource="#Stakeholder"/>
           </owl:Class>
           <owl:Class rdf:about="#CommunityBasedOrganization">
               <rdfs:subClassOf rdf:resource="#Stakeholder"/>
```

Fig. 6. OWL Ontology Imported with JCreator


Ontology1276361390448.owl* - Microsoft Visual Studio
File Edit View Project Debug XML Tools Window Community Help
Ontology1276361390448.owl*
Ready

```
<?xml version="1.0"?>

 <owl:Class rdf:about="#DevelopmentProject"/>
   <owl:Class rdf:about="#Person">
         <rdfs:subClassOf rdf:resource="#DevelopmentProject"/>
         <owl:Class rdf:about="#CommunityLeader">
             <rdfs:subClassOf rdf:resource="#Person"/>
       </owl:Class>
         <owl:Class rdf:about="#CommunityWorker">
           <rdfs:subClassOf rdf:resource="#Person"/>
       </owl:Class>
       <owl:Class rdf:about="#ProjectStaff">
           <rdfs:subClassOf rdf:resource="#Person"/>
       </owl:Class>
       <owl:Class rdf:about="#TraditionalLeader">
           <rdfs:subClassOf rdf:resource="#Person"/>
       </owl:Class>
       </owl:Class>
       <owl:Class rdf:about="#Financier">
         <rdfs:subClassOf rdf:resource="#DevelopmentProject"/>
         <owl:Class rdf:about="#Donor">
           <rdfs:subClassOf rdf:resource="#Financier"/>
       </owl:Class>
       <owl:Class rdf:about="#Government">
           <rdfs:subClassOf rdf:resource="#Financier"/>
         </owl:Class>
         <owl:Class rdf:about="#Department">
             <rdfs:subClassOf rdf:resource="#Government"/>
         </owl:Class>
         <owl:Class rdf:about="#Municipality">
             <rdfs:subClassOf rdf:resource="#Government"/>
         </owl:Class>
         <owl:Class rdf:about="#Agency">
             <rdfs:subClassOf rdf:resource="#Government"/>
         </owl:Class>
        <owl:Class rdf:about="#NonGovernmentalOrganization">
           <rdfs:subClassOf rdf:resource="#Financier"/>
        </owl:Class>
     </owl:Class>
     <owl:Class rdf:about="#Stakeholder">
         <rdfs:subClassOf rdf:resource="#DevelopmentProject"/>
         <owl:Class rdf:about="#CivilSociety">
           <rdfs:subClassOf rdf:resource="#Stakeholder"/>
       </owl:Class>
       <owl:Class rdf:about="#PrivateCompany">
           <rdfs:subClassOf rdf:resource="#Stakeholder"/>
           <owl:Class rdf:about="#Consultant">
             <rdfs:subClassOf rdf:resource="#PrivateCompany"/>
         </owl:Class>
```

Fig. 7. OWL Ontology Imported with Microsoft Visual Studio

paper is an important step towards the development of Semantic Web applications as e-government applications, which have potential to perform semantic inference and reasoning over the OWL ontologies and facilitate software components integration and interoperability. Moreover, many platforms as Java API, .NET, ASP and so forth, exist for developing Semantic Web applications based on OWL ontologies [4][25].

VIII. Conclusion

This study has presented a framework for constructing semantic model ontologies in OWL Web Service Standard for e-government applications. The proposed framework uses simple ontology engineering techniques (modelling and representation techniques) to capture the semantic content of an e-government service business domain; this makes the framework easy to understand and user-friendly. Furthermore, the platform employed includes Protégé, JCreator, and JGrasp, to create and import the OWL ontology. These are mainly open source software; which make the framework usable by the broader e-government community, particularly e-government developers from the developing countries where there is little or no practice of semantic content representation for e-government systems.

REFERENCES

[1] T. Lee, C.T. Hon and D. Cheung, “XML Schema Design and Management for E-government Data Interoperability,” *Electronic Journal of E-government*, Vol. 7, No. 4, pp. 381-390, 2009.

[2] J. Choudrie and V. Weerrakody, “Horrizontal Process Integration in E-government: The perspective of UK Local Authority,” *International Journal of Electronic Government Research*, Vol. 3, No. 3, pp. 22-39, July-September 2007.

[3] S. Muthaiyah and L. Kerschberg, “Achieving Interoperability in E-government Services with two Modes of Semantic Bridging: SRS and SWRL,” *Journal of Theoritical and Applied Electronic Commerce Research*, Vol. 3, No 3, pp. 52-63, December, 2008.

[4] A. Saekow and C. Boonmee, “A Practical Approach to Interoperability Practical Implementation Support (IPIS) for E-government Interoperability,” *Electronic Journal of E-government*, Vol. 7, No. 4, pp. 403-414, 2009.

[5] M. A. Wimmer, “Integrated Service Modelling for Online One-Stop Government,” *Electronic Markets*, Vol. 12, No. 3, pp.149-156, 2002.

[6] P. Salhofer, B. Stadlhofer and G. Tretter, “Ontology Driven E-government,” *Electronic Journal of E-government*, Vol. 7, No. 4, pp. 415-424, 2009.

[7] L.M.A. Sabucedo, L.E.A. Rifon, F. Corradini, A. Polzonetti and B. Re, “Knowledge-based Platform for E-government Agents: A Web-based Solution Using Semantic Technologies,” *Journal of Expert Systems with Applications, Elsevier Inc.*, Vol. 2010, No. 37, pp. 3647-3656, 2010.

[8] D. Apostolou, L. Stojanovic, T.P. Lobo, J.C. Miro and A. Papadakis “Configuring E-government Services Using Ontologies,” *IFIP International Federation for Information Processing, Springer Boston*, Vol. 2005, No. 189, pp. 1571-5736, 2005.

[9] D. Apostolou, L. Stojanovic, T.P Lobo and B. Thoensen, “Towards a Semantically-Driven Software Engineering Envirionment for E-government,” *IFIP International Federation for Information Processing, M. Bohlen (Eds), TCGOV 2005*, LNAI 3416, pp. 157-168, 2005.

[10] F. Sanati and J. Lu, “Multilevel Life-event Abstraction Framework for E-government Service Integration,” *In Proceedings of the 9th European Conference on E-government 2009 (ECEG 2009)*, London, UK, pp. 550-558, 29-30 June, 2009.

[11] Y. Xiao, M. Xioa and H. Zhao, “An Ontology for E-government Knowledge Modelling and Interoperability,” *In Proceedings of IEEE International Conference on Wireless Communications, Networking and Mobile Computing, (WiCOM 2007)* , Shanghai, pp. 3600-3603, 21-25 September, 2007.

[12] A. Gomez-Perez and V.R. Benjamins, “Overview of knowledge Sharing and Reuse Components: Ontology and Problem-Solving Methods,” *In Proceedings of the IJCAI-99 workshop on Ontologies and Problem-Solving Methods (KRR5)*, Stockholm, Sweden, pp. 1-15, 2 August, 1999.

[13] T.R Gruber, “Toward Principles for the Design of Ontologies used for Knowledge Sharing,” *International Journal Human-Computer Studies*, Vol. 43, pp. 907-928, 1993.

[14] M. Uschold, “Building Ontologies: Towards a Unified Methodology,” *In Proceedings of Expert Systems 96, the 16th Annual Conference of British Computer Society Specialist Group Expert Systems*, Cambridge, UK, pp. 1-18, 16-18 December, 1996.

[15] C.A Welty, “Ontology Research,” *AI Magazine*, Vol. 24, No. 3, pp. 11-12, 2003.

[16] L. Ceccaroni and E. Kendall, “A Semantically-Rich, Graphical Environment for Collaborative Ontology Development en Agentcities,” *iD3*, Barcelona, Spain, pp. 1-6, 2003.

[17] J.A.M. Usero and M.P.B. Orenes, “Ontologies in the Context of Knowledge Organization and Interoperability in e-Government Services,” *IRFD World Forum 2005 - Conference on Digital Divide, Global Development and the Information Society*, November 14-16. Tunis, Tunisia, pp. 1-8, 2005.

[18] C. Calero, F. Ruiz and M. Piattini, “Ontologies for Software Engineering and Software Technology,” *Calero.Ruiz.Piattini (Eds.)*, Springer-Verlag Berlin Heidelberg, 2006.

[19] F. Bettahar, C. Moulin and J.P. Barthes, “Ontologies Supporting E-government Services,” *In Proceedings of the IEEE Artificial Intelligence Conference*, Corvilha, Portuguese, pp. 100-1005, 5-8 December,2005.

[20] A. Mondorf and T. Herborn, “Ontology-based process mediation in the European Project BRITE,” *In the Proceedings of MKWI 2008*, Munich, pp. 341-352, 26-28 February, 2008.

[21] M. Laclavik, “Ontology and Agent Based Approach for Knowledge Management,” *PhD Thesis*, Institute of Informatics, Slovak Academy of Sciences, June, 2005.

[22] J.V. Fonou-Dombeu, “A Conceptual Ontology for E-government Monitoring of Development Projects in Sub Saharan Africa,” *In Proceedings of the Information Society Technologies of Africa 2010 (IST-Africa 2010) Conference*, Paul Cunningham and Miriam Cunningham (Eds), IIMC International Information Management Corporation, Durban, South Africa, pp. 1-8, 19-21 May, 2010.

[23] M. Horridge, H. Knublauch, A. Rector, R. Stevens and C. Wroe, “A Practical Guide to Building OWL Ontoloies Using the Protégé-OWL Plugin and CO-ODE Tools Edition 1.0,” *Research Report*, University of Manchester, UK, 27 August, 2004.

[24] M. Singh and S.K. Malik, “Constructing Ontologies in OWL Using Protégé-2000,” *In Proceedings of the 2nd National Conference on Challenges and Opportunities in Information Technology 2008 (COIT 2008)*, Mandi Gobindgarh, Punjab-India, pp. 1-4, 2008.

[25] H. Knublauch, M. Horridge, M. Musen, A. Rector, R. Stevens, N. Drummond, P. Lord, N.F. Noy, J. Seidenberg and H. Wang, “The Protégé OWL Experience,” *Workshop - OWL: Experiences and Directions (OWLED-2005)*, Galway, Ireland, November, pp. 1-11, 2005.

[26] A. Cregan, M. Mochol, D. Vrandecic and S. Bechhofer, “Pushing the Limits of OWL, Rules and Protégé A Simple Example,” *Workshop - OWL: Experiences and Directions (OWLED-2005)*, Galway, Ireland, pp. 1-10, November, 2005.

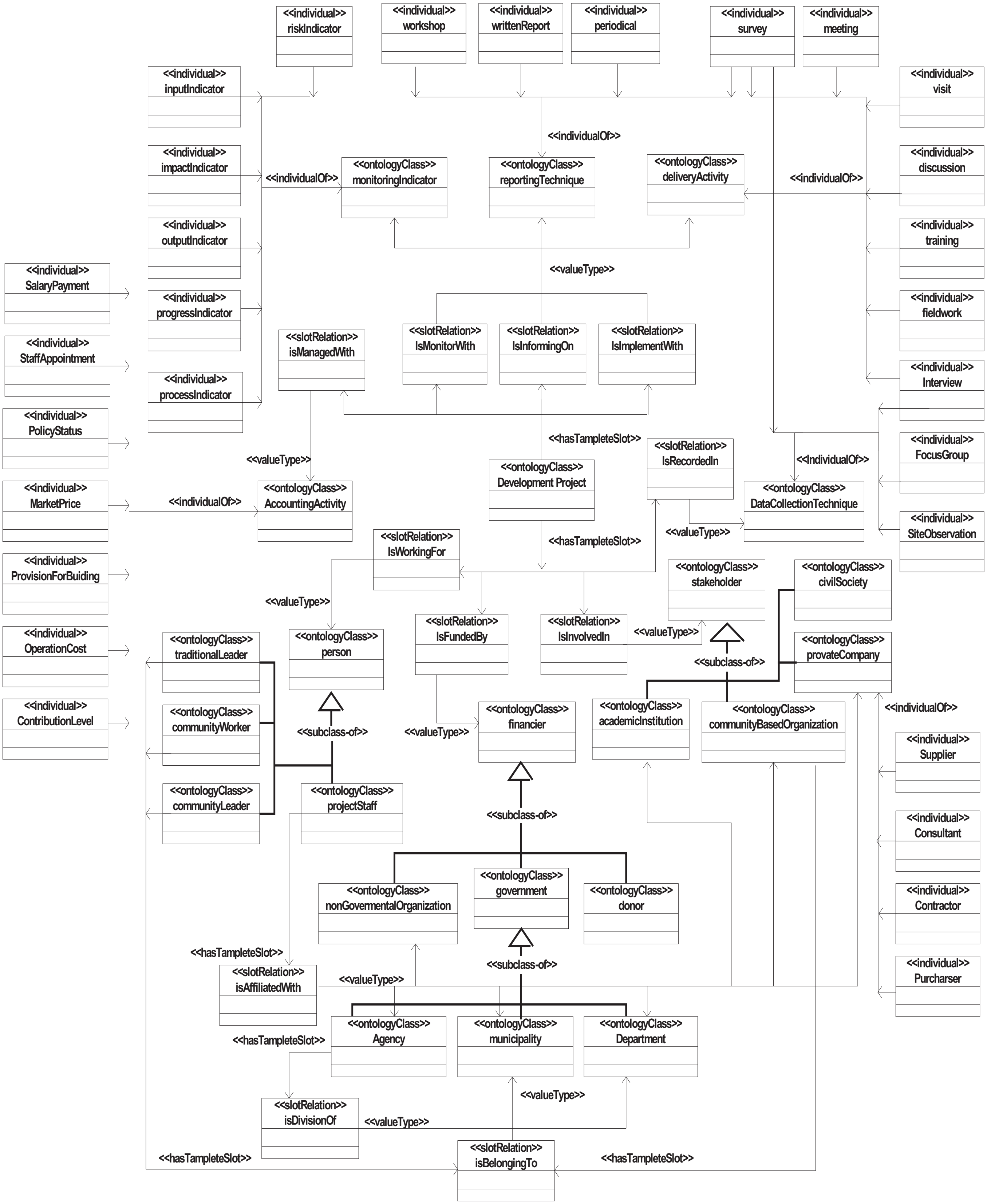

<<individual>> riskIndicator
<<individual>> workshop
<<individual>> writtenReport
<<individual>> periodical
<<individual>> survey
<<individual>> meeting
<<individual>> inputIndicator
<<individual>> visit
<<individualOf>>
<<individual>> impactIndicator
<<individualOf>>
<<ontologyClass>> monitoringIndicator
<<ontologyClass>> reportingTechnique
<<ontologyClass>> deliveryActivity
<<individualOf>>
<<individual>> discussion
<<individual>> outputIndicator
<<individual>> training
<<valueType>>
<<individual>> SalaryPayment
<<individual>> progressIndicator
<<individual>> fieldwork
<<slotRelation>> isManagedWith
<<slotRelation>> IsMonitorWith
<<slotRelation>> IsInformingOn
<<slotRelation>> IsImplementWith
<<individual>> StaffAppointment
<<individual>> processIndicator
<<individual>> Interview
<<individual>> PolicyStatus
<<hasTampleteSlot>>
<<slotRelation>> IsRecordedIn
<<IndividualOf>>
<<individual>> FocusGroup
<<valueType>>
<<ontologyClass>> Development Project
<<individual>> MarketPrice
<<individualOf>>
<<ontologyClass>> AccountingActivity
<<ontologyClass>> DataCollectionTechnique
<<valueType>>
<<individual>> SiteObservation
<<slotRelation>> IsWorkingFor
<<hasTampleteSlot>>
<<individual>> ProvisionForBuiding
<<ontologyClass>> stakeholder
<<ontologyClass>> civilSociety
<<valueType>>
<<slotRelation>> IsFundedBy
<<slotRelation>> IsInvolvedIn
<<valueType>>
<<individual>> OperationCost
<<ontologyClass>> traditionalLeader
<<ontologyClass>> person
<<ontologyClass>> provateCompany
<<subclass-of>>
<<individual>> ContributionLevel
<<ontologyClass>> academicInstitution
<<individualOf>>
<<ontologyClass>> financier
<<ontologyClass>> communityBasedOrganization
<<ontologyClass>> communityWorker
<<subclass-of>>
<<valueType>>
<<individual>> Supplier
<<ontologyClass>> communityLeader
<<ontologyClass>> projectStaff
<<subclass-of>>
<<individual>> Consultant
<<ontologyClass>> government
<<ontologyClass>> nonGovermentalOrganization
<<ontologyClass>> donor
<<individual>> Contractor
<<hasTampleteSlot>>
<<subclass-of>>
<<slotRelation>> isAffiliatedWith
<<valueType>>
<<individual>> Purcharser
<<ontologyClass>> Agency
<<ontologyClass>> municipality
<<ontologyClass>> Department
<<hasTampleteSlot>>
<<valueType>>
<<slotRelation>> isDivisionOf
<<valueType>>
<<slotRelation>> isBelongingTo
<<hasTampleteSlot>>
<<hasTampleteSlot>>

Fig. 3. Semi-formal UML Representation of the OntoDPM

Security and Confidentiality in Interconnected Networks

Aljoša Jerman Blažič
SETCCE
Ljubljana, Slovenia
aljosa@setcce.si

Svetlana Šaljić
SETCCE
Ljubljana, Slovenia
svetlana.saljic@setcce.si

***Abstract*— With the rapid development of interconnected environments many technical and organizational aspects are being addressed in relation to atomized services operating across networked domains. Some of those aspects still need further examination and evaluation also from security perspective. Data may traverse several organizational, security or information domains in order to be processed and results delivered. This calls for proper technical and organizational design approaches, which include means for secure data exchange. Various aspects of secure information exchange are already being addressed by different business, administration, defense and other professional initiatives. The aim of this paper is to present results of the ongoing activities for development and deployment of secure interconnected networks and to demonstrate a framework that shows the relations between relevant security requirements and security mechanisms that can be applied to fulfill the requirements of interconnected domains. The scope of security requirements and mechanisms spans the network, application and information layers.**

Keywords-component; seruity domains, confidentiality, secure data exchnage, security requirments, security mechanisms

I. Introduction

The importance of sharing information across organizational or technical domains is commonly recognized in many professional areas. Novel business or administration concepts supported by technical infrastructures rely on disperse data processing, with services deployed location independent across open networks. These requirements have also been moved to other fields with high security requirements and standards such as governmental or defense organizations [1], which today rely on infrastructures with limited or no connectivity. Data being pushed across network boundaries must be addressed form organizational and technical aspects [4, 5, 13, 15]. In many cases these data may carry sensitive information such as financial statements, governmental decisions or military orders.

In the vision of future networks everyone and everything is connected in order to allow for information sharing: the right information, at the right place, at the right time. Furthermore, dispersed processing power means that information can be processed simultaneously or in sequence at different locations or within different domains and results then combined. Ever increasing throughput, new (web) services based frameworks or cloud computing concepts for instance present foundations for information exchange, shared data processing and distributed storage. However, when it comes to sensitive information, e.g. personal, financial, governmental or even military, sharing requires harmonization and change of the existing capabilities and change in terms of doctrine, processes, personnel, culture and organization.

Examples of information sharing are to be found in many domains e.g. health services, which may combine several organizations exchanging information in order to provide proper and professional health support. Telemonitoring for example requires health services to reach patients across several networks and organizations. Health parameters may be collected from patients on the field by sensor and mobile services providers and then processed by e.g. public health institutions. Corresponding service response may then be delivered by specialized health service provider, which may be operated on private basis or patient’s relatives and professional health personnel in geographical proximity are informed on downgraded health status. It is obvious that in such scenarios data flows may traverse several organizational and security domains and be processed on distributed basis, which only imposes the risk of sensitive information leakage.

Another important concept in this context is Network-Enabled Capabilities (NEC) defined by defense and coalition organizations such as NATO [1, 12, 13, 15]. Defense operations may require fetching data from different organizations such as national weather system, then processed by military geographic data provider and final results on contaminated areas delivered by defense organization to public announcement system in case of e.g. chemical warfare. However, more recent military operations require cooperation of different organizations not only on national basis but on much wider scope between coalition partners. Such future networks are known by the term Network and Information Infrastructure (NII), which is in fact an ever-developing coalition-wide multinational military intranet alike network.

Better integrated networks mean that sharing relevant information will be easier and quicker and that more people will be reached than before. The technical basis for networked operations lies in a secure, robust and extensive federation of networks, a large network consisting of a number of smaller individual networks. This is the case for any area and domain, i.e. business, public administration, non-for-profit organizations and defense organizations.

While technical concepts and mechanisms that support secure data exchange (e.g. access control, encryption, confidentiality labeling) are already well understood and a plethora of technical solutions are being widely implemented and put in practice such as [16], [17], [18], [19], [20], [21], etc., requirements of network and information connections and their functional security in the context of interconnected networks still remains unknown or poorly understood. In this paper we present the attempt to design a framework for collecting requirements, which focus on security mechanisms needed on both the network connection level and the information exchange level. The ultimate aim of the on-going work is to develop a methodology for designing high level system architectures for network and information connections and their associated security mechanisms that support a controlled exchange of information in different contexts. The work presented has background on coalition needs to deploy NII infrastructures, while results presented have much broader impact and may be used in any multi-domain scenarios with sensitive information flows. The primary focus of the paper is to describe a framework, which is supported by identifying some key security requirements and indicating how the framework could be developed and used to address the issues of secure networks interconnection.

II. Sharing Information Across Domain Boundaries

In this paper we address the problem of secure information exchange across domain borders. For this purpose we distinguish four different types of domains:

- Security domains; these domains are defined by the security requirements that apply, e.g. based on a security policy or classification level implemented by a single domain.
- Organizational domains; these domains span the collection of information, systems and infrastructure for which a single organization is responsible, e.g. a domain operated by a single organization.
- Technical domains; these are defined by a collection of technical means used to enable information processing and communication, e.g. local area network.
- Information domains; these are defined by a set of information that is used in a specific functional context by a community of interest (CoI), e.g. information related to a specific business, administration or defense process.

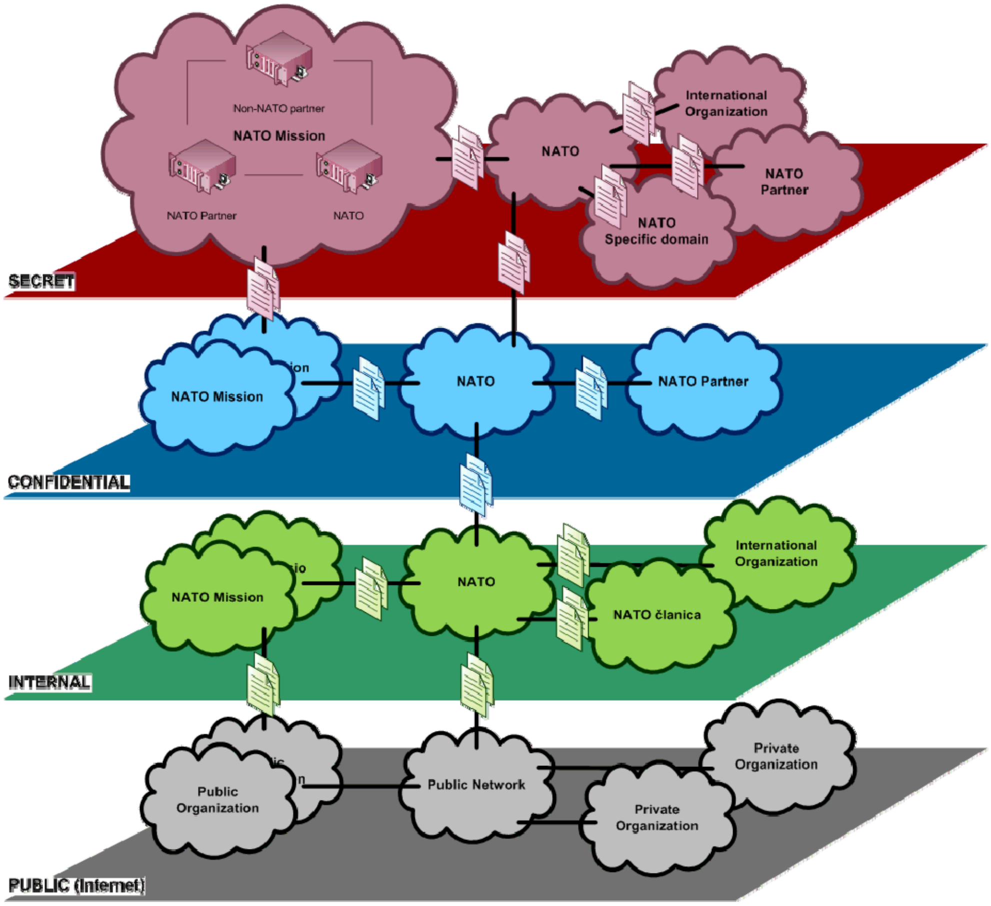


Figure 1. Information flows crossing different domain boundaries.

These domain types provide different perspectives on an information infrastructure. A reason for distinguishing between domain types is that they pose different security requirements with respect to information exchange. When exchanging information one should be aware of which domain boundaries are crossed. Subsequently, relevant security requirements should be identified for each domain boundary individually. Figure 1 illustrates how domains can interrelate in case of coalition organization such as NATO. Information flows may cross different security or organizational domains. In order to prevent unauthorized access or interception in such conditions, data exchange is to be performed in a controlled manner using appropriate security means and organizational measures.

Secure data exchange depends on the context, which addresses crossing at least one boundary (one domain), while many scenarios exist where more than one boundary is being crossed. When multiple domains of different types are crossed an order for the resulting transition sequence must be identified. This can be interpreted by a multidimensional system, where each axis presents one domain type and points in the coordination system data origin, data final destination and data boundaries crossings. Figure 2 presents multi-domain crossing in a three way coordination system.

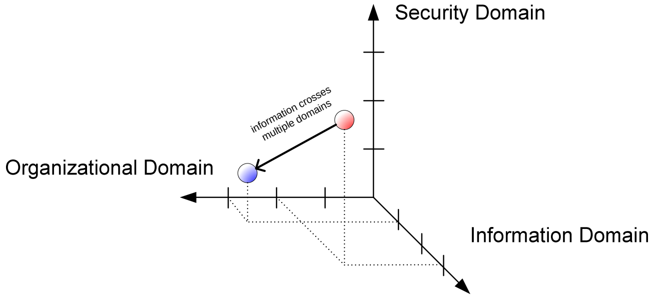


Figure 2. Interpretation of information flow domain crossing in three dimensional coordination system.

Sharing information between security domains presumes a way of determining whether a specific information object may or may not be shared. Each information object has a set of security properties that are relevant in this process.

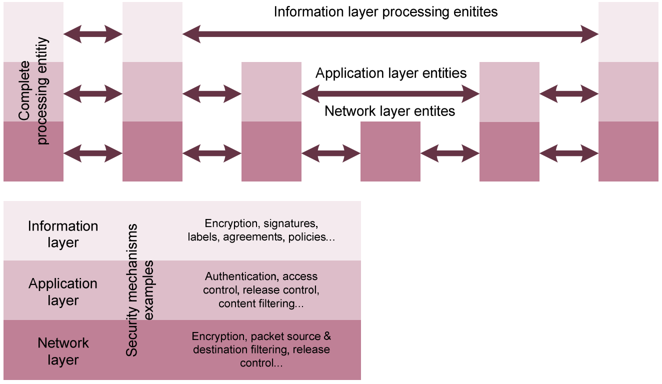
When sharing information a number of general security requirements apply:

1. Making sure that information in a domain can be accessed (community of interest) while access to confidential information is limited according to a set policy (need to know).
2. Allowing information objects to be shared with anybody outside the security domain, while confidential information that should not be shared is protected against unintended release.
3. Managing the flow of information objects over different security domains, based on a shared security policy, e.g. mission-specific or process-specific classifications may be used to support information sharing within a mission or a process while preventing data leakage outside the mission or the process.

Setting up technical and organizational environment that supports such crossings is based on a framework of security mechanisms. Framework is a structured approach to select applicable security mechanisms which operate on different (technical) layers. For the purpose of this paper we differentiate between the network, application and information layers:

- Network layer addresses physical, (inter)network and transport protocols, such as TCP/IP;
- Application layer addresses application-specific protocols such as HTTP, MMHS but also proprietary system interfaces.
- Information layer addresses the actual information payload which may be encoded and/or encrypted.

Security mechanisms may come in various forms and may be applied in different combinations and technical implementations. Their role and function is presented by diagram showing to what layer each mechanism could be applied. It is important to understand that implementing a security mechanism on one layer may impact mechanisms working on other layers. An example is the application of network encryption, which renders application layer packet filtering useless.



Information layer	Encryption, signatures, labels, agreements, policies...
Application layer	Authentication, access control, release control, content filtering...
Network layer	Encryption, packet source & destination filtering, release control...

Figure 3. Security mechanimsms and their application on different layers.

Figure 4 presents an example of secure information exchange between two different security domains. In this case, the following mechanisms are being used: confidentially labeling for selecting information classification level [1], XML guard as release control mechanism [1, 15] and data encryption [18] as access prevention or control mechanism. Both domains have their security policies aligned and classification categories are adjusted. Data being sent from domain A to domain B means that information flow is supported form a higher to a lower classification network. XML Guard is deployed as release control mechanism, encryption is used to support data exchange across unknown networks and classification label is used to support exchange of information on data confidentiality level.

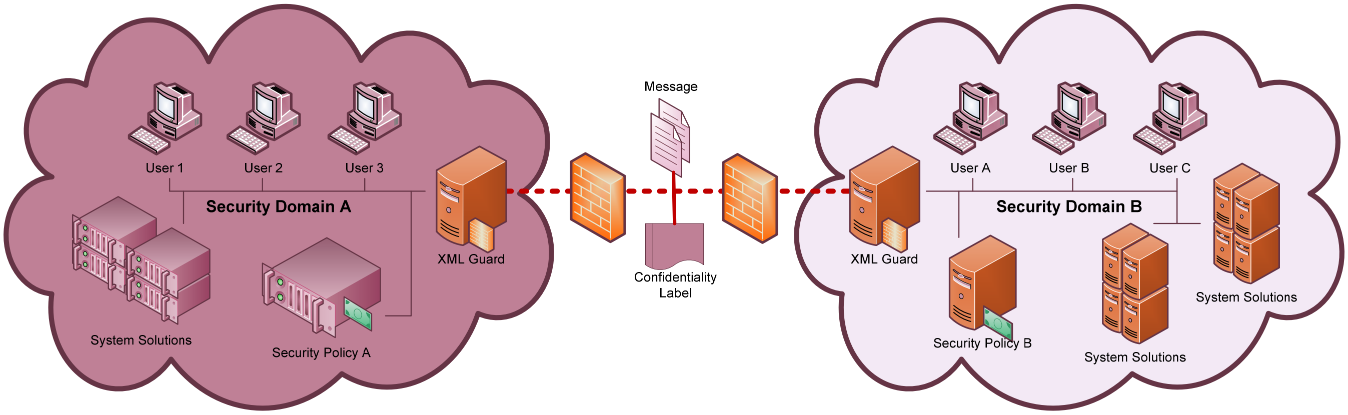


Figure 4. Secure data exchange between different security domains..

III. Security Requirements

Setting up secure environment for inter domain information exchange requires framework of organizational measures and technical means. The primary scope of this paper is to present findings on security requirements and security mechanisms related to secure data exchange, and their role within the interconnected domains. The work relates to national defense and coalition initiatives on network enabled capabilities concepts and interconnected networks.

In order to support secure information exchange across different domains data security (exchange) context has to be first defined. Data security context is associated with many scenarios and has to be addressed with for example exchange of information during defense or tactical operations. The security context includes parameters on degree of data sensitivity in terms of unauthorized access (this is usually defined with the classification category) and how classified data is to be managed. These are the basic considerations based on which security requirements are then collected.

For better understanding the following topics are not in scope for the framework. They do however need to be addressed when the framework is implemented in practical situations.

A. *Policies*

Policies are a set of decisions that are made on the organization level. Security requirements and related security mechanisms can be used to enforce these decisions. It is important to note the fundamental difference between a decision made by a responsible entity (i.e. to give a document a certain classification) and the enforcement of such a decision within the infrastructure (i.e. by either blocking or allowing a document to be transferred to another domain).

B. *Relation to organization and management of information*

With the increasing introduction of enforcement mechanisms in the infrastructure the link between infrastructure, organization and management of information becomes more interrelated. These relations can present complex and dynamics characteristics and for the purpose of the work presented relations are static and known. The actual organization and management of information and related processes are therefore outside the scope of this paper.

C. *Non-security requirements*

The requirements presented in this paper are limited to examples within the security scope. In order to limit the negative impacts e.g. functional and interoperability issues, additional requirements are needed.

Results presented in this paper do not have explicit intention to give a comprehensive overview of all possible security contexts. Rather it describes the basic problems of cross boundaries affection when data are exchanged across domains. In the table below key requirements for secure data exchange applicable to e.g. defense scenarios are summarized.

TABLE I. A LIST OF IDENTIFIED REQUIRMENTS FOR SECURE INFORMATION EXCHANGE ACROSS DOMAIN BOUNDARIES

No.	Requirement	Explanation
1	Information confidentiality	General requirement for data being exchanged between domains
2	Confidentiality breaches detection	Detection of unauthorized downgraded data classification
3	Change/alteration detection	Applicable for complete lifecycle of classified data
4	Information integrity	Applicable for complete lifecycle of classified data
5	(Authorized) changes propagation	Propagation (of security context changes) to relevant entities in communication
6	Trust-relation(s)	A trust-relation between parties established before exchanging information
7	Information non-repudiation	Applies for originating and receiving domain
8	Policies enforcement	Decisions on information

		exchange must not be in violation against policies
9	Policies availability	Availability of all relevant policies for an information set is ensured at all times
10	Explicit policies decisions	Information can be exchanged only when classification level is defined and adjusted
11	Deviations of policies	Deviations from policy should be possible but always detectable
12	Anonymity	Ensure anonymity of data sources (entities)
13	Limited access	Only specified subjects should have access to specified objects – need to know principle
14	Audit trail and integrity demonstration	Information actions trusted log and demonstration of information integrity

The list of collected requirements presented is in scope of defense administration, operation and decision making processes. The intention of listed requirements is not to present a comprehensive list applicable to any scenario or professional area. Other areas may need to implement additional or different requirements. However, most of the requirements are applicable to other areas, such as financial or governmental institutions, while presented collection is used primarily to demonstrate how to set up a framework for secure information exchange in defense scenarios.

IV. SECURITY MECHANISMS

Next, a set of security mechanisms that support secure information exchange in multi domain environment is collected. For each mechanism a short description is given in the table below. This is not architectural proposal but a basic collection of mechanisms needed to support secure data exchange. Conceptual and final architecture must address and include mechanisms, which are selected according to the requirements identified for specific scenario of secure data exchange.

TABLE II. A LIST OF SECURITY MECHANIMSM RELEVNT FOR SECURE INFORMATION EXCHANGE ACROSS DOMAIN BOUNDARIES

No.	Mechanism	Description
1	Authentication	Authentication of subjects for access control purposes
2	Access control	Authorized access to the information and resources (conforming to a policy)
3	Release control	Authorized released of information (conforming to a policy)
4	Security assertions for integrity	Applying additional (meta) information to demonstrate integrity of information in a form of e.g. digital signature or label.
5	Data encryption	Prevent access to information for anyone except by entities in possession of special knowledge (decryption key).
6	Hash function	Integrity demonstration.
7	Confidentiality labeling	Providing information on data classification level(s), classification marking rules and classification parameters.
8	Agreement	Establishing formal relation according to policy requirements between two or more entities.
9	Policy enforcement	Ensure that all decisions concerning secure information exchange are conforming to policy requirements.
10	Policy translation	Interpretation of different policies and establishing mutual understanding of their consequences.
11	Referencing and binding	Describe implicit or explicit relationship between resources or portions of resources.
12	Trusted binding	Support trustworthy relationships between resources or portions of resources
13	Source and destination filtering	Control of information flows based on authorized source and/or destination entities.
14	Content filtering	Excluding specific information elements form information flow based on policy requirements.
15	Segmentation	Partitioning of domains and/or entities in order contain risks of security breaches.
16	Validation	Determining or demonstrating integrity according to a predetermined set of requirements.
17	Time stamping	Process of securely applying or delivering trusted time meta data through the lifecycle of information.
18	Validity period marking	Process of securely keeping track of validity events and defining future validity in the lifecycle of information.
19	Sticky policies	Ensuring that policies are bounded to information during secure exchange.
20	Audit trails and change logs	Chronological sequence based on audit records and other relevant information.

V. FRAMEWORK COMPOSITION

Once security requirements are collected and security mechanisms selected the framework for secure information exchange in multi domain environment can be composed. This part of the process includes several steps, whose result should deliver conceptual model for deployment of proper organizational and technical infrastructure. Bellow is an example of how to define relations between security requirements and security mechanisms.

Step 1 requires determining what domains will be crossed (see chapter II). In this example we foresee a scenario crossing where security and organization domains are crossed: A NATO nation needs to send a part of a national confidential document to NATO. The document has to be filtered for non releasable information, reclassified for NATO confidential.

The next step involves selection of security requirements applicable to the foreseen scenario. In the case of NATO confidential information exchange, the requirements are R1, R2, R4, R6, R7, R8, R9, R10, R11, R13, R14.

In order to show the relation between listed security requirements and the list of security mechanisms we use the following approach: first we select a security requirement and match this to appropriate security mechanisms. Subsequently we add a requirement and matching mechanisms. Then we check if there are any conflicts in the mapping of the mechanisms to the requirements. Implementation layer of security mechanisms is finally added. Table 3 presents the output of the process.

TABLE III. AN EXAMPLE OF RELATION BETWEEN SECURITY REQUIRMENTS AND SECURITY MECHANISMS IN SCENARIOS OF CONFIDENTIAL DATA EXCHENAGE WITHIN COALITON PARTNER AND COALITION.

Mechanism	Requirement	Applicable layer		
		Information	Application	Network
M1	R1, R8, R13		•	
M2	R1, R8, R13		•	•
M3	R1, R8, R13		•	•
M4	R4, R6, R7, R14	•	•	
M5	R1, R8		•	•
M7	R1	•	•	
M8	R6, R8, R9	•		
M9	R8	•	•	
M10	R8, R9, R10		•	•
M11	R1, R8, R10	•	•	
M12	R1, R8, R10	•	•	
M13	R1, R7, R8, R10, R13			•
M14	R1, R7, R8, R10, R13	•	•	
M16	R4, R14	•	•	
M20	R2, R1	•	•	•

Figure 5. An example of relation between security requirments and security mechanisms in scenarios of confidential data exchenage within coaliton partner and coalition.

The example above presents how security mechanisms can be selected and integrated in the infrastructure. The result of the exercise does not deliver a final architectural model. This is to be done in the final step, which is however not relevant for framework development.

When composing a framework, it is important to perform steps of consistency. It might happen that selected security mechanisms collide or that a combination of security mechanisms is affecting the process in a way it is non-executable or that security implications have been tampered. Systematic approach to address these issues is still to be developed.

It is important to understand that security mechanisms are not to be confused with security services or solutions implementation. Utilization of security mechanisms may happen in various combinations, where some instances can cover a significant part of the listed security mechanisms. Therefore, the methodology presented, takes into account only logical components of the final architecture. Further development should therefore among others be focus on interpreting the results and their translation to security implementations.

VI. CONCLUSIONS

The work presented does not only illustrate how to address secure data exchange in interconnected infrastructures but how to understand implications of data being exchanged across different domains. Connecting secure networks has a long history and has been around since the introduction of open systems. However, local networked environments have long ago become part of global infrastructures and the last barriers for unconnected high risk networked environments are coming down. This is aligned with the key strategy of coalition partners to utilize a single network supporting order and decision processes.

Through the research work performed on security in interconnected networks it has been proved that a single security architecture is not feasible since an appropriate architecture depends on the (security) context (scenario) to which it needs to be applied. Rather it is essential to identify security context for scenario or a collection of scenarios. This can be very demanding process and proposed example, which illustrates approach taken, only shows that a significant amount of resources are required to identify proper architectural concept for a specific scenario.

Further work lies ahead, mostly focused on development of methodologies, which will ease security context definition, security requirements and security mechanisms mapping. Another research topic should focus on security profiles, which address common scenarios, and security contexts. Profiles may also deliver predefined architecture compositions for secure data exchange based on a specific context. Similar work has already been done for the security patterns through European Commission funded 6th Framework programme integrated project SERENITY [26]. The project was focused on developing methodologies and languages for capturing security and dependability knowledge in heterogeneous and pervasive environments. The result of 3 years research was a platform for collecting security requirements and library of patterns [23, 24]. Patterns are layered through business and organization levels, workflows and processes level, devices and networks level. Specific language was developed in order to interpret patterns, which provide security properties (e.g. information confidentiality), context in which a pattern is to be used (e.g. communication over IP protocol) and one or more implementations (e.g. data encryption using IPSec).

Another important research topic is security mechanisms utilization and standardization of security solutions orchestration. In scenarios of dynamic services and concepts such as cloud computing, security requirements and supportive security mechanisms must be selected and composed in consistent services solution on the fly. This is why it is of utmost importance for future research in this field to address above issues and propose standardized methodologies for setting up frameworks for secure information exchange in interconnected environments.

REFERENCES

[1] B.J. te Paske, D. Boonstra, D.H. Hut, and H.A. Schotanus, "Cross-Domain Solutions, A conceptual model", Whitepaper, TNO, 2009.

[2] P. Hoffman, "Enhanced Security Services for S/MIME", RFC 2634, IETF, June 1999.

[3] National Institute of Standards and Technology, "Standard Security Label for Information Transfer", FIPS 188, NIST, 1994.

[4] ISO/IEC, “Information technology Portable Operating System Interface (POSIX)”, ISO/IEC 9945 and IEEE 1003-1, ISO/IEEE, 2003

[5] ISO, "Information Technology – Open Systems Interconnection – Security Frameworks for Open Systems: Overview", ISO 10181-1, ISO, 1996.

[6] ISO, "Information Technology – Open Systems Interconnection – Security Frameworks for Open Systems: Access Control Framework", ISO 10181-3, ISO, 1996.

[7] National Security Agency, "Common Criteria Labeled Security Protection Profile", NSA, 1999.

[8] ISO/ITU, "Information Technology – Security Techniques – Security Information Objects for Access Control", ISO/IEC 15816, ITU-T X.841, ISO/ITU, 2002.

[9] W. Nicolls, "Implementing Company Classification Policy with the S/MIME Security Label", RFC 3114, May 2002.

[10] A. Thümmel and K. Eckstein, "Design and Implementation of a File Transfer and Web Services Guard Employing Crypto-graphically Secured XML Security Labels", Proceedings of the 7th IEEE Workshop on Information Assurance, U.S. Military Academy, West Point, NY, pp. 26 – 33, IEEE, 2006.

[11] OASIS, "eXtensible Access Control Markup Language v2.0", XACML, 2005.

[12] NATO, "Information Exchange Gateways Reference Architecture" March 2009

[13] NATO, "Core Enterprise Services Framework V1.2", January 2009.

[14] http://www.xmlspif.org/, "Open XML SPIF - XML Schema for Security Policy Information File". Last access November 2010.

[15] http://www.isode.com/whitepapers/isode-security-infrastructure.html, "Isode Security Policy, Security Label and Security Clearance Infrastructure", August 2008. Last access November 2010.

[16] D. Eastlake, J. Reagle, and D. Solo, "XML-Signature Syntax and Processing", RFC 3275, IETF, 2002.

[17] T. Imamura, B. Dillaway, and E. Simon, "XML Encryption Syntax and Processing", XMLEnc, W3C, 2002.

[18] T. Dierks and E. Rescorla, "The Transport Layer Security (TLS) Protocol", RFC 5246, IETF, 2008

[19] S. Kent and R. Atkinson, "Security Architecture for the Internet Protocol", RFC 2401, IETF, 1998

[20] M. Wahl, T. Howes, and S. Kille, "Lightweight Directory Access Protocol", RFC 2251, 1997

[21] R. Housley, W. Ford, W. Polk, and D. Solo, "Internet X.509 Public Key Infrastructure Certificate and CRL Profile", RFC 2459, IETF, 1999

[22] J. Porekar, K. Dolinar, and A. Jerman Blažič, "Design Patterns for a Systemic Privacy Protection", Journal On Advances in Security, vol 2, no 2&3, pp. 267 – 287, IAIRA, September 2009.

[23] J. Porekar, A. Jerman Blažič, and T. Klobučar, "Towards Organizational Privacy Patterns", Proceedings of The Second International Conference on the Digital Society, pp. 15 – 19, IEEE, February 2008

[24] European Comission FP6 Intergated Project SERENITY, http://www.serenity-project.org/, accessed october 2010.

ISBN: 978-1-61208-116-8

Online Banking Fraud Detection Based on Local and Global Behavior

Stephan Kovach
Laboratory of Computer Architecture and Networks
Department of Computer and Digital System
Engineering, Polytechnic School of São Paulo
São Paulo, Brazil
skovach@larc.usp.br

Wilson Vicente Ruggiero
Laboratory of Computer Architecture and Networks
Department of Computer and Digital System
Engineering, Polytechnic School of São Paulo
São Paulo, Brazil
wilson@larc.usp.br

***Abstract* – This paper presents a fraud detection system proposed for online banking that is based on local and global observations of users' behavior. Differential analysis is used to obtain local evidence of fraud where a significant deviation from normal behavior indicates a potential fraud. This evidence is strengthened or weakened by the user's global behavior. In this case, the evidence of fraud is based on the number of accesses performed by the user and by a probability value that varies over time. The Dempster's rule of combination is applied to these evidences for final suspicion score of fraud. Our main contribution is a fraud detection method based on effective identification of devices used to access the accounts and assessing the likelihood of being a fraud by tracking the number of different accounts accessed by each device.**

Keywords- differential analysis; local and global behavior; device identification; Dempster-Shafer theory

I. INTRODUCTION

Fraud prevention describes the security measures to avoid unauthorized individuals from initiating transactions on an account for which they are not authorized [1]. In spite of many advanced mechanisms available for fraud prevention for online banking applications, it can fail. Fraud detection consists in identifying such unauthorized activity once the fraud prevention has failed. In practice, fraud detection must be used continuously, since the system is unaware that fraud prevention has failed [1]. Among the approaches used by fraudsters, phishing is one of the most common forms for stealing account details for authentication from the customers. Social engineering is the most common method used in phishing. Social engineering usually comes in the form of e-mails trying to convince users to open attachments or by directing them to some fraudulent site**,** and most of the time it is so well designed that many costumers are led to informing their account details**.**

This paper presents a framework, and the corresponding system, for online banking fraud detection in real time. It uses two complementary approaches for fraud detection. In the differential analysis approach, the account usage patterns are monitored and compared with the history of its usage, which represent the user's normal behavior. Any significant deviation from the normal behavior indicates a potential fraud [2].

In the global analysis approach, each device is monitored and classified as legitimate or fraudulent with certain probability based on global information. This is based on three assumptions. First, it is assumed that each device used for online banking has a single identification. The second assumption is based on the fact that the probability of a transaction being a fraud increases with the number of accounts accessed by the same source that requested the current transaction. The third assumption comes from the fact that the only way to know that a fraud has been perpetrated is when the customer reports it.

The major contribution of this paper is the finding, by empirical analysis of a real-world transaction dataset, that the effective identification of access devices and monitoring the number of different accounts accessed by each device is a very promising supplement for other methods in detecting fraudulent behavior in online banking applications.

This paper is organized as follows: Section 2 presents an overview of related work on fraud detection. Section 3 describes some characteristics of online banking frauds. Section 4 details the proposed fraud detection system. Section 5 concludes the paper and outlines future work.

II. RELATED WORK

There are few published works about fraud detection within the domain of online banking applications. This is most likely due to the privacy, the secrecy and the commercial interests concerning this domain, rather the absence of research [3]. Therefore, due to the limited exchange of ideas, the development of new fraud detection methods in the banking area is difficult. Most published work is related to the domain of credit card, computer intrusion and mobile communication. Some relevant works on fraud detection are reviewed next.

Credit card frauds- Most of the works on preventing and detecting credit card fraud were carried out with special emphasis on data mining and neural networks. Aleskerov, Freisleben and Rao [4] describe a neural network based database mining system in which a neural network is trained with the past data of a particular customer and the current spending patterns is processed to detect possible anomalies. However, Bolton and Hand [5] proposed a detection technique in which break point analysis is used to identify changes in spending behavior.

ISBN: 978-1-61208-116-8

Computer intrusion- Intrusion detection approaches in computers is broadly classified into two categories based on a model of intrusions: misuse and anomaly detection. Misuse detection attempts to recognize the attacks of previously observed intrusions in the form of a pattern or a signature and then monitors such occurrence. Anomaly detection tries to establish a historical normal profile for each user, and then uses sufficiently large deviation from the profile to indicate possible intrusions [6]. Denning [7] presents a statistical model for real-time intrusion detection based in anomaly detection. Ghosh and Schwrtzbard [8] describe an approach that employs artificial neural networks used for both anomaly and misuse detection.

Mobile communication frauds - Fraud in communication networks refers to the illegal access to the network and the use of its services. Cortes and Pregibon [9] define statistical summaries, denominated signatures, of users over two time windows, namely, current and historical, respectively. The current network activity is compared with the historical activity for any deviation. Fawcett and Provost [10] present rule-based methods and neural networks for detecting fraudulent calls based on profiling subscriber behavior.

In all domains above mentioned, fraudsters tends to adapt to new prevention and detection measures. In the same way, legitimate users may gradually change their behavior over a longer period of time. Therefore, fraud detection techniques need to be adaptive and to evolve over time in order to avoid false alarms. Models can be updated at fixed time points or continuously over time [9][10].

Panigrahi, Kundu, Sural, and Majumdar [11] describe a framework for fraud detection in mobile communication networks using rule-based deviation method. The main point of this paper is the detailed description of the use of Dempster-Shafer theory in order to combine the evidences of fraud given by two rules.

The system proposed in this paper combines three different approaches: (1) differential analysis using statistical models in order to detect local evidence of fraud; (2) an innovative approach using a probabilistic model for evaluating the likelihood of a transaction being a fraud based on its global behavior; and (3) Dempster-Shafer theory for combining evidences of fraud.

III. ONLINE BANKING FRAUD CHARACTERISTICS

An empirical analysis performed on real-world transactions datasets revealed that most of frauds had the following behavior characteristics:

- Large number of different accounts accessed by a single fraudster;
- Transactions involving small values in many accounts;
- More payment transactions than usual in a single account;
- Increased number of password failures before the occurrence of frauds.

While the latter two characteristics can be detected by differential analysis using local attributes, the first two characteristics need information about similar attacks in other accounts. The fraud detection system described in the next section takes these characteristics into account.

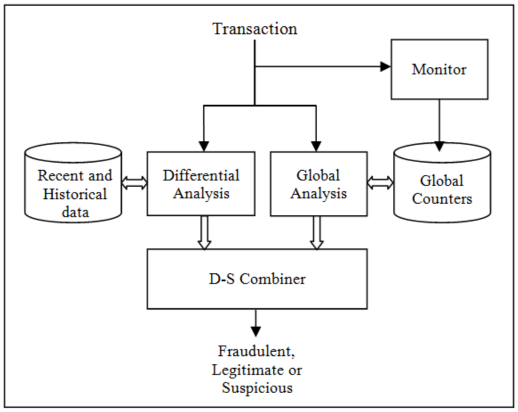


Figure 1. The general architecture of the system

IV. THE FRAUD DETECTION SYSTEM

The general architecture of the proposed fraud detector is illustrated in Fig. 1. In this architecture, each access device from which transactions are performed is supposed to have an identity. These identities are used along with a set of counters to monitor the number of different accounts accessed by each device. The system uses two independent approaches for detecting frauds: a differential analysis approach that detects significant changes in transaction patterns in individual accounts, and a global analysis approach that uses the set of counters to detect unusual number of accounts accessed by a single device. The fraud evidences determined by the two approaches are then combined in order to determine an overall score that may trigger an alarm depending on a prefixed threshold. The main issues of the architecture are discussed in the next subsections.

A. Device Identification

The proposed fraud detection technique has as a core concept, the notion of access device identity.

In the domain of online banking, where accesses are made through the Internet, the identification of source devices based in IP address only is difficult since it can change over time. In the proposed approach, the identification of the access device is made by a component that must be downloaded and installed in the client device. This component generates a *fingerprint* of the access device and sends it to the bank site as part of each transaction data. The *fingerprint* is calculated by applying a cryptographic function on the hardware and software information, as the processor and the operating system serial numbers, MAC address, and some configuration details.

The implementation details of the component are out of the scope of this paper. It assumes that the component is implemented with three main requirements:

1. It generates a different fingerprint for each different access device;

2. It introduces some randomness during the fingerprint generation in order to difficult its spoofing by other devices;
3. It informs the new fingerprint whenever the configuration of the device changes.

Actually the proposed system is based on the component that is already being used by the actual online banking system.

B. Global Behavior and the Monitor

The observation of the user's global behavior plays a major role in the fraud detection system proposed herein. An example of global behavior that may evidence a fraud is the large number of different accounts accessed by a single device. Another example is the occurrence of login fails over many accounts using a single trial password. A set of counters are used in order to verify the global behavior of the users. As shown in Fig. 1, the monitor accounts for updating these counters at each incoming transaction.

C. Differential Analysis

In the differential analysis approach, an incoming transaction is examined against a set of profiles that characterize the normal usage pattern of a legitimate customer. If the current usage pattern deviates significantly from the customer’s average usage pattern, it may indicate a potential fraud. In order to calculate this deviation, two buffers are used in such a way that all transactions submitted in the current session enter the first buffer. The second buffer keeps a certain number of most recent transactions. The transactions in the first and second buffer are used to calculate the current usage pattern and the customer’s average usage pattern, respectively. Then, the deviation is calculated using a statistical method the result of which is a probabilistic value that gives a degree of belief in the evidence of fraud. If this session is classified as legitimate, all the transactions of the first buffer are inserted in the second buffer and the oldest transactions are removed from it [12].

Some of the profiles monitored by this module are described below:

- *Payment transaction frequency*. This profile is monitored in order to detect the sudden increase of payment transactions, which is unusual to legitimate a user.
- *Password failures*. The measure of password failures at login time is compared against a fixed limit that is determined from prior observations. This profile is useful for detecting attempted break-ins [7].
- *Login frequency*. Profiles for login frequencies by day and time are monitored to detect fraudsters who try to log into an unauthorized account during a period of time when the legitimate user is not expected to be using the account [7].

D. Global Analysis

The purpose of the global analysis module is for strengthening or weakening the evidence of a fraud determined by the differential analysis module. It is performed by evaluating a new evidence of fraud based on global observation of the user’s device behavior. The evidence of fraud, given by a probability, is determined by means of three lists: Black List, White List and Suspect List. The Black List contains the identity of devices associated to transactions that have already been classified as fraudulent. The White List contains the identities of the devices, as well as the account numbers accessed by them, associated to transactions classified as legitimate. The Suspect List contains the identities of devices the transactions of which have not yet been classified. The assignment of the devices to one of those lists and the determination of its fraud probability score are driven by rules described as follows:

For each incoming transaction,

- If the current device is in the Black List, then the fraud probability is assigned to one meaning that the transaction is fraudulent with a high level of evidence;
- If the current device and the account number accessed by it are in the White List, then the fraud probability is assigned to zero denoting that the transaction is legitimate with high level of evidence. Note that the device identity may be associated with one or more accounts in the White List. This is the case in which a single user has many accounts;
- If the current device is neither in the Black List nor in the White List, then the device identity and the accessed account number are included in the Suspect List. While in this list, the fraud probability of this transaction is determined by an exponentially decaying function described in the next section.

E. The Suspect List and the Exponentially Decaying Function

If the incoming transaction device is inserted in the Suspect List, it will remain there until explicitly classified as fraudulent or legitimate, when the associated device identity and account number are inserted in the Black or White List, respectively. The idea behind this rule comes from the fact that a given transaction can only be assured as fraudulent by the customer himself/herself.

If no fraud is reported until the end of a prefixed period of time, nothing can be said about the trustiness of this device. In this case, the device will be moved to the White List since it is more likely to be legitimate based on the analysis made on real-world transactions dataset. However, a flag is set indicating that this device was moved to the White List at the end of predefined period of time and not explicitly classified as legitimate. This flag is used by the fraud analyst if a fraud performed from this device is detected later. Since the device had been moved to the White List, the next transactions performed by this device will be regarded as legitimate by this module.

The elapsed time since the occurrence of a fraud and its detection by the customer can take more than a month. According to the information from analysts of a real banking institution, there are some cases in which it takes up to two month to be reported by the customer. The reason for this delay is due to the fact that many fraudulent transactions go unnoticed since the values involved in individual transactions are usually very small.

When a device is included in the Suspect List, an initial value is assigned to the fraud probability. This value is calculated by an exponentially decaying function that depends on the number of different accounts that were accessed by this device the transactions of which have not yet been classified. If a fraud on any of these accounts is reported by the customer, the associated device identity will be moved to the Black List.

The exponentially decaying function was chosen due to the fact that most of the frauds are reported as soon as they were committed and very few at the end of some period of time, for example, two months later. In other words, the probability of being a fraud is higher at the beginning of a transaction, decaying at a fast rate along the time.

The exponentially decaying function is expressed as

$$P(t) = P_{max} \,.\, e^{-\lambda t} \qquad (1)$$

where,

P_{max} is the maximum probability value assigned to the device when it is included in the list. It depends on the number of different accounts accessed by the device (N), since the probability of being a fraud increases with this number. For the initial trial, P_{max} was chosen as being equal to N/10 for $1 < N < 8$, and 0.9 for $N \geq 9$. The maximum value of 0.9 was chosen since 1.0 is reserved for assured fraudulent devices, i.e., included in the Black List.

λ is calculated such that at the end of the period (t_{end}), the probability value reaches an arbitrary low value. Assuming t_{end} = 60 days and $P(t_{end}) = 0.01$, then

$$\lambda = - (1/60).\ln(0.01/P_{max}) \qquad (2)$$

Fig. 2 shows the exponentially decaying curves for each value of N.

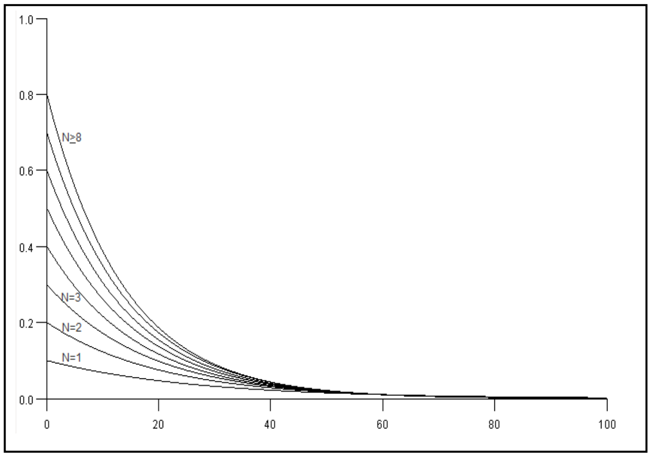


Figure 2. Exponentially decaying curves

The dashed line in Fig. 3 shows an example of the probability values assigned to a device that varies over the time. Note that when the number of accounts accessed by it increases, the probability value jumps to its correspondig maximum value.

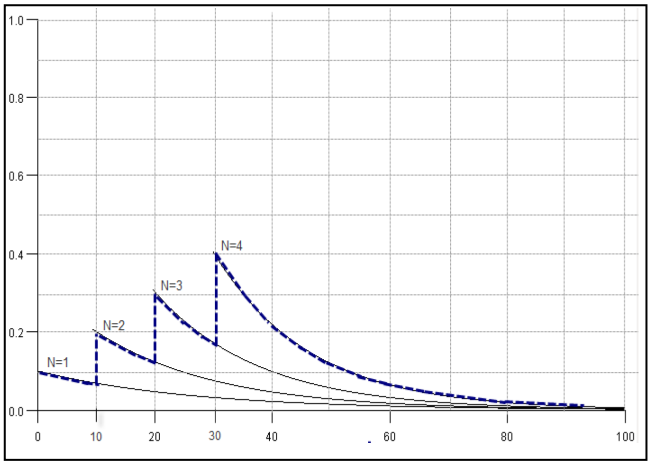


Figure 3. The fraud probability under exponential decaying function

F. Dempster-Shafer Combiner

This module uses the Dempster-Shafer theory to combine the evidences of fraud estimated by differential and global analysis modules and computes the overall suspicion score of a transaction.

The Dempster-Shafer (D-S) theory is a mathematical theory of evidence that provides a formal framework for combining sources of evidence [13].

The main difference between the D-S theory and the probability theory is that the former allows the explicit representation of uncertainty. The other difference is that the D-S theory requires no knowledge of prior probabilities.

The D-S framework is based on a view that hypotheses can be regarded as a subset of a given set of mutually exclusive and exhaustive possibilities named a *frame of discernment* [13]. For the fraud detection domain, the frame of discernment Θ is consisted by two mutually exclusive values, given as: $\Theta = \{$*fraud, -fraud*$\}$. The set of all possible hypotheses of Θ corresponds to all subsets of Θ including itself, denoted by 2^{Θ}. In the case of fraud, the power set is consisted by three possible hypotheses: {*fraud*}, {*-fraud*} and Θ = {*fraud,-fraud*} (denoting the uncertainty).

A probability number *m(h)* between 0 and 1, expressing an estimative of confidence or belief, is assigned to a hypothesis *h*. This number is called *basic probability assignment (bpa)* or *mass*. In our system, the probabilistic values computed by the local and global analysis modules are applied to basic probability assignments.

Two functions are defined in the D-S theory in order to express uncertainty: *Belief function (Bel) and Plausibility function (Pls)*

ISBN: 978-1-61208-116-8

Belief function (Bel) is the total belief committed to a hypothesis. It sums the mass of all non-empty subsets of the hypothesis and the mass of hypothesis itself.

Plausibility function (Pls) takes into account all the masses assigned to a hypothesis and those that can be plausibly transferred to it in the light of new information [13]. It defines the maximum belief that can be committed to a hypothesis.

Belief and Plausibility functions are related as follow:

$Pl(H) = 1 - Bel(-H);$

$U(H) = Pl(H) - Bel(H).$

where, *Bel(-H)* means *disbelief of H*, i.e., belief that refutes the hypothesis *H*; and *U(H)* means uncertainty of *H*.

Bel(H) and Pls(H) represents the upper and lower bounds in the evidence of hypothesis *H*.

The Dempster's rule of combination gives a function for evaluating an overall score from two evidences. Given two basic probability assignment of evidences $m_1(h)$ and $m_2(h)$, they may be combined into a third basic probability assignment $m_3(h)$ by the expression below:

$$m_3(h) = m_1(h) \oplus m_2(h) = \frac{\sum_{x \cap y = h} m_1(x) \cdot m_2(y)}{1 - \sum_{x \cap y = \varnothing} m_1(x) \cdot m_2(y)}$$

where, the symbol ⊕ denotes orthogonal sum.

This rule can be used for combining basic probability assignment of all features monitored by the local and global analysis modules and then obtaining overall summary values for each module. These summary values from both modules are then combined to provide the final suspicion score [14]. Based on this score, a transaction on a given account can be detected as fraudulent, legitimate or suspicious.

V. CONCLUSION AND FUTURE WORK

In this paper, we have introduced a novel approach for fraud detection in online banking transactions by using global counters and an effective identification of access devices. The idea behind this approach comes from the fact that fraud suspicion in a transaction increases with the number of accounts accessed from the same source. The effective identification of devices is made by a component that is downloaded and installed in each device during its first access to the bank. A monitor counts the number of different accounts accessed by each device. These counters are then used by the global analysis module that estimates the likelihood of a transaction being a fraud. The paper describes the details of how this likelihood is evaluated. A differential analysis is also performed on each transaction against a set of customer profiles. This approach is based on the proposition in which any significant deviation from the normal behavior indicates an evidence of fraud. The Dempster's rule combines the resulting fraud evidences from global and differential analysis to calculate the overall suspicion score of each transaction.

The proposed system is very promising in detecting fraudulent transactions in online banking applications with low rate of false alarms.

The benefit of this approach comes from the fact that most of fraudsters do not attack a single account, but many accounts from a single device. Therefore, the simple observation of a device's global behavior, such as the number of different accounts that has been accessed by it, can bring more evidences rather than just applying complex statistical methods on its local parameters.

Currently, the system is in its final stage of development. It is being validated and its parameters adjusted using a real-world transaction dataset.

Among the directions for future work we are regarding the development of a *simulator* that produces different patterns of legitimate and fraudulent transactions in any proportion and randomness in order to evaluate the best threshold values for low rate of false alarms, and the study of new algorithms and probabilistic functions for global analysis.

REFERENCES

[1] R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review". Statistical Science. Vol. 17, No. 3, 2002, pp 235-255.

[2] U. Murad and G. Pinkas, "Unsupervised profiling for identifying superimposed fraud", in Proceedings of the 3rd European Conference on Principles of Data Mining and Knowledge Discovery, 1999, pp. 251-266

[3] K. N. Karsen and T. G. Killingberg, "Profile based intrusion detection for Internet banking systems", Master Thesis, Norwegian University of Science and Technology, Norway, 2008

[4] E. Aleskerov, B. Freisleben, and B. Rao, "CARDWATCH: A neural network based database mining system for credit card fraud detection", in Computational Intelligence for Financial Engineering. Proceedings of the IEEE/IAFE, 1997, pp 220- 226. IEEE, Piscataway, NJ.

[5] R. J. Bolton and D. J. Hand, "Unsupervised profiling methods for fraud detection", in Conference on Credit Scoring and Credit Control 7", Edinburgh, UK, 5-7 Sept., 2001.

[6] Y. Kou, C.T. Lu, S. Sirwonqattana, and Y.P. Huanq, "Survey of fraud detection techniques", in Proceedings of the IEEE International Conference on Networking, Sensing and Control, vol. 1, 2004, pp. 749-754.

[7] D. E. Denning. "An intrusion detection model". IEEE Transactions on Software Engineering, 13:222-232, February 1987.

[8] A. K. Ghosh and A. Schwartzbaxd. "A study in using neural networks for anomaly and misuse detection", in Proceedings of the 8th USENIX Security Symposium, 1999.

[9] C. Cortes and D. Pregibon, "Signature-based methods for data streams," Data Mining and Knowledge Discovery, vol. 5, no. 3, pp. 167-182, 2001.

[10] T. Fawcett and F. Provost, "Adaptive fraud detection", Data Mining and Knowledge Discovery Journal, Kluwer Academic Publishers, Vol. 1, No. 3, 1997, pp. 291-316.

[11] S. Panigrahi, A. Kundu, S. Sural, and A. K Majumbar, "Use of Dempster-Shafer theory and Bayesian inferencing for fraud detection in communication networks", Lecture Notes in Computer Science, Spring Berlin/ Heidelberg, Vol. 4586, , 2007, p.446-460.

[12] P. Burge and J. Shawe-Taylor, "Detecting cellular fraud using adaptive prototypes", Proceedings of the AAAI-97 Workshop and AI Approaches to Fault Detection and Risk Management. Mento Park, CA: AAAI Press, 1997, pp. 9-13.

[13] Y. Wang, H. Yang, X. Wang, and R. Zhang, "Distributed Intrusion detections Based on data fusion method.", in Proceedings of the 5th World Congress on Intelligent Control and Automation, 2004, pp. 4331–4334.

[14] Q. Chen and U. Aickelin, "Anomaly detection using the Dempster-Shafer method," in Proc. of the 2006 International Conference on Data Mining, DMIN 2006, 2006, pp. 232–240.

Measuring the Impact of Different Metrics on Software Quality: a Case Study in the Open Source Domain

Valentino Sartori, Birhanu Mekuria Eshete, Adolfo Villafiorita
Fondazione Bruno Kessler
via Sommarive, 18
38123 Trento, Italy
valentino.sartori@tin.it, eshete@fbk.eu, adolfo.villafiorita@fbk.eu

***Abstract* — Knowledge about the expected impact of different project and technological choices is fundamental for project planning, resource allocation, and quality of the final software product. The latter property, in particular, is essential to gain users' trust and confidence. In this paper we present some preliminary results about a study we are conducting on open source web applications available in SourceForge. The ultimate goal is providing tools to support project managers and team in making choices that, being all other factors the same, increase the probability of delivering higher quality software products.**

***Keywords* - Software Robustness; Software Metrics; Software Quality; Project Metrics.**

I. INTRODUCTION

The last decade has seen a steady growth of web applications and services. Their popularity is due to many factors, among which we mention making content on the web easier to update (e.g., Content Management Systems), enabling forms of remote collaboration (e.g., with Wikis), delivering in a more efficient ways e-Government services to citizens (e.g., with web portals), and providing a new way to deploy and make available complex applications (e.g., Google Docs).

The development of web applications, however, is rather complex, since it nearly always requires the integration and harmonization of code written in different programming languages. A typical web application could, for instance, be written in PHP, use a MySQL database for storage, and deliver information to the user with pages written with HTML, CSS, and JavaScript. On top of that, the programming languages used to code the applications' logic (e.g., PHP, Ruby, Perl) do not have features, such as, for instance, type and range checking, that help programmers spot and correct errors before the application is deployed. As a result, several applications have vulnerabilities that could be exploited to, e.g., expose or steal sensitive data. Although one could claim that these problems could be mitigated using languages with more stringent syntax checks, such as Java, practical issues often make the development of web applications with these technologies unfeasible (e.g., lack of trained resources) or less attractive (the vast majority of service providers, for instance, do not offer deployment of Java applications).

Growing popularity of web applications and the flexibility granted by the different technological layers that can be used to deliver web applications have resulted in a rich array of different frameworks. We mention, as an example, ASP, C#, Java, Ruby, PHP, Perl, Python, and Javascript. When starting the development of a new web application, thus, the project team might be faced with the necessity of choosing one among the various frameworks and programming languages available for development. In such a scenario, knowledge about the impact of different technological choices could become a strategic tool to guide the selection of the technology to adopt.

This paper presents some preliminary results related to a study we have conducted on web applications made available by SourceForge. We are, in particular, interested in understanding relationship between some technological choices (e.g., the main programming language used for the development of the web application) and the quality of the corresponding product. To do so, we collected data and historical data about several applications under the "web application" category of SourceForge and tried to link the data to the quality of the product. We need to remark and emphasize that this paper is a first step toward a more systematic and complete analysis of the data. More in depth analyses, therefore, will be needed to further validate the interpretation of (some) data and (some) results we present here; the extension of the results to a wider class of applications and to a wider set of variables will help consolidate the assumptions made in this paper.

The paper is structured as follows. Section II presents the data sources and the data collection tools we used. Section III defines the goal of this work. The actual results are shown in Section IV, where we characterize the applications we have included in our study and Section V, where we show the data we have obtained. Finally, Section VI presents some related work and Section VII draws some conclusions.

II. SOURCEFORGE AND DATA COLLECTION

SourceForge [1] is a repository of open source software that provides tools for managing a software development project and distributing applications for free. Tools made available by SourceForge to support project teams include versioning and bug tracking systems, wikis, forums, and repositories to distribute different releases of a system. In the words of the owners, “SourceForge.net is the world's largest open source software development web site. As of August 2010 more than 240,000 software projects have been registered to use our services by more than 2.6 million registered users, making SourceForge.net the largest collection of open source tools and applications on the net.”

Given availability of data and number of projects, SourceForge is a great opportunity for researchers to analyze trends in (open source) software development. Matter of fact SourceForge has been used in the past for analyses and studies by several authors: we mention [2], [3], [4], [5], [6], and [7].

For various metrics, such as the number of downloads and team size, however, the data made available by SourceForge is the most recent value. Moreover extraction of the data requires to parse the HTML web pages of the SourceForge website. To simplify the analysis work, repositories containing dumps of the SourceForge database are available to researchers (see, e.g., [8] and [9]).

We used the SourceForge Research Data Archive (SRDA) [10]. The SRDA repository, available to registered users, provides a web form that allows to query a database containing monthly dumps of the SourceForge database. From SRDA one can get a vast amount of descriptive and statistical data about SourceForge projects and users [11]. Not all information is however made available by SRDA. In particular, no data about source code metrics, necessary for our work, was available when we performed the analyses.

To support our data collection needs, that requires downloading big amounts of data and integrating information from different sources, we developed a small system, whose architecture is depicted in Fig. 1. The left hand side of the picture shows the data sources we use, namely SourceForge (through the web pages made available on the Internet) and SRDA (through the web form made available to registered users). The right hand side of the picture shows the tools we developed:

- a *Parser*, written in Java and based on *Jtidy* [12] and on a DOM inspector, that we use to extract information from SourceForge’s web pages.
- A *Repo Client* that we use to automate calls to *cvs* and *svn*, to download the source code of the projects we analyze. The source code is analyzed using *cloc* [13], a tool to compute basic metrics about source code (size, expressed in lines of code and comments).
- A *database* (*Local DB*, in Fig. 1), which we use to locally integrate and store all the information we need. The database is then queried by users to perform analyses.

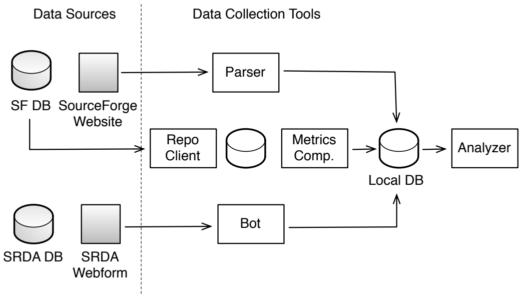


Figure 1. The system architecture

III. GOAL OF THE ANALYSIS AND DOMAIN

We performed analyses on the SourceForge database to achieve the following goals:

- Goal 1. Provide a characterization of Source Forge systems available in the “web application” category. The goals in this area include: understanding what categories of applications are most represented, what programming languages are used, and their growth over time; understanding whether there is a relationship between technology adopted and system size, measured both in terms of lines of code and in terms of Function Points [11].
- Goal 2. Provide a characterization of the quality of systems available in the “web application” category of SourceForge. Software quality is a difficult property to measure. Various approaches have been proposed emphasizing various dimensions, that include internal properties (e.g., software maintainability) and external properties (e.g., usability, reliability, availability, security). However, an assessment involving all these aspects can hardly be automated. We decided, therefore, to limit our attention to the number of bugs, their evolution over time, and the time taken by the project team to fix bugs. Since security bugs are of particular interests for web applications, we also distinguished and computed specific data for them.
- Goal 3. Highlight patterns between some of system characteristics (e.g., system size, programming language used) and software “quality” (in the sense of the previous goal).

The data we analyzed include all projects under the category “web based” which had released at least one version from January 2006 to May 2010 and for which there is at least one bug filed in the project’s bug tracking system. This screening is necessary to select projects which have had some active development. Various projects in SourceForge, in fact, are just “placeholders” for ideas that never get developed or that will be developed in the future.

IV. A CHARACTERIZATION OF SOURCEFORGE’S WEBAPPS

Fig. 2 and Fig. 3 provide a simple characterization of web applications in SourceForge. The data, collected from SRDA, spans from January 2006 (the first snapshot

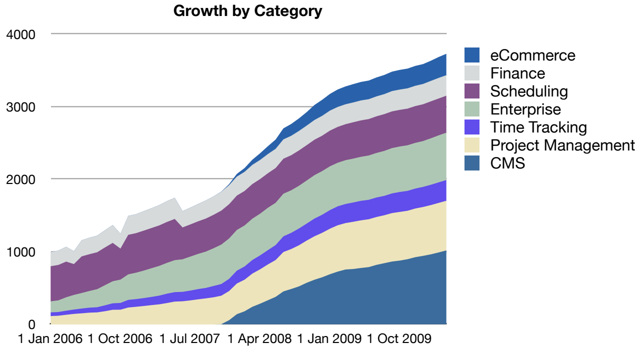


Figure 2. Growth of web applications by category

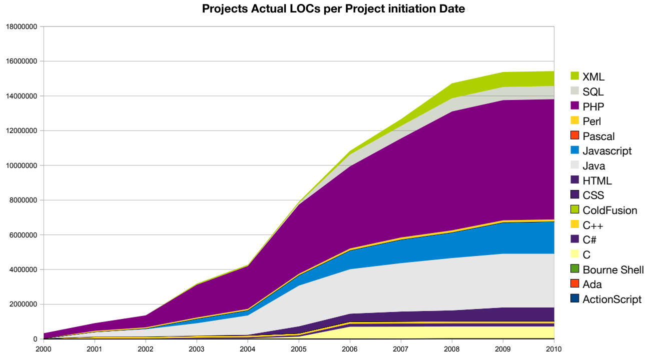


Figure 3. Source Lines of Code by Technology and Project

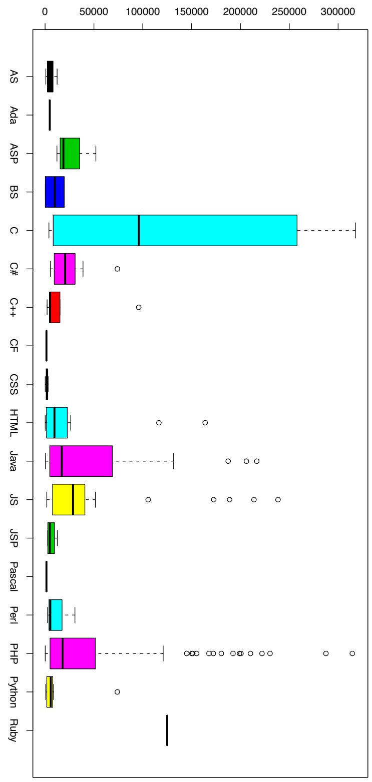


Figure 4. Source Lines of Code by Technology and Project

available) to May 2010 (the last snapshot available when we performed the analyses).

All projects in SourceForge are assigned by the project leader to one or more categories. The graph in Fig. 2 shows the growth, over time, of the categories under the “web based” umbrella. (Notice that the total sum in the graph is bigger than the actual number of projects, given the fact that one application can be assigned different categories.) Maybe not surprisingly the categories collecting the highest number of applications are CMS (Content Management Systems), Project Management, and Enterprise applications.

Fig. 2 weights all projects equally, independent from size and complexity. Applications in SourceForge, however, range from simple scripts no bigger than a few hundred lines of code to complex applications in the range of hundred of thousands of lines of code. A more accurate measure of growth, therefore should take into account also the size. This is shown in Fig. 3 where we measure, for each project, the size (in SLOC) as of May 2010 of the different technologies used in the projects. It has to be remarked that some of these technologies (e.g., CSS, XML) are not programming languages, although they can still be the “targets” of bugs and bug reports. The data of each project is shown on the date the project was started. Values accumulate over time. Thus, for instance, the data on the year 2000 (the first value on the x axis) shows the size (in SLOC) reached in May 2010 by all the projects that were started in that year. The data in 2001, as a second example, shows the size reached in May 2010 by all the projects started in 2001 together with those started in 2000. This explains the asymptotic nature of the graph, since projects started later will have had less time to evolve (and, hence grow).

The graph clearly shows that a wide variety of languages are used for developing web applications. Among them PHP is, by far, the programming language of choice, followed by Java, and Javascript. The graph omits some environments, such as Ruby and Python, which are scarcely represented in SourceForge. Notice also that the data refers to SourceForge only and that it might not be representative of the overall popularity of programming environments. Ruby on rails applications, for instance, have in RubyForge (a provider alternative to SourceForge) their repository of choice. Thus Ruby will tend to be underrepresented in SourceForge. Notice also that some of the programming languages represented in the figure, such as C and C++, come both from CGI-based web applications and from projects related to the development of desktop applications that, on the side, provide also some kind of web interface or service. Finally, we remark that the total number of lines of code we measured is about 16 million.

The technologies chosen to develop an application depend upon many factors, among which training and skills, legacy, and availability of libraries, to name a few. Fig. 3 shows the “popularity” of different technologies in SourceForge, but it does not tell us anything about whether there is a consistent usage of certain programming languages given some specific project characteristics, such as, for instance, system’s size. This is shown in Fig. 4, where, we measure the size of projects for each different programming language. Data is presented with a box plot, that allows us to show the median value (the bold vertical line in the box), the

TABLE I. DENSITY OF BUGS AND SECURITY BUGS PER (MAIN) PROGRAMMING LANGUAGE

Project's Main Programming Language	***Bugs per KSLOC***	***Security Bugs per KSLOC***
Javascript	0.2	0
SQL	0.5	0.1
JSP	0.7	0
Java	0.9	0
C#	1.4	0.1
Perl	2.2	0
C	3.4	0.4
C++	4.9	0
Bourne Shell	6.6	3
Action Script	8.1	0
PHP	16.6	7

TABLE II. TIME REQUIRED TO CLOSE A BUG

Project's Main Programming Language	***Average Time to close a Bug***	***Average Time to close a Security Bug***
Javascript	74	46
SQL	54	8
JSP	9	7
Java	61	76
C#	31	3
Perl	156	201
C	322	25
C++	149	2
Bourne Shell	15	16
Action Script	48	90
PHP	76	124

are where the majority of the population lies (the box), the minimum and the maximum values (the "T"s at the extremes of the box plot). By looking at the diagram, C seems to be the most "flexible" language, since it appears in a wide variety of projects, ranging from small applications to systems in the range of 300K SLOC. Java and PHP are closely related, with similar patterns with respect to the size of projects in which these programming languages are used. In both cases the vast majority of applications written in PHP or Java is below 100K SLOCs, with PHP being the language with the most exceptions.

V. WHERE ARE THE BUGS?

Projects in SourceForge can use the SourceForge's bug tracking system to maintain track of the bugs discovered during development or usage of the system. Like many other similar systems, the bug tracker in SourceForge allows one to assign a description, a priority, a status, a category, a person responsible for the resolution, among other things. In the second part of our work we tried to correlate information about bugs and, more specifically, security bugs, with the technologies used to develop an application. We had, however, to face the following two issues related to data quality and availability:

1. Not all fields in SourceForge's bug tracker are compulsory and many projects do not file information about the category of the bug. To distinguish between security-related and non-security related bugs, when the category is not available we used a simple classification algorithm that measures the presence of specific words in the bug description. In particular, if the bug description contains some (key)words typically associated to security problems, we classify the bug as a security bug. The (key)words include, for instance: login, logout, session, phishing, penetration. The set of keywords is synthesized based on Sans Security Terms Glossary [14]. The approach is similar to [18]. See Section VI for more details.
2. The information about the file in which a bug is located is not available in the bug tracking system. Moreover, commit messages in several cases do not mention the bug they fix. As a result it is often impossible to assign a bug to a specific file and, hence, to a specific programming language. All projects, however, have a main programming language. To allocate bugs to a specific technology, therefore, we made the (strong) hypothesis that all bugs reported in a project refer to the main programming language used. Thus, for each project, we identified the main programming language (that is the programming language with the greatest number of SLOCs) and assumed all bugs referred to it.

The results of the analyses are shown in Table 1, where we report the number of bugs per thousand lines of code. Some software engineers estimate the defect density of well-written code to be between 3 and 6 per thousand lines of code [15]. Our data shows quite a few values outside the predicted range. Although one explanation could be that we refine the work in [15], a more likely explanation is due to the "noise" in our data (not all bugs refer to the main programming language) and to the great variety of applications hosted by SourceForge (which include both high and low quality software). That said, the table seems to show that Java, a language with a rather strict syntax, shows a lower density of bugs than languages with a relaxed constructs, such as PHP and Bourne Shell. There are some exceptions: Javascript seems to perform better than Java; C++, in spite of being object oriented, worse than C.

Table 2, finally, reports the average time required to close both non-security and security related bugs. The table shows the elapsed time and not the actual effort spent on fixing the bug. Thus the values in the table should be interpreted more as the combination of priority and complexity, rather than a simple measure of complexity.

Also in this case results are not definitive. The table shows that security bugs tend to be fixed in a shorter time for some technologies, but not for all of them. The interpretation is not clear: the most likely reasons could include complexity in fixing certain security bugs and some differences in the data available (e.g., the huge difference between the time take to fix bugs and security bugs in C++ could be due to the fact that there are few projects written in C++). Further analysis is needed.

VI. Related Work

In [16] the authors analyze the correlation among different object-oriented metrics. The goal is identifying dependent metrics to reduce the burden of metrics computation and to define statistically significant quality threshold for Java software. The analysis, conducted on 146 open source Java projects downloadable from SourceForge, for a total amount of over 70,000 classes and over 11 million lines of code. The author show a strong correlation among metrics in five different cases and, in the process, identify actual ranges of values for several metrics. Our work, by contrast, focuses on correlation between project choices (such as the programming language) and the corresponding quality of the end system.

In [17] the authors examine the code base of the OpenBSD operating system to determine whether its security is increasing over time. They do so by measuring the rate at which new code has been introduced and the rate at which vulnerabilities have been reported over the last 7.5 years and fifteen versions. Some of the questions the authors try to answer include aspects related to whether legacy code influences security and whether software and software development practices are leading to the development of more secure software. The authors show that the majority of security bugs are in foundational code (that is code released with the first versions of a system).

In [18] the authors use text-mining techniques to classify some bugs as security bugs. The results of the classification is then validated with software engineers yielding a 77% of correct classification. Our method for the classification of security bugs is inspired by that of the authors, although simpler in scope and lacking the manual validation phase.

Finally, in [19] the authors report on data collected during corrective maintenance and refactoring of a complex system to improve software quality. In the case of [19] the association between bugs fixed and changes to the code was possible due to the practices adopted by the development team, that required to state in the commit messages that issues being addressed.

VII. Conclusions

Knowledge about the expected impact of different project and technological choices is fundamental for project planning, resource allocation, and quality of the final software product. Open Source Repositories, such as SourceForge, not only deliver high-value services to support teams and individuals interested in open source development, but they also provide a wealth of information about software projects and development practices.

In this paper we have presented a study we conducted to understand whether some simple technological choices, such as the programming language adopted to develop an application, provide a clear advantage to control the complexity of development and increase a system's quality. We chose to analyze web applications hosted by SourceForge. The choice was made for various reasons, among which complexity of web application development and the wide choice of technologies to develop them. To understand whether some technologies consistently outperform others, we used some crude indicators, such as the density of bugs and the time required to fix bugs. The results we got, in our opinion, provide some preliminary insights.

Further work is needed to consolidate the results presented in this paper. The directions include: the input domain, the interpretation of some metrics, and the consolidation of the analyses. Concerning the first point, two obvious areas of improvement are the enlargement to a wider set of applications (e.g. by including other repositories, such as RubyForge) and the reduction of "noise" from the data (e.g., removal of projects that are not active). Concerning the second point (the interpretation of some metrics), we could extend analyses to other metrics usually related to software quality (e.g., inner quality metrics). Concerning the third point, more work is needed to systematically analyze the correlation among the different variables characterizing (SourceForge) projects.

These are some of the necessary steps to build a solid ground upon which we could eventually come with a set of rules of the thumb to guide technological choices to increase the quality of software artifacts.

References

[1] SourceForge website, Available at http://sourceforge.net. Last accessed December 20, 2010.

[2] English, R. and Schweik, C. M. "Identifying success and tragedy of floss commons: A preliminary classification of sourceforge.net projects". In *FLOSS '07: Proceedings of the First International Workshop on Emerging Trends in FLOSS Research and Development* (Washington, DC, USA, 2007), IEEE Computer Society, p. 11.

[3] Grechanik, M., McMillan, C., DeFerrari, L., Comi, M., Crespi, S., Poshyvanyk, D., Fu, C., Xie, Q., and Ghezzi, C. "An empirical investigation into a large-scale java open source code repository". In *ESEM '10: Proceedings of the 2010 ACM-IEEE International Symposium on Empirical Software Engineering and Measurement* (New York, NY, USA, 2010), ACM, pp. 1–10.

[4] Li, Y., Tan, C.-H., Teo, H.-H., and Mattar, A. T. "Motivating open source software developers: influence of transformational and transactional leaderships". In *SIGMIS CPR '06: Proceedings of the 2006 ACM SIGMIS CPR Conference on Computer Personnel Research* (New York, NY, USA, 2006), ACM, pp. 34–43.

[5] Robles, G., and Gonzalez-Barahona, J. M. "Geographic location of developers at sourceforge". In *MSR '06: Proceedings of the 2006 International Workshop on Mining Software Repositories* (New York, NY, USA, 2006), ACM, pp. 144–150.

[6] Van Antwerp, M., and Madey, G. "The importance of social network structure in the open source software developer community". In *HICSS '10: Proceedings of the 2010 43rd Hawaii International Conference on System Sciences* (Washington, DC, USA, 2010), IEEE Computer Society, pp. 1–10.

[7] Gao, Y. and Madey, G. R. “Towards understanding: a study of the SourceForge.net community using modeling and simulation”. In *SpringSim* (2) (2007), M. J. Ades, Ed., SCS/ACM, pp. 145–150.

[8] J. Howison, M. Conklin, and K. Crowston. “Flossmole: A collaborative repository for FLOSS research data and analyses”. In *International Journal of Information Technology and Web Engineering*, 1(3), pp 17–26, 2006.

[9] C. Daffara and J. Gonzalez-Barahona. “Flossmetrics Project”, 2007. Available at http://www.flossmetrics.org/. Last accessed December 20, 2010.

[10] Van Antwerp, M. and Madey, G., “Advances in the SourceForge Research Data Archive (SRDA)”, The *4th International Conference on Open Source Systems - (WoPDaSD 2008)*, Milan, Italy, September 2008. Also available at http://www.nd.edu/~oss/Papers/srda_final.pdf, last accessed December 20, 2010.

[11] Albrecht, A. J., “Measuring Application Development Productivity,” *Proceedings of the Joint SHARE, GUIDE, and IBM Application Development Symposium*, Monterey, California, October 14–17, IBM Corporation (1979), pp. 83–92.

[12] JTidy - An HTML Parser and Pretty Printer in Java. Available at http://jtidy.sourceforge.net/howto.html. Last accessed December 20th, 2010.

[13] CLOC - Count Lines of Code. Available at http://cloc.sourceforge.net/. Last accessed December 20, 2010.

[14] Glossary of security terms. Available at: http://www.sans.org/security-resources/glossary-of-terms/. Last accessed December 20, 2010.

[15] Hatton, L. “Re-examining the fault density - component size connection”. *IEEE Software* 14, 2 (1997), pp. 89–97

[16] Barkmann, H., Lincke, R., and Lowe, W. “Quantitative evaluation of software quality metrics in open-source projects”. In *WAINA ’09: Proceedings of the 2009 International Conference on Advanced Information Networking and Applications Workshops* (Washington, DC, USA, 2009), IEEE Computer Society, pp. 1067–1072.

[17] Ozment, A. and Schechter, S., “Milk or Wine: Does Software Security Improve with Age?" *Proceedings of the 15th Usenix Security Symposium*, *Usenix*, 2006, pp. 93–104.

[18] Gegick, M., Rotella, P., and Xie, T. “Identifying security bug reports via text mining: An industrial case study”. In *Proceedings of the 7th International Working Conference on Mining Software Repositories, MSR 2010* (Co-located with ICSE), Cape Town, South Africa, May 2-3, 2010, pp. 11–20.

[19] Longo, F., Tiella, R., Tonella, P., and Villafiorita, A. “Measuring the impact of different categories of software evolution”. In *Software Process and Product Measurement, International Conferences: IWSM 2008, Metrikon 2008, and Mensura 2008*, Munich, Germany, November 18-19, 2008. Proceedings (2008), vol. 5338 of Lecture Notes in Computer Science, Springer, pp. 344–351.

Software Platform Architecture for Ubiquitous City Management

Kyung-Won Nam
Fusion Technology Team, R&D Division, LG CNS
Seoul, Korea
lenny@lgcns.com

Jin-Su Park
Fusion Technology Team, R&D Division, LG CNS
Seoul, Korea
jsupark@lgcns.com

***Abstract*— U-City (Ubiquitous City), which is a city or a region with ubiquitous information technology, has recently emerged as an important government initiative for urban management in Republic of Korea. There are various kind of information related to the u-city and these information can be gathered through technologies such as wireless networking, RFID (Radio Frequency Identification) tags, CCTV (closed-circuit television) and sensors. In addition, information from existing legacy systems (e.g. Intelligent Transport System, Urban Facility Management System) can be related to the previously mentioned information. In the concept of U-City, U-services (Ubiquitous urban services) like U-Facility, U-Transport, U-Security and U-Environment are defined as services which process a large number of multiple events occurred in urban areas by using these information. In the past years, Events or information from U-services has been separately managed causing an increase in system complexity, overlapping investment costs and so on. Also, due to the lack of suitable urban management system, there has been difficulty in integrating with other urban management solutions. In terms of urban management, there must be a system that can store and process the massive amount of data collected from U-services. This paper discusses the flexible and extensible software platform architecture to manage these ubiquitous information more effectively.**

Keywords-Ubiquitous City; Ubiquitous Urban Services; Software Platform; Flexible and Extensible Architecture; Ubiquitous Information

I. INTRODUCTION

In the past, there have been problems like the regional unbalance, environmental pollution and so on among cities [3]. As the search has been on for alternative city management to these problems, paradigm for city management has shifted from the conventional urban management system to the new intelligent management system based on the ubiquitous technologies [5]. With the rapid growth of ubiquitous computing systems, it is possible to provide citizen with the information services for urban situation at any time and monitor situation around cities in real time [4]. In the Republic of Korea, the word “U-City (Ubiquitous City)” refers to a city or region with ubiquitous information technologies, as shown in Figure 1. All information systems are linked, and virtually everything is linked to an information system through technologies such as wireless network, RFID (Radio Frequency Identification) tags or readers, CCTV (closed-circuit television) cameras. Using these information and technologies, it is also possible to develop cities and the quality of life among citizen. Thus, adoption of the new software platform architecture for urban management system is necessary to manage various kind of information from U-city effectively.

The following sections of this paper discuss the U-City infrastructure in South Korea and its components. This paper then introduces the flexible and extensible software platform architecture.

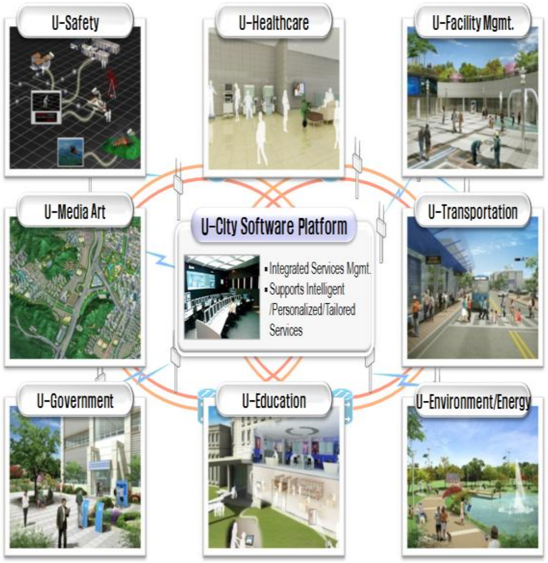


Figure 1. The concept of ubiquitous city

This paper is organized as follows: Section Ⅱ describes the U-City Infrastructure in South Korea. Then in Section Ⅲ, the design principles of software platform architecture is presented. In Section Ⅳ, we suggest the software platform architecture for urban management system. Finally Section Ⅴ concludes the paper and discusses future work.

II. U-CITY INFRASTRUCTURE

Figure 2 shows the U-City IT framework which consists of the followings:

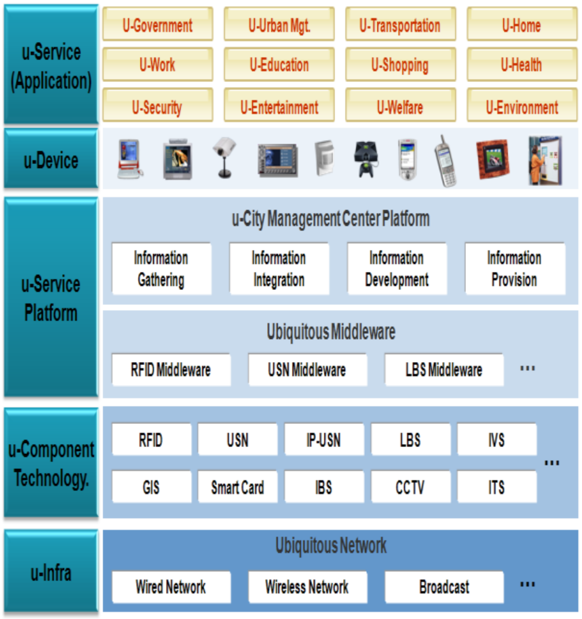


Figure 2. **U-City IT** framework

- U-Service is the application set which provide citizen with a variety of urban-related services. U-Safety enables IP based broadband integrated surveillance, automatic object recognition and tracking using pattern recognition technology. U-Health is a medical data sharing system between organizations and clients integrating separated information hospital projects. U-Facility is a real-time urban monitoring and remote controlling system based on 2D or 3D GIS map. U-Transportation is tailored/real time services for customers establishing national integrated transportation information center. U-Government, which is paperless government, civil affair services, is integrated government services focused to citizens. U-Education is personalized/tailored learning services enabling decreasing expenses on private education. Learning everywhere is possible by providing mobile devices and network infra. U-Education is increasing confidence of public education with intelligent and systemic education.
- U-Device refers to a device which presents information and data collected from U-Services by providing user interface or a device which creates information and data like CCTV cameras, sensors. Wired/wireless U-Device is combined with each other to make ubiquitous computing available for the citizen
- U-Service platform makes each U-Service to interact with U-devices easily hiding the details of a particular set of functionality.
- The most widely used technologies like RFID (Radio Frequency Identification), USN (Ubiquitous Sensor Network), GIS (Geographic Information System), GPS (Global Positioning System) are grouped into U-Component Technology including 3rd party solutions.

III. DESIGN PRINCIPLES OF SOFTWARE PLATFORM ARCHITECTURE

As mentioned above, the massive amount of data can be collected from U-Services or many kinds of ubiquitous devices. Therefore, a suitable urban management system based on layered architecture is necessary to process those data. For the effectiveness of urban management, adoption of flexible and extensible software platform architecture in urban management systems is emerging as a major issue lately.

The flexible and extensible software platform architecture is required to meet the following requirements:

- standard interface to various kinds of U-services,
- standard interface to heterogeneous ubiquitous devices using multiple protocol,
- creating and processing business events through collected data,
- interface to other relevant external systems,
- interface support for compatibility with third party solutions,
- user interface for monitoring data collected from U-services and devices,
- guaranteeing extensibility for processing large amount of data,
- standard security policy for complex interface,

We focus on two key aspects of design principles when adopting architecture for urban management system.

A. *Distributed and loosely-coupled System*

By adopting distributed and loosely-coupled system, it is possible to distribute data and traffic load [1]. Each component can easily be altered to accommodate changes in system capability and system requirements [1][6]. When the some aspect of the system is scaled to a larger size, it can operate correctly.

B. *Service Oriented Architecture*

SOA (Service Oriented Architecture) is a flexible set of design principles used during the phase of systems development and integration in computing. A system based on SOA architecture will provide a loosely-coupled suite of services that can be used within multiple separate systems from several business domains. SOA advocates an approach in which a software component provides its functionality as a service that can be leveraged by other software components. SOA allows the integration of existing systems, applications and users into a flexible architecture that can easily accommodate changing needs It can also provide various kinds of applications and other different platform with standard interface. There are lots of advantages including the ability to monitor and track the transaction among systems in terms of service oriented architecture based on ESB (Enterprise service Bus).

An ESB is an architectural pattern and a key enabler in implementing the infrastructure for a service oriented architecture. The increasing adoption of SOA and the proliferation of Web services have revealed an ever growing need to provide a managed layer between services and their consumers.

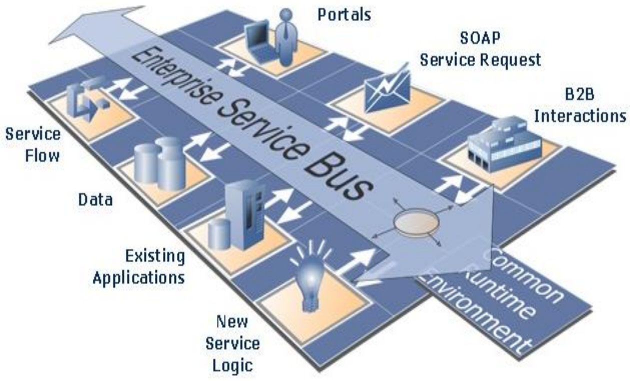


Figure 3. The Enterprise Service Bus (Mahesh H. Dodani, "From Objects to Services: A Journey in Search of Component Reuse Nirvana", IBM Software)

IV. Software Platform Architecture For Urban Management System

A. *Layer Design of Software Platform Architecture*

A multi-layer system is using different layers for allocating the responsibilities of an application. It helps to structure applications that can be decomposed into groups of subtasks in which each group of subtasks is at a particular level of abstraction. It has several benefits than other existing system implementing its layers as a monolithic block. In multi-layer systems, each layer can be reused if it has well-defined interface. And also, changes in each layer affect only one layer. Thus, developers can adapt affected layers without altering the remaining layers.

Considering the advantage of a multi-layer system mentioned above, adopting multi-layer system is necessary to improve system performance and reduce system complexity because the urban management system has to process large amount of data gathered all around the city. Platform-independent technologies can be provided through this layered architecture and also loosely-coupled system is needed to meet the requirements of many applications.

As shown in Figure 4, in this section, we introduce layer design of software platform architecture. The architecture consists of 3-level layer based on distributed system and each layer interacts with other layer by standard interface like web services. Each layer consists of stand-alone components which have similar responsibility resulting in developing each layer easily and load balancing.

1) U-Infra Abstraction Layer: This layer collects data through interaction with sensing collection server like RFID middleware, USN (Ubiquitous Sensor Network) middleware, media board and CCTV server. It provides standard APIs (Application Programming Interface) to control U-devices (ubiquitous devices). For example, it receives response from sensors after sending request for controlling sensors.

2) Platform Service Layer: This layer stores and processes data collected after sending commands to U-Infra Abstraction Layer. On the request from U-Service Integration Layer, it processes data and events to U-Service Integration Layer.

3) U-Service Integration Layer: U-Services interacts with other layers through U-Service Integration Layer. This layer proposes internal/external interface standard.

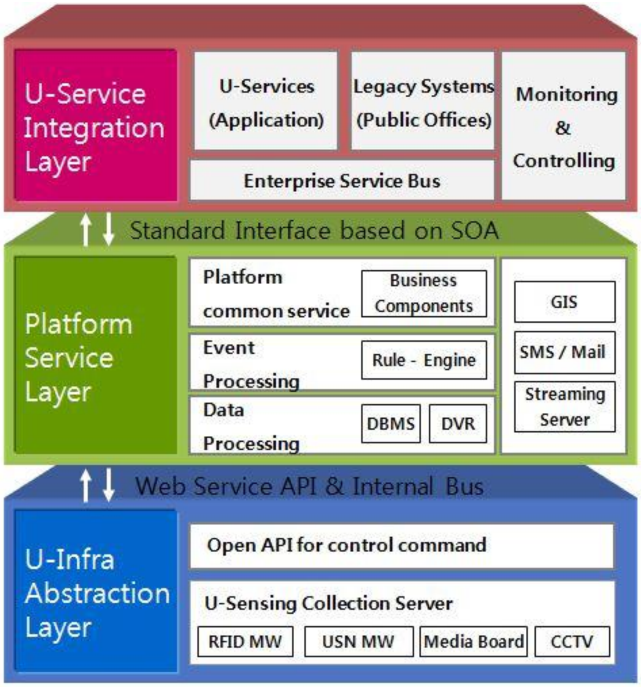


Figure 4. Layer design of software platform architecture

Each layer can be installed independently or integrated into a single machine depending on the capacity of the system. The software components belonging to one layer communicate with the other layers in a way of pre-defined interfaces. For example, a component of Platform Service Layer can call Web Service APIs of the U-Infra Abstraction Layer, which displays alert text message of the possibility of flooding through the media board (also known as digital signage) around the street. In the same manner, USN Middleware of U-Infra Abstraction Layer can send the water level information to the subscribers (U-Service application or Monitoring System) using the Internal Bus that can be implemented with Message Oriented Middleware.

V. CONCLUSION AND FUTURE WORK

In Korea, most of urban development project recently established is being promoted by national U-City strategy. With increasing number of U-City construction projects, it is expected that the total number of population residing in U-Cities in 2015 will reach at about 2.3 million. The U-City act proposed by the Korean Ministry of Land, Transport and Maritime Affairs was enacted through discussion with related departments and public hearing. As the demand for U-City rises, the demand for the robust software system which is capable of managing information and the devices city-wide will become a greater priority.

In this paper, we proposed software architecture for U-City information system but more work is needed. We need to find and develop new service models for U-City residents. Standardization of the software platform would be an issue for government to prevent overlapped investment. More than all, the work should be for the convenient, safe, and pleasant life of citizens.

ACKNOWLEDGMENT

This research was supported by a grant (07High Tech A01) from High tech Urban Development Program funded by Ministry of Land, Transportation and Maritime Affairs of Korean government.

REFERENCES

[1] Birman, K., Reliable Distributed Systems:Technologies, Web Services and Applications., New York: Springer-Verlag, 2005.

[2] Mahesh Dodani, *From Objects to Services: A Journey in Search of Component Reuse Nirvana*, in Journal of Object Technology, vol. 3, September-October 2004, pp. 49-54.

[3] D. H. Shin, "Ubiquitous city:Urban technologies, urban infrastructure and urban informatics," *Journal of Information Science 2009*, vol. 35, pp. 515-526.

[4] Vassilis Kostakos, T Nicolai, E Yoneki, Eamonn O'Neill, H Kenn, and J. Crowcroft, "Understanding and measuring the urban pervasive infrastructure" Personal and Ubiquitous Computing 2010, vol 13, pp. 355-364.

[5] Lee, S. H, Yigitcanlar, Tan, Han, Jung-Hoon, Leem, and Youn-Taik (2008), "*Ubiquitous urban infrastructure : Infrastructure planning and development in Korea*," Innovation: Management, Policy and Practice, 10(2/3). pp. 282-292.

[6] M. van Steen, F. J. Hauck, and A. S. Tanenbaum. "A Model for Worldwide Tracking of Distributed Objects." In Proc. TINA'96 Conference, Heidelberg (Germany), Sept. 1996.

[7] W. T. Tsai, Miroslaw Malek, Yinong Chen, and Farokh Bastani. "Perspectives on Service-Oriented Computing and Service-Oriented System Engineering" The 2nd IEEE International Workshop on ServiceOriented System Engineering SOSE 06 (2006). pp3-10.

PROMETHEUS: a web platform for supporting knowledge management in an environment based on experience factory

Pasquale Ardimento, Nicola Boffoli, Vito Nicola Convertini, Giuseppe Visaggio
Department of Informatics
University of Bari Aldo Moro
Bari, Italy
ardimento@di.uniba.it, boffoli@di.uniba.it, convertini@di.uniba.it, visaggio@di.uniba.it

***Abstract* — This paper presents a Knowledge Management System (KMS), called PROMETHEUS, which consists of a set of processes that constitute the Experience Factory (EF) and a platform that is the Knowledge Experience Base (KEB), which collects Knowledge Experience Packages (KEP). The KMS thus formed supports the formalization and packaging of knowledge and experience of producers and innovation transferors encouraging gradual explanation of tacit information of bearers of knowledge to facilitate the transfer. The KMS enables the cooperative production of KEP between different authors contributing to the production of KEP and users of the latter. The paper describes the approach outlined in the PROMETHEUS Project and the precautions taken in the design of KEP to ensure that: the experience contained in it, even when collected through projects executed by many person-years, can be quickly acquired by the user, contains the tools to facilitate the acquisition of knowledge innovation support to transfer.**

Keywords - Business mode; Experience Factory; Knowledge Management

I. Introduction

The knowledge of its software engineers and developers is the most relevant asset of a software company. However, handling such knowledge properly is a complex task. Several studies and experimentations have been conducted on how to share and increase such knowledge. “Among them there is the ground breaking work of Basili on the experience factory” [27]. In this paper, we present a knowledge experience factory, called PROMETHEUS, to enact a knowledge management system within a software company. The framework is made up of four major sections: Contents, Attributes, Educational and Training E-Learning, Taxonomy.

The PROMETHEUS (Practices Process and Methods Evolution Through Experience Unfolded Systematically) Project [1], [2], [21], [22], [23], [24], [26] is a model of Experience Factory (EF) to collect experimental knowledge in a repository Knowledge Experience Base (KEB) in the form of Knowledge Experience Package (KEP). The KEP is the vehicle suggested for the transfer of knowledge while the EF is the set of processes that make the Open Innovation.

This paper describes the structure of the KEP and the features that make the contents to be tailored and attractive for the target of the innovation.

The rest of the paper is structured as follows: the next section discusses related works and research activities; third section presents the major concepts implemented in PROMETHEUS, 4th section describes the incremental production of KEPs. Finally, in the conclusions some observations are made about PROMETHEUS and possible future research pathways are identified.

II. Related Works

The aim of experience factory [28] is to provide an infrastructure that supports project developments by analyzing and synthesizing all kinds of experience, acting as a repository for such experience, and supplying that experience to various projects on demand. Introduced in late eighties, the concept of experience factory has been implemented in many organizations [8], [10], [11], [15]. Unfortunately, there is still no exact well established technique that would lead to a guaranteed success in adopting the concept of experience factory in a company and to this regard the human factor is pointed out as the main cause [28].

Our approach focuses on a knowledge base whose contents make it easier to achieve knowledge transfer among research centres; between research centres and production processes; among production processes. The knowledge base must be hybrid, public, as we wish, or private, depending on KEP authors preferences. The public KEB allows one or more interested communities also included public administrations, to develop around it and exchange knowledge.

III. PROMETHEUS

The Authors use the term knowledge package to refer to an organized set of: knowledge content, teaching units on the use of the demonstration prototypes or tools and all other information that may strengthen the package’s ability to achieve the proposed goal. The KEP must be usable independently of its author or authors and for this purpose the content must have a particular structure: distance education and training must be available through an e-learning system. In short, the proposed knowledge package contains knowledge content integrated with an e-learning function.

In PROMETHEUS, the KEP must include all the components shown in Figure 1. A user can access one of the package components and then navigate along all the components of the same package according to her/his needs.

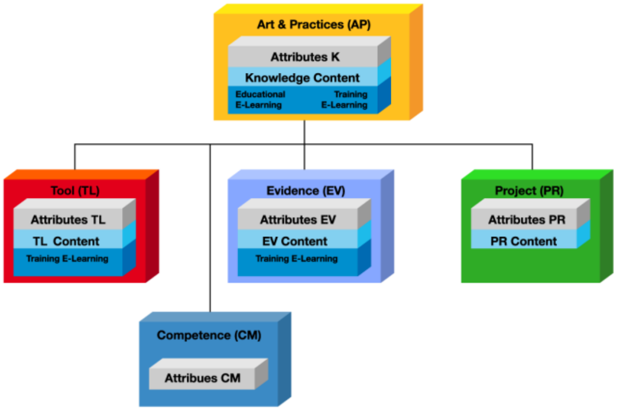


Figure 1. Diagram of a Knowledge/Experience package

The KEP does not contain the conceptual basis of the subject, because it is considered as the background of the user's knowledge, and can be found in conventional sources of knowledge such as technical reports, papers and books. Anyway, when users need some of the basic concepts for understanding the contents of KEP they can use educational e-learning course. And if users should need more information, they can use the "attachments" regarding reports, papers and books about basic topics of KEP. Instead, if the use of a demonstrational prototype is required to become operational, the same package will point to a training in e-learning course. As stated above, the use of these courses is flexible, to meet individual user's needs.

When a package also has support tools, rather than merely demonstration prototypes, Knowledge Content (KC) links the user to the available tool. For the sake of clarity, we point out that this is the case when the knowledge package has become an industrial practice, so that the demonstration prototypes included in the archetype they derived from have become industrial tools. The tools are collected in the Tools Component (TL). Each tool available is associated to an educational course, again of a flexible nature, in the use of the correlated training e-learning course.

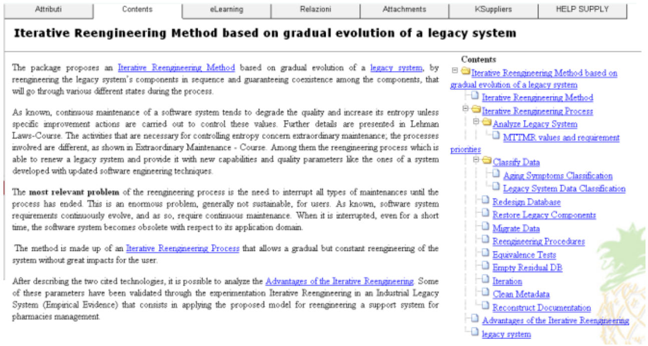
Should the user need support from whom has knowledge of the contents of KEP, a list of resources is a reference. The list is collected in the Competence component (CM).

A. Knowledge Content

It can be seen in the Figure 1 that the Art & Practices KC is the central one. It contains the KEP expressed in a hypermedia form in order to include figures, graphs, formulas and whatever else may help to understand the content. The KC is organized as a tree that starting from the root (level 0) descent to the lower levels ($level_1$, $level_2$, …, $level_n$) through pointers (Figure 2). The higher the level of a node the lower the abstraction of the content, which focuses more and more on operative elements. The root node of KC is made up of the following sections:

- Thoughtful Index: tells the reader how the package suggested will practically change, with a list of processes and activities, case the whole process is not innovating or has to be modified.
- Problem (one or more): describes the problem of KEP. A problem may belong to one of the two following types: decision and optimization. If the problem is the decision there should be the possibility to make a choice, and the aim for this choice. If the problem is optimization, the resources you want to improve the performance and the objective function of optimization have to be indicated. For each problem, the context has to be defined, that is to say all facts and circumstances which cause and condition a certain problem.

The leaf nodes have the answers to the problems: the solution or solutions suggested for each problem set. Figure 2 shows an example of KC.



Iterative Reengineering Method based on gradual evolution of a legacy system

The package proposes an Iterative Reengineering Method based on gradual evolution of a legacy system, by reengineering the legacy system's components in sequence and guaranteeing coexistence among the components, that will go through various different states during the process.

As known, continuous maintenance of a software system tends to degrade the quality and increase its entropy unless specific improvement actions are carried out to control these values. Further details are presented in Lehman Laws-Course. The activities that are necessary for controlling entropy concern extraordinary maintenance; the processes involved are different, as shown in Extraordinary Maintenance - Course. Among them the reengineering process which is able to renew a legacy system and provide it with new capabilities and quality parameters like the ones of a system developed with updated software engineering techniques.

The **most relevant problem** of the reengineering process is the need to interrupt all types of maintenances until the process has ended. This is an enormous problem, generally not sustainable, for users. As known, software system requirements continuously evolve, and as so, require continuous maintenance. When it is interrupted, even for a short time, the software system becomes obsolete with respect to its application domain.

The method is made up of an Iterative Reengineering Process that allows a gradual but constant reengineering of the system without great impacts for the user.

After describing the two cited technologies, it is possible to analyze the Advantages of the Iterative Reengineering. Some of these parameters have been validated through the experimentation Iterative Reengineering in an Industrial Legacy System (Empirical Evidence) that consists in applying the proposed model for reengineering a support system for pharmacies management.

Contents

- Iterative Reengineering Method based on gradual evolution of a legacy system
  - Iterative Reengineering Method
  - Iterative Reengineering Process
    - Analyze Legacy System
      - MTTMR values and requirement priorities
    - Classify Data
      - Aging Symptoms Classification
      - Legacy System Data Classification
    - Redesign Database
    - Restore Legacy Components
    - Migrate Data
    - Reengineering Procedures
    - Equivalence Tests
    - Empty Residual DB
    - Iteration
    - Clean Metadata
    - Reconstruct Documentation
  - Advantages of the Iterative Reengineering
  - legacy system

Figure 2. Sample of Knowledge Content of a KEP

To ensure control of completeness and lack of ambiguity in the contents of KEP, the vocabulary of KEP, i.e. concepts and relations between there meanings, has been formalized by the W3C XML Schema [20], in short XSD to obtain for each KC the following advantages:

1. The full list of concepts (elements) which have to be declared with obligatorily, multiplicity and default values of the elements / concepts, relationships between elements / concepts, type of elements, attributes defined for each element, type of attributes, ...;
2. Elimination of ambiguity, incompleteness, verbosity due to Informal definitions;
3. Verification of the correct syntax ;
4. Interoperability of the KEP, at the syntactic level between background of experience that share the structure proposed by us, leading to an independence of the software that produces them.

The research results integrated by a KEP may be contained within the same knowledge base or derive from other knowledge bases or other laboratories. If the knowledge package being read uses knowledge packages located in the same experience base, the relations will be explicitly highlighted. In Figure 3 is graphically shown the KEP structure.

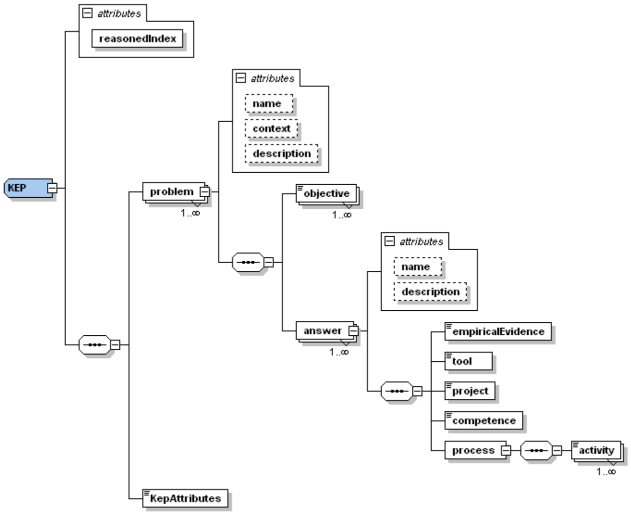


Figure 3. Sample of content of a Knowledge/Experience package

B. *Attributes*

Search inside the package starting from any of its components is facilitated by the component's Attributes. As shown in Figure 1, each component in the knowledge package has its own attributes structure. For all the components, attributes allow rapid selection of the relative elements in the knowledge base. Attributes have already been defined in [25], [2]. To facilitate the research, we used a set of selection classifiers and a set of descriptors summarizing the contents. The summary descriptors include: a brief summary of the content and a history of the essential events occurring during the life cycle of the package, giving the reader an idea of how it has been applied, improved, and how mature it is. The history may also include information telling the reader that the content of all or some parts of the package are currently undergoing improvements.

The classifiers include:

- The adoption risks of the technological innovation where it is provider;
- The mitigation initiatives of risk that assure a better performance of the KEP in the solution of main problems
- The impact that the KEP will have on the active processes of the production lines where it will be applied, supposed that the problems to solve correspond to the ones in the KEP
- A forecast of the Return of Investment that the new introduction will have in the company. For this reason the economical impact of the KEP as well as its impact on the value chain are specified;
- The acquisition plan of the methods of the KEP;
- The history of the KEP, i.e. the set of practices that have required the use and the results following to their application in order to assure a higher perception of reuse of the KEP.

The interested reader can find further details on the contents of the KEP and the management and use of KEB on the technical report [18].

C. *Educational and Training E-Learning*

PROMETHEUS, as shown in Figure 1, is made up of several sections for each component provided by the structure of the KEP. The division into sections enables beneficiaries to cut, and then adapt the learning to their training needs. In the interface for access to each section of PROMETHEUS there are links to resources and relationships of the component selected with the rest of KEP.

In order to support the beneficiaries for the acquisition of the KEP, PROMETHEUS helps them select a training program best suited to their knowledge. PROMETHEUS, in fact, provides for each KEP, in addition to training materials and training, tests to assess skills for the adoption of KEP by companies, research and government institutions interested in them. If tests were to detect any skill gaps, models are used to fill them, suggesting the more appropriate decision.

For each competence PROMETHEUS, in fact, provides a set of learning units. Each learning unit aims to train the user of a KEP on one or more items of the KEP of interest. Therefore, it is possible to attach to each teaching unit a test plan to verify that the user already has, or has acquired, the corresponding part of competence.

Such a model predicts that each jurisdiction has an associated evaluation questionnaire and a decision model. In the questionnaires the self-assessment tests result in the user evaluation of the KEP and guidelines to improve the training of the user. The test assesses the competencies and skills attained by the user and the gap between them and the ones expected. The model of decision interprets the level of acquisition of the skills of users receiving the evaluation questionnaires and suggests actions to be undertaken to fill any gap between skills expected and skills acquired by the learner. The model decision is made by the decision tables.

Operationally, for each competency $C_{(i)}$ a specific evaluation model is planned (Figure 4). In this model, the responses gathered by $QC_{(i)}$ the evaluation questionnaires provided, are interpreted by an appropriate set of decision tables DT (Figure 4). There is more than one DT where each DT aims to interpret the answers of the corresponding teaching unit UD.

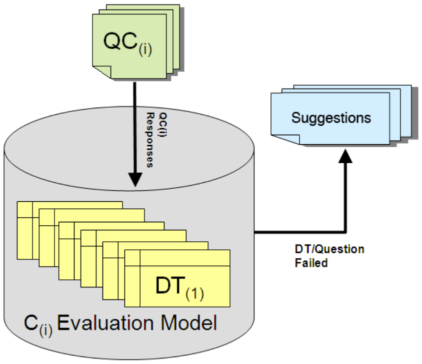


Figure 4. Evaluation Model

D. PROMETHEUS Taxonomy

Taxonomy is the practice and science of classification according to natural relationships. In PROMETHEUS, taxonomy is used by administrators to organize content and is created from 'Vocabularies' that contain related 'Terms'.

The PROMETHEUS logic implemented for taxonomy allows a vocabulary to be set up with either tags defined by user (also known as folksonomy) or terms defined by administrator.

When users view a KEP to which a term has been assigned, along with the KEP, many themes will generally display the node's term(s). Each term appears as a link. Clicking the link displays a page showing the other KEPs with the same term.

The PROMETHEUS Taxonomy organizes taxonomies into vocabularies which consist of one or more terms. Vocabularies group terms that describe an aspect of the content. Each vocabulary consists of a set of terms.

PROMETHEUS can have an unlimited number of vocabularies each containing an unlimited number of terms.

Within a vocabulary terms can be ordered into hierarchies. In PROMETHEUS, all vocabularies are hierarchical, in other words, you can simply arrange items in a hierarchy.

Vocabularies may be designated as “free” tagging in which users creating new content don't have to classify it with terms from a "controlled vocabulary", previously defined. Instead users can freely define terms, or "tags".

Vocabularies can be set to allow terms to define related terms. This function is similar to "see also" in a dictionary.

Vocabularies define whether users may attach only a single term to a node or whether users may attach multiple terms to a node.

Simple implementations might create a set of terms without hierarchies. More complex implementations, in the future, might use hierarchies of terms.

By using multiple vocabularies it is possible to classify an individual content in multiple ways. It’s also possible tie the vocabulary to particular content type: AP, TL, EV, PR, CM (see Figure 1) or whatever. Then when users create content of a particular type, they'll see a list of the vocabulary terms that go with it. Users can then categorize their post by choosing from the list (you can also give your vocabulary a help text to help your users choose).

Vocabularies can have hierarchies of terms. In PROMETHEUS, administrator simply arranges items to create a hierarchy: “single select” allows terms to be nested but each sub-term is associated on only one parent; “multiple select” allows a term to be associated to multiple parents. With either single or multiple a vocabularies can have as many levels as desired. If administrator allow "free tagging", when users create content they can make up their own terms as they go along, instead of having to choose from a list.

By choosing "multiple select", administrator can allow users to put a post into more than one category at once by tagging it with more than one vocabulary term.

Also, if appropriate, administrator can require to users that create content of a certain "content type" they assign at least one of this vocabulary's terms.

Administrator can decide the order in which vocabulary will appear in lists by assigning a "weight" to it.

Finally, administrator can delete the vocabulary altogether, thereby also deleting all its terms (but not the content to which they were assigned).

Administrator must assign your term a name (you have to do it. There's no such thing as a "nameless term").

Administrator can list synonyms for a term (this creates what is known as a "thesaurus"), decide the order in which the term will appear in lists by assigning it a "weight" and also delete a term altogether.

Moreover the menu of PROMETHEUS can call the KEPs that match terms of a specific taxonomy, those terms named categories by administrators.

IV. INCREMENTAL PRODUCTION OF KEPS

A KEP is generally based on conjectures, hypotheses and principles. As they mature, their contents must all become principle-based. The transformation of a statement from conjecture through hypothesis to principle must be based on experimentation showing evidence of its validity. The experimentation, details of its execution and relative results, are collected in the Evidence component (EV), and duly pointed to by the knowledge package.

Finally, a mature knowledge package is used in one or more projects, by one or more firms. At this stage the details describing the project and all the measurements made during its execution that express the efficacy of use of the package are collected in the Projects component (PR) associated with the package. A KEP is undergoing a process of incremental improvement that aims to reach all parts described above. The incremental completion is performed by different authors who cooperate but that are geographically and tempo rally spread.

As shown in Figure 5 Author(s) produce the KEP with their own knowledge. Researchers and practitioners, beneficiaries of the contents of KEP, reported as Recipient(s), acquire the innovation contents contained in KEP, whatever stage they are. The KEP evolve since then, through their research or their experiments becoming their own authors. The results of the research or experiments, properly formalized, enrich the KEP.

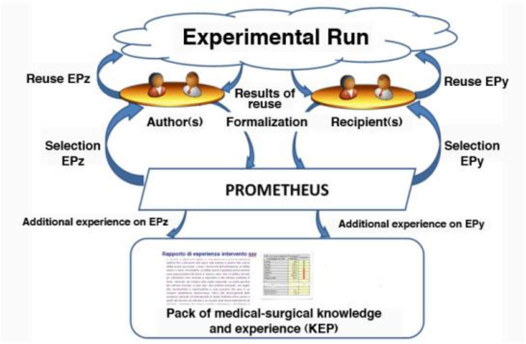


Figure 5. Incremental Production of KEPs

V. CONCLUSIONS

In this paper, we presented a web-based platform aimed for use in companies implementing the concept of experience factory. We provided a general overview of PROMETHEUS knowledge management framework, and described the main features and structure of KEP, which are Contents, Attributes, Educational and Training E-Learning, Taxonomy. We also showed that PROMETHEUS integrates a Knowledge Management System and a Learning System, allowing navigation among all its components.

We have already validated PROMETHEUS in academic environment [21], [24], [26] but it is necessary in order to generalize the validity of the KEP proposed in this work that it is validated by empirical studies in non-academic environments. Obviously, in order to demonstrate the validity of PROMETHEUS many other empirical investigations and studies are needed, in particular industrial context. For this reason, the authors intend plan and execute experiments, make instruments and material available to other interested researchers.

REFERENCES

[1] P. Ardimento, N. Boffoli, M. Cimitile, A. Persico, and A. Tammaro, “Knowledge Packaging supporting Risk Management in Software Processes”, Proceedings of IASTED International Conference on Software Engineering SEA, Dallas, pp. 30-36, November 2006.

[2] P. Ardimento, M. Cimitile, and G. Visaggio, “La Fabbrica dell’Esperienza nell’Open Innovation”, proceedings of A.I.C.A., Benevento (Italy), September 2004.

[3] H. Chen and Z. Wu, "On Case-Based Knowledge Sharing in Semantic Web," ictai, pp.200, 15th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'03), 2003

[4] H.W. Chesbrough, “Open Innovation: The New Imperative for Creating and Profiting from Technology”,Harvard Business School Press, 2003.

[5] C. Edquist, “Systems of Innovations Approaches: Their Emergence and Characteristics”. In Edquist, C. (Ed.) Systems of Innovation: Technologies, Organizations and Institutions. London: Pinter Publishers/Cassell Academic, 1997

[6] R. L. Glass, "A Sad SAC Story about the State of the Practice," IEEE Software, vol. 22, no. 4, pp. 120, 119, July/Aug. 2005, doi:10.1109/MS.2005.82

[7] W. Huang, M. O'Dea, and A. Mille, “ConKMeL: A Contextual Knowledge Management Framework to Support Intelligent Multimedia e-Learning”. Proceedings of IEEE Fifth International Symposium on Multimedia Software Engineering, pp 223-230, 2003, doi: 10.1109/MMSE.2003.1254445

[8] A. Jedlitschka and D. Pfahl, “Experience-Based Model-Driven Improvement Management with Combined Data Sources from Industry and Academia”. Proceedings of the 2003 International Symposium on Empirical Software Engineering, pp. 154-161, 2003.

[9] K.D. Joshi, S. Sarker, and S. Sarker, "The Impact of Knowledge, Source, Situational and Relational Context on Knowledge Transfer During ISD Process," hicss, vol. 8, pp.252c, Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05), doi: 10.1109/HICSS.2005.597.

[10] M. Klein, “Combining and relating ontologies: an analysis of problems and solutions”. In IJCAI-2001 Workshop on Ontologies and Information Sharing, pp. 53--62, Seattle, WA, 2001.

[11] T.W. Malone, K. Crowston, and G.A. Herman, “Organizing Business Knowledge-The MIT Process Handbook”, MIT Press Cambridge, 2003.

[12] C.A. O’Reilly and L.M. Tushman, “The ambidextrous organization” Harvard Business Review 82 (4):74-81, 2004.

[13] D. J. Reifer, “Is the Software Engineering State of the Practice Getting Closer to the State of the Art?”, IEEE Software, Volume 20 Issue 6, November 2003, doi: 10.1109/MS.2003.1241370.

[14] F. Tao, D. Millard, A. Woukeu, H. Davis, “Managing the Semantic Aspects of Learning Using the Knowledge Life Cycle”. Fifth IEEE International Conference on Advanced Learning Technologies (ICALT'05), pp. 575-579, 2005, doi: 10.1109/ICALT.2005.195.

[15] K. Schneider and T. Schwinn, ”Maturing Experience Base Concepts at DaimlerChrysler”, Software Process Improvement and Practice, pp. 85–96, 2001.

[16] G. Xiangyang, H. Linpeng, and L. Dong, “Intelligent Data Transferring Based on Semantic Web Services”, Services Computing, 2004 IEEE International Conference on (SCC'04), pp.463-466, 2004.

[17] R. Agarwal and E. Carmel, ”Tactical Approaches for alleviating Distance in Global Software Development”, IEEE Software, pp. 22-29, Mar-Apr 2001.

[18] G. Visaggio, “Knowledge Experience Base and Experience Factory”, available at http://serlab.di.uniba.it/images/stories/Serlab/knowledgebaseandexperiencefactory.pdf, retrieved on 01 October 2010.

[19] D.M. Amidon, “Blueprint for 21st Century Innovation Management”, Journal of Knowledge Management, Volume 2, Number 1, pp 23-31, September 1998.

[20] W3C XML Schema available at http://www.w3.org/XML/Schema retrieved on 01 October 2010

[21] P.Ardimento, D.Caivano, M.Cimitile, and G.Visaggio, “Empirical Investigation of the Efficacy and Efficiency of tools for transferring software engineering knowledge”, Journal of Information & Knowledge Management, Volume 7, ISSUE 3, September 2008, pp. 197-208.

[22] P. Ardimento, M. Cimitile, and G. Visaggio, “Knowledge Management integrated with e-Learning in Open Innovation”, Journal of e-Learning and Knowledge Society, Vol. 2, n.3, Erickson edition, pp. 343-354, 2006.

[23] PROMETHEUS available at http://serandp.com/prometheus retrieved on 01 October 2010

[24] P. Ardimento, M.T. Baldassarre, M. Cimitile, and G. Visaggio “Empirical Experimentation for Validating the Usability of Knowledge Packages in the Innovation Transfer”, Communications in Computer and Information Science, ISSN: 1865-0929 (Print) 1865-0937 (Online), Volume 22, pp. 357-370, November 2008, Springer Berlin Heidelberg.

[25] P. Ardimento and M.Cimitile, “An Empirical Study on Software Engineering Knowledge/Experience Packages”, 9th International Conference on Product Focused Software Process Improvement (PROFES 2008), Roma 23-25 June 2008, Lecture Notes in Computer Science 5089 Springer 2008, ISBN 978-3-540-69564-6 pp. 298-303.

[26] P. Ardimento, M.T.Baldassarre, M.Cimitile, and G.Visaggio, “Empirical Validation on Knowledge Packaging supporting knowledge transfer“, 2nd International Conference on Software and Data Technologies (ICSOFT) 2007, pp 212-219, ISBN: 978-989-8111-05-0, 2007, Volume PL/DPS/KE/MUSE.

[27] V.R. Basili, G. Caldiera, and H. D. Rombach, “Experience Factory”. In Marciniak, J.J. (ed.), Encyclopedia of SE, vol 1, John Wiley & Sons; pp. 469-476, 1994

[28] A. Koennecker, R. Jeffery, and G. Low, “Lessons Learned from the Failure of an Experience Base Initiative Using Bottom-up Development Paradigm”. 24th Annual Software Engineering Workshop. 1999.

A Performance Analysis of Snort and Suricata Network Intrusion Detection and Prevention Engines

David J. Day
School of Computing and Mathematics
University of Derby
Derby, UK
d.day@derby.ac.uk

Benjamin M. Burns
School of Computing and Mathematics
University of Derby
Derby, UK
benburns01@googlemail.com

Abstract - **Recently, there has been shift to multi-core processors and consequently multithreaded application design. Multithreaded Network Intrusion Detection and Prevention Systems (NIDPS) are now being considered. Suricata is a multithreaded open source NIDPS, being developed via the Open Information Security Forum (OISF). It is increasing in popularity, as it free to use under the General Public Licence (GPL), with open source code. This paper describes an experiment, comprising of a series of innovative tests to establish whether Suricata shows an increase in accuracy and system performance over the de facto standard, single threaded NIDPS Snort. Results indicate that Snort has a lower system overhead than Suricata and this translates to fewer false negatives utilising a single core, stressed environment. However, Suricata is shown to be more accurate in environments where multi-cores are available. Suricata is shown to be scalable through increased performance when running on four cores; however, even when running on four cores its ability to process a 2Mb pcap file is still less than Snort. In this regard, there is no benefit to utilising multi-cores when running a single instance of Snort.**

Keywords – snort; suricata; performance; NIDS; NIDPS; multithreaded; multi-core; comparison; experiment

I. Introduction

Nielsen's Law states that the bandwidth available to users increases by 50% annually [1]. This exponential growth perpetrates design challenges for developers of Network Intrusion Detection and Prevention Systems (NIDPS). Once traffic levels exceed operational boundaries, packets are dropped and the system becomes ineffective. With pattern matching taking up to 70% of the total processing time [2];[3], copious research is focused on reducing the pattern matching overhead with inventive algorithms. Alternatively, some NIDPSs utilise specialist hardware such as Application Specific Integrated Circuits (ASICS) and Field Programmable Gateway Arrays (FPGA) providing parallelism to increase throughput [4]. However, these systems are costly, leaving some organisations restricted to using single threaded PC-based freeware, such as Snort.

With Internet bandwidth accelerating and Central Processing Unit (CPU) core speeds reaching a plateau, it is unlikely that a single threaded solution will remain effective. The relative influence of Moore's Law [5] on single threaded application performance is reducing and this is responsible for a developmental shift toward increasing power-density for multithreaded processing [6]. Consequently, almost all PC CPUs are now multi-core. However, multi-core processors are only as valuable as the multithreading software utilising them and Snort is not multithreaded.

To address this, Suricata has been released by the Open Information Security Foundation (OISF). It is an open source NIDS promising multi-threading and graphics card acceleration in the form of CUDA (Computer Unified Device Architecture) and OpenCL [7]. Other feature benefits include: Gzip Decompression, Automatic Protocol Detection, Flow Variables, Independent HTP library and Fast IP (Internet Protocol) Matching [8]. If Suricata delivers on the promises of the OISF it may meet the demands caused through exponential increases in network traffic.

This paper describes an evaluation of Suricata through critical comparison of both Suricata and Snort NIDPSs. The remainder of this paper is organised as follows: Section II describes the experiment design including empirical metrics, test-bed development, system stressing, traffic generation and attack methods. In Section III, the experiments are described and the results are reported. In Section IV, we present an analysis of the results, and finally, in Section V, we offer our conclusions.

II. Experiment Design

A. Metrics

Antonatos et al. [3] suggest that the metrics to be used for measuring the performance of an NIDPS should be the attack detection rate, false positives and capacity. Limitations in capacity imply false negatives; once a NIDPS exceeds its capacity, packets will be dropped and subsequently any malicious content within them will not be detected. Mell et al. [9] define the quantitative metrics used for evaluating NIDPS accuracy as follows: coverage (amount of attacks possible to detect), probability of false alarm, probability of detection, attack resistance, ability to handle high bandwidth traffic and capacity. With regard to capacity, it has a number of constituent components and thus, it is not a single metric. Table 1, informed by Hall & Wiley [10], illustrates some of the metrics that constitute capacity.

The above research informed that the following capacity metrics should be recorded: Bytes per second, packets per second and quantity of network attacks. In addition, for each NIDPS, the number of packets dropped, true positives, true negatives, false negatives, and the total amount of alarms were also recorded. Finally, the host resources monitored were, CPU and memory utilisation, persistent storage, network interface bandwidth and page file statistics.

TABLE 1 METRICS OF CAPACITY

Test Metrics	Resources Used
Packets per Second	CPU Cycles, network interface bandwidth, memory bus bandwidth.
Bytes per second (average packet size)	CPU Cycles, network interface bandwidth, memory bus bandwidth.
Protocol Mix	CPU cycles and memory bus bandwidth.
Number of unique hosts	Memory size, CPU cycles, memory bus bandwidth.
Number of new connections per second	CPU cycles and memory bus bandwidth.
Number of concurrent connections	Memory size, CPU cycles, memory bus bandwidth.
Alarms per second	Memory size, CPU cycles, memory bus bandwidth.

B. Test-bed

The test-bed was setup in a virtual environment, facilitating experiment portability and security. It also allowed for faster experiment initialisation. This was necessary for frequent repetition and re-configuration of the experiment tests.

Vmware workstation 6.5 was used as the virtualisation platform, largely due to superior IO and disk performance over competitors Virtual Box and Virtual PC [11]. Ubuntu 10.4 TLS 32 bit was chosen as the operating system. Ubuntu is frequently updated and has a good community base. Further it is the most popular Linux operating system [12]

The default NIDPS hardware configuration was a 2.8GHz (E5462) Quad-Core Intel Xeon, running with 1-4 cores and 3GB of DDR2 800MHz fully-buffered memory. Each system also had a maximum hard-drive capacity of 20GB. The network traffic was replayed for each system separately. The system used to replay the network traffic utilises a single core, as well as 1 GB of memory. The VMware host operating system utilised 2GB of memory and 1 core preventing the host from having any performance impacts on the test-bed.

Snort and Suricata were configured to run using identical rule-sets. Suricata uses a different classification configuration to Snort, which uses 134 decoder and 174 pre-processor rules. Both NIDPSs were using identical logging methods, namely, Barnyard, MYSQL and AcidBase. The versions of Snort and Suricata used were v2.8.5.2 and v1.0.2 respectively. Both systems used the Snort v2.8.5.2 VRT rule set, combined with the Emerging Threats rule set. After all rules were loaded, Suricata had 11039 detection rules loaded against Snorts 11065. This discrepancy was due to Suricata's failure to parse certain VRT rules.

C. Traffic

There are a number of considerations when choosing network traffic for NIDPS testing. Firstly, attack traffic can be used, either on its own, or, with the added context of background traffic. When using background traffic, this can either be real or simulated. If it is real, it could be left intact or alternatively, sanitised [9] i.e. payload and ip address information removed.

For the test to be useful, it is deemed desirable to use real network background traffic. However, repetition of the experiments, using real-time network traffic, would be unpredictable due to its dynamic nature. Our solution was to use traffic that had been captured to a pcap (packet capture) file. This facilitated their processing by the NIDPSs in offline mode, allowing for replay on the network at different speeds, using TCPReplay [13]. Further, any risk to mission critical networks was removed.

There are numerous test traffic sources available online for download, unfortunately, these are often sanitised. This renders them useless for evaluating content matching NIDPS, which perform deep packet inspection. Tools do exist which can add random payloads into sanitised data, e.g., TCPdump Randomiser [14], however, the realism of such modified data becomes questionable. Hacking contests also offer sources of traffic capture, although the traffic content is not documented, hence this must be predetermined prior to use, e.g., which attacks were used and which were successful. As a result of these issues, it was decided to capture background traffic from a busy universities web and application server. This was then merged with exploit traffic, created using the Metasploit Framework [15]. The Metasploit Framework contains a total of 587 exploit modules [15], allowing attack data to be easily generated in quantity.

The exploit traffic was captured by performing attacks via Metasploit to a Microsoft Windows 2000 machine. Windows 2000 was chosen as there are more Metasploit exploits for this operating system than any other. Numerous services and discontinued applications where installed to facilitate as many of these attacks as possible. Regrettably, they could not all be obtained. The attacks perpetrated are shown in Table 2, captured using Wireshark [16]. With the background and attack traffic captured, the two were combined. Part of the Wireshark application, Edicap, was used to modify the timestamp of the exploit traffic, to correlate with the background traffic. With this done, the two were merged together in chronological order, such that the attack traffic was distributed within the background traffic.

D. Stressing the system

The capacity of a NIDPS is closely connected to the CPU capacity of the system [2]. Thus, Snort and Suricata should be subjected to CPU impairment, to evaluate their efficacy under stressful conditions.

TABLE 2 EXPLOITS PERFORMED

Code	Name	Description
ms03_026_dcom	Microsoft RPC DCOM Interface Overflow	Module exploits a stack buffer overflow in the RPCSS service
ms05_039_pnp	Microsoft Plug and Play Service Overflow	Stack buffer overflow in the Windows Plug and Play service
ms05_047_pnp	Microsoft Plug and Play Service Registry Overflow	Stack buffer overflow in Windows PnP services. Causes Reboot
ms06_040_netapi	Microsoft Server Service NetpwPathCanonicalize Overflow	Stack buffer overflow in the NetApi32 CanonicalizePathName() function using the NetpwPathCanonicalize RPC call in the Server Service
ms05_017_msmq	Microsoft Message Queueing Service Path Overflow	Exploits a stack buffer overflow in the RPC interface to the Microsoft Message Queueing service
ms01_033_idq	Microsoft IIS 5.0 IDQ Path Overflow	exploits a stack buffer overflow in the IDQ ISAPI handler for Microsoft Index Server

VMware was used to allow the number of logical and physical cores to be reduced. The cores themselves were stressed by generating threads, causing an adjustable and measureable workload. This was performed using the application cpulimit [17], which generates configurable workloads across the processor, allowing for the total amount of stress applied by each thread, to be limited by a percentage of the CPU capacity.

Snort and Suricata both provide the ability to replay pcap files internally. This is done at the maximum speed possible for the NIDPS, providing a good metric as to the performance of a system. Yet, using this method the maximum loss free rate (MLFR) cannot be accounted for. Therefore, TCPReplay [13] was used to control the traffic replay rate, thereby allowing for stress testing under network load.

E. System Monitoring

The following resources were monitored: CPU utilisation, memory utilisation, persistent storage bandwidth and network interface bandwidth. This was performed using the Linux command line utility dstat.

F. Experiment protocol

Throughput speeds are increasing [18], and the MLFR of NIDPSs is affected by both the utilisation of the CPU and the throughput of the traffic [10]; [3]. Thus, the experiment was designed to provide data regarding how each system performs, with increased throughput and under increased CPU stress.

Attack traffic was played to both NIDPSs, with varying CPU configurations. These were: core configuration of 2 processing cores, 1 core, 50% and 75% load. The ability for the NIDPSs to read the packets, along with the accuracy of alerts, was measured, with special attention being paid to the false negative rate. The test traffic was replayed into the environment through TCPReplay, at a multiplier of 40, i.e., replayed 40 times faster than it was captured. This results in a reported playback throughput of 3.1 Mbps, and a packet drop rate of under 2%. This ensured the experiments could be completed in a timely fashion, on the threshold of packet loss.

Each time a test was run, the start and end times of the NIDPS start-up and traffic replay were recorded. This provided a good reference point when analysing the alerts and system statistics. For each test run, the alert output information was recorded using acidbase, as well as the unified2 alert output file being archived for future reference. Statistics produced on NIDPS close down, reported, the number of generated alerts, how many packets were processed, and what ratio of network protocols were handled. Any traffic travelling from hosts 192.168.16.2 and 192.168.16.128 was known to be malicious traffic.

III. RESULTS

This section reports and analyses the results for each NIDPS, for accuracy, dropped packet rate, system utilisation and offline speed. Each of these is now discussed in turn.

A. Accuracy

To determine accuracy, control alerts were used. These are alerts generated without system stress, used as a baseline. Deviation from the baseline under stress is an indication of a change in detection accuracy.

Table 3 shows the number of alert types generated when the attacks were performed against each NIDPS. Figure 1 shows Suricata alerted on every exploit, under all configurations, yet some alerts types were lost, resulting in a reduction of detection breadth [19].

TABLE 3 ALERTS GENERATED BY SNORT AND SURICATA

Alert	Snort	Suricata
ms05_040_pnp	4	4
ms05_047_pnp	1	1
ms05_039_pnp	1	6
ms03_026_dcom	1	2
ms01_033_1dq	2	4
ms05_017_msmq	2	3

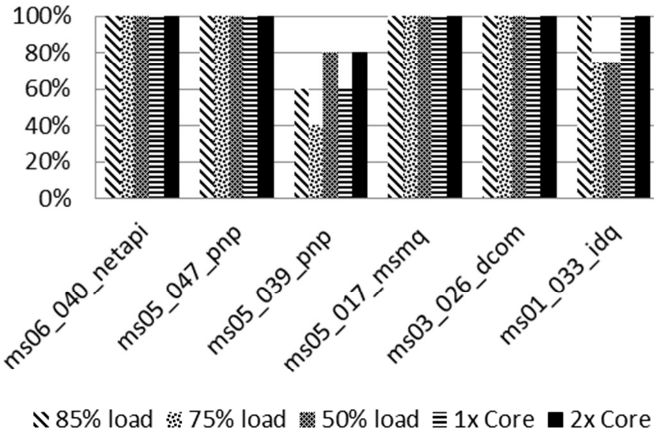


Figure 1 Suricata Alerts

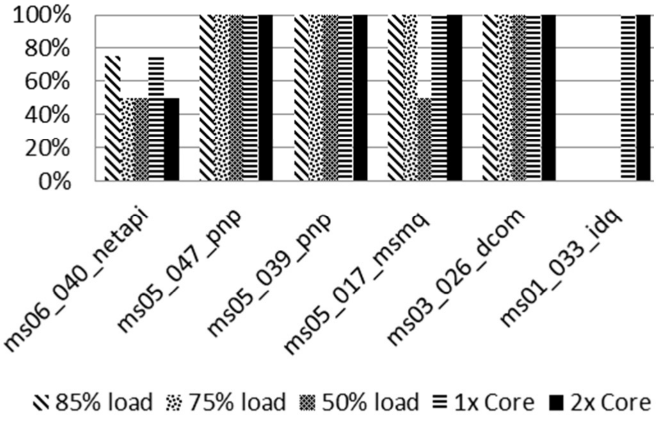


Figure 2 Snort Alerts

Figure 2 shows Snort fails to alert on ms01_033_idq. This is a false negative caused by excessive load.

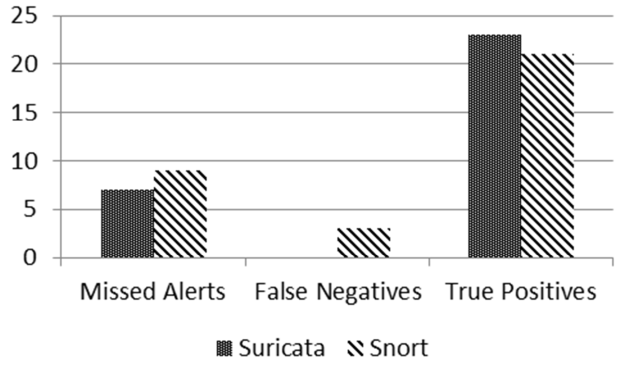


Figure 3 Attack accuracy measurements

Figure 3 shows the number of false positives (fp) and true positives (tp) for both NIDPSs, relative to the number of missed alerts by each system.

B. Dropped Rate

False negatives (fn) can be caused by dropped packets. Figure 4 plots the amount of packets dropped by Snort and Suricata as the CPU availability drops. While Snorts percentage drop is largely linear, Suricata's performance diminishes significantly once the CPU availability reduces below one core. Figure 5 shows how reducing the number of cores, and stressing the CPU, effects false negatives on both systems.

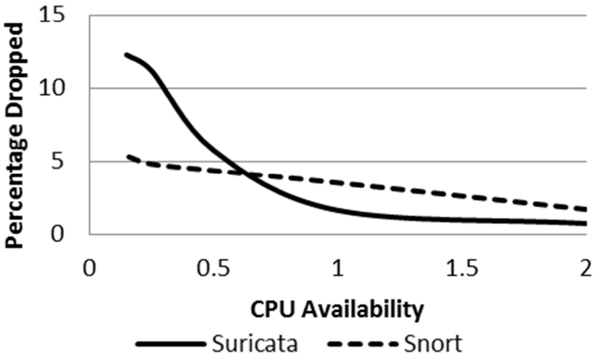


Figure 4 Packet loss at 3.2 MBps

C. System utilisation

Figure 6 shows the relationship between CPU utilisation, and network throughput, on both Suricata and Snort. It depicts how CPU load increases relative to network throughput. This behaviour is more prominent whilst running Suricata, with Snort exhibited similar behaviour on much smaller scale.

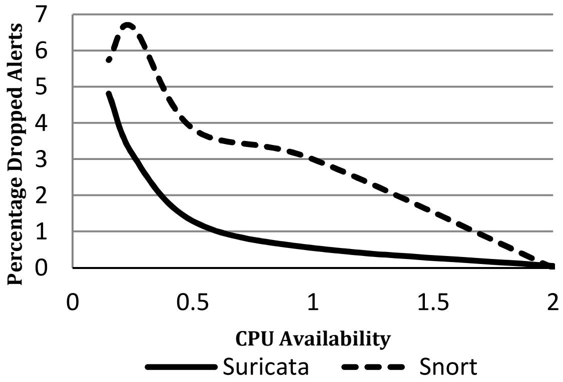


Figure 5 False Negatives (dropped alerts)

With dual-cores available, Suricata has a lower drop rate than Snort. To investigate why, both systems were evaluated for their ability to utilise both cores. Figures 7 and 8 show how Snort and Suricata (respectively), utilise dual core processors.

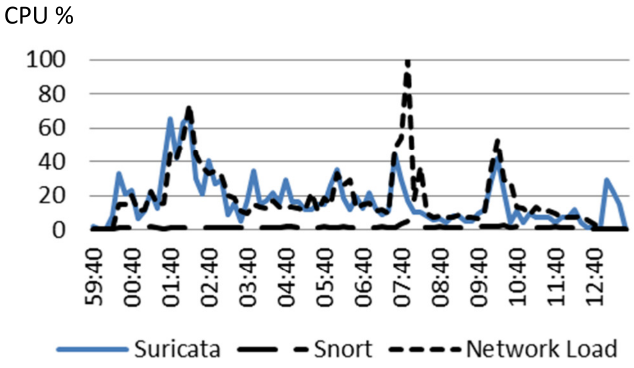


Figure 6 Network throughput and CPU utilisation for the Single Core Configuration

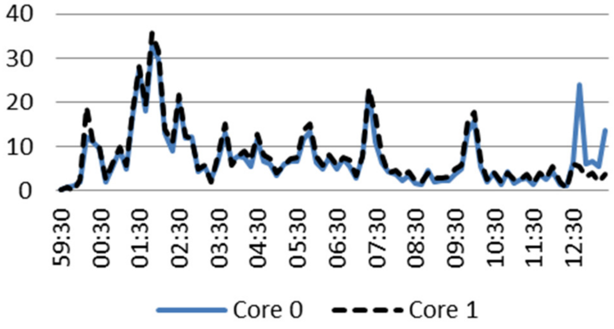


Figure 7 Suricata utilising dual cores

Figure 7 shows that Suricata utilises the 2 cores uniformly, compared to Snorts more erratic load balancing, Figure 8. This is consistent with expectations due to Suricata's multithreaded design.

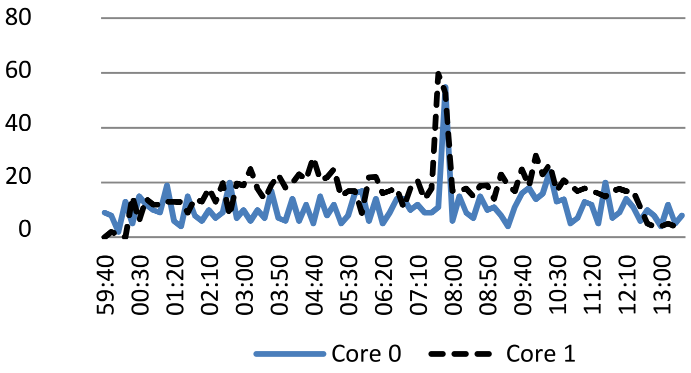


Figure 8 Snort utilising dual cores

Both NIDPSs have the ability to process traffic in offline mode by receiving a pcap file and processing it at maximum capacity. This was performed to identify the speed in which both systems can process traffic. The test was performed for both NIDPSs, using the same pcap file. The time each system took to process the file is displayed in Figure 12.

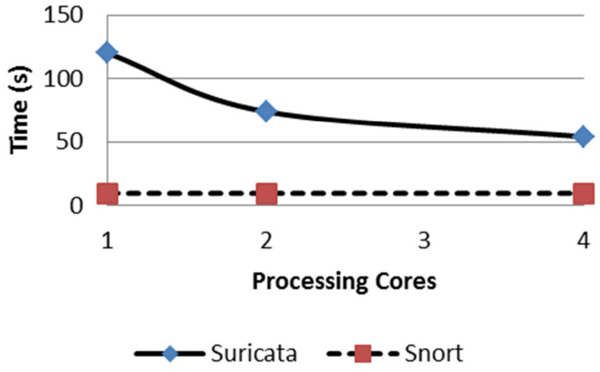


Figure 9 Pcap processing time (offline mode)

Additional cores did not improve Snorts processing time, although Suicata's performance increased by 220%, when using four cores, compared to one. Again this is expected, considering Suricata's multithreaded design.

IV Result analysis

Arguably the most important metric to evaluate NIDPSs, is accuracy. This has been described as the system's attack coverage, false positives, false negatives, capacity, and ability to handle high bandwidth traffic [9]. The experiments outlined provided details regarding all of these.

The developers of Suricata have stated that their primary focus is improving NIDPS accuracy [20]. With Suricata having a higher accuracy than Snort, our experiments show that they have had some success. This is evident in Figures 1,2 and 3, including data showing that Snort failed to alert on the ms01_033_idq exploit with the processor loaded at 50% or above. A partial reason is Snort having less control alerts triggered by the attack than Suricata (two compared with four). Snort failed to alert on ms01_033_idq using two rules from the VRT rules set, i.e., ID 1245 and 1244. Suricata succeeded in that these same alerts were triggered.

Larger processing requirements demanded by Suricata caused it to reach its operational capacity quicker than Snort, explaining the greater number of dropped packets under stress. By comparison, Snort places less demand on the system, enabling it to have a reduced packet drop rate at peak system loads. Figure 4 shows the percentage of dropped packets increasing steeply, once CPU availability is reduced to a single stressed core. The proportional relationship between dropped packets, and false negatives, is demonstrated for both systems in Figure 5.

When Suricata is run on a multi-core configuration it has a lower packet loss rate than Snort. Figures 7 and 8 show that Suricata uses available cores, on a dual core system, in a more uniform fashion. Offline tests show that Suricata was considerably slower than Snort. Alhough multiple cores relates to a more marked improvement with Suricata, than Snort, see Figures 4,5 and 9. In this sense, it could be argued that, Suricata possesses an improved ability to provide scalability. Nevertheless, in circumstances when the bandwidth received is greater than Snort can handle, the recommendation is to run multiple instances of Snort on multiple processor cores [21]. This could provide scalability similar to that of Suricata, albeit with added cost of processing a single threaded application over multiple cores.

V. *Conclusions*

The analysis of the results has shown that Suricata has a higher accuracy rate than Snort, although this comes at the cost of putting an increased relative demand on the CPU. The results show that, due to utilising multiple cores more uniformly, Suricata has the potential to be a more scalable and efficient, where multiple cores are available. However, due to the higher resource demands of Suricata, the accuracy would be expected to diminish, when used in low commodity, single core, deployment.

This research has endeavoured to classify the performance benefits of the innovative Suricata engine. Whilst the potential of Suricata is significant, at present, its development is incomplete. Since acceptable performance is not guaranteed, trial implementations of the engine would be advised. This would provide an opportunity for feedback; accelerating the developmental process of this pivotal detection and prevention engine. In addition to this, future research should pay attention to documenting Suricata's performance, whilst utilising even larger numbers of cores.

REFERENCES

[1] J. Nielsen, "Nielsen's Law of Internet Bandwidth," *useit.com: Jakon Nielsen's Website,* [Online] 5 April 1998, [Cited: 4 January 2011.] http://www.useit.com/alertbox/980405.html.

[2] J.B.D. Cabrera, J.Gosar, and R.K. Mehra, "On the statistical distribution of processing times in network intrusion detection," 43rd IEEE Conference on Decision and Control, vol. 1, IEEE Press, 2004, pp. 75-80, doi: 10.1109/CDC.2004.1428609.

[3] S.Antonatos, K.Anagnostakis, and E. Markatos, "Generating realistic workloads for network intrusion detection systems," Proceedings of the 4th ACM workshop on software and performance, ACM, 2004, pp. 207-215, doi: 10.1145/974043.974078

[4] M. Abishek, W. Najjar, and L.Bhuyan, "Compiling PCRE to FPGA for accelerating SNORT IDS," ACM, 2007, Proceedings of the 3rd ACM/IEEE Symposium on Architecture for networking and communications systems , pp. 127-136. doi: 10.1145/1323548.1323571.

[5] G. Moore, "Cramming more components onto integrated circuits," Electronics, McGraw Hill, Vol. 38, Num. 8, 19 April 1965

[6] A. Ghuloum, "Face the inevitable, embrace parallelism," Communications, vol. 52, ACM, September 2009, pp. 36-38. doi: 10.1145/1562164.1562179.

[7] M. Jonkman, "Suricata IDS Available for Download," *Seclists.org*, [Online] 2009, [Cited: 12 May 2010.] http://seclists.org/snort/2009/q4/599.

[8] OISF, "The open information security foundation," [Online] 2010, [Cited: 4 October 2010.] http://www.openinfosecfoundation.org/index.php?start=15.

[9] P. Mell, V. Hu, R. Lippmann, J. Haines, and M. Zissman, "An Overview of Issues in Testing Intrusion Detection Systems," [Online], [Cited: 16 September 2010.] http://csrc.nist.gov/publications/nistir/nistir-7007.pdf.

[10] M. Hall, and K. Wiley, "Capacity Verification for High Speed Network Intrusion Detection Systems," Lecture notes in computer science, Springer, 2002, vol. 2516, pp.239-251. doi: 10.1007/3-540-36084-0_13.

[11] P. Domingues, F. Araujo, and L. Silva, "Evaluating the Performance and Intrusiveness of Virtual Machines for Desktop Grid Computing," Proceedings of the 2009 IEEE International Symposium on Parallel&Distributed Processing, IEEE, 2009, pp. 1-8. doi: 10.1109/IPDPS.2009.5161134.

[12] L. Bodnar, "Page hit ranking," *DistroWatch.com,* [Online] 2010, [Cited: 20 April 2010.] http://distrowatch.com/.

[13] A. Turner, "TCPReplay pcap editing & replay tools for *NIX," *TCPRepla,* [Online] 23 August 2010, [Cited: 13 December 2010.] http://tcpreplay.synfin.net/.

[14] Institure of Computer Science, "Network monitoring for security: intrusion detection systems*" Institure of Computer Science*, [Online] 6 August 2007, [Cited: 12 December 2010.], http://dcs.ics.forth.gr/dcs/Activities/Projects/ids.html.

[15] Rapid 7, "Metasploit – penetration testing resources," *Metasploit,* [Online] 2010, [Cited: 1 October 2010.] http://www.metasploit.com/.

[16] Wireshark.org.uk, "Wireshark," *Wireshark.org.uk,* [Online] [Cited: 14 April 2010.] http://www.wireshark.org/.

[17] A. Marletta, "CPU usage limiter for Linux," *Sourceforge.net,* [Online] 29 November 2010, [Cited: 13 December 2010.] http://cpulimit.sourceforge.net/.

[18] M. Cloppert, "Detection, Bandwidth, and Moore's Law," *SANS Computer Forensic Investigations and Incident Response Blog,* [Online] 05 Jan 2010, [Cited: 05 May 2010.] https://blogs.sans.org/computer-forensics/2010/01/05/.

[19] C. Jordan, "Writing detection signatures," USENIX, December 2005, ;login, vol. 30, pp. 55-61.

[20] V. Julien, "On Suricata Performance," Inliniac [Online] 2010, [Cited: 06 October 2010.] http://www.inliniac.net/blog/2010/07/22/on-suricata-performance.html.

[21] N. Houghton, "Single Threaded Data Processing Pipelines and Intel Achitectures," *VRT*, [Online] Vulnerability Research Team, 7 June 2010, [Cited: 2010 12 17.], http://vrt-sourcefire.blogspot.com/2010/06/single-threaded-data-processing.html.

Fast Singular Value Decomposition for Large-scale Growing Data

Jengnan Tzeng
Department of Mathematical Sciences, National Chengchi University
Taipei, Taiwan
jengnan@math.nccu.edu.tw

***Abstract*—Singular value decomposition (SVD) is a fundamental technique in linear algebra, and it is widely applied in many modern information technologies, for example, high dimensional data visualization, dimension reduction, data mining, latent semantic analysis, etc. However, when the matrix size of the data is huge and continuously growing, the matrix can not be loaded all at once into the computer memory and $O(n^3)$ computational cost of SVD becomes infeasible. To resolve this problem, we will adapt a fast multidimensional scaling method to obtain a fast SVD method, given that the significant rank of a huge matrix is small. This proposed fast SVD method can be easily implemented via parallel computing. We also propose a fast update method to be applied when the huge data is updated continuously. We will demonstrate that the approximated SVD result is sufficiently accurate, and most importantly it can be derived very efficiently. Using this fast update method, many modern techniques based on SVD which were infeasible will become viable.**

Keywords-Singular value decomposition; multidimensional scaling; parallel computing; huge matrix.

I. Introduction

Singular value decomposition (SVD) and Principle component analysis (PCA) are two fundamental techniques in linear algebra and statistics. There are many modern applications based on these two tools, such as linear discriminate analysis [1], multidimensional scaling analysis [2], feature extraction, high dimensional data visualization, etc. In recent years, digital information has been proliferating and many analytic methods based on PCA and SVD are facing the challenge of their significant computational cost. Thus, it is crucial to develop a fast method of PCA and SVD.

In 2008, Tzeng et al. [4] developed a fast multidimensional scaling (MDS) method which turned the classical $O(n^3)$ MDS method to be linear. MDS is a method to represent the high dimensional data into the low dimensional configuration. Because of the phenomenon of the curse of dimensionality, MDS is widely used in data mining, clustering, and many recommendation systems for web services. When the data configuration is Euclidean, MDS is similar to principle component analysis (PCA), in that both can remove inherent noise with its compact representation of data. The $O(n^3)$ computational complexity makes it infeasible to apply to huge data, for example, when the sample size is more than one million.

In the following section, we will show how to adapt classical MDS to be a fast Split-and-combine MDS (SCMDS). And using this SCMDS, we can modify the PCA and SVD method to become fast methods.

II. Methodology

In 2008, we adapted the classical MDS so as to reduce the original $O(n^3)$ complexity to $O(n)$ [4], in which we have proved that when the data dimension is significantly smaller than the number of data entries, there is a fast linear method for classical MDS. The following section begins with a review of SCMDS. Then we will demonstrate how to adapt SCMDS method to become the fast PCA, and with further modification, the fast PCA can become the fast SVD.

A. From MDS to SCMDS

The main idea of fast MDS is using statistical resampling to split data into overlapping subsets. We perform the classical MDS on each subset and get the compact Euclidean configuration. Then we use the overlapping information to combine each configuration of subsets to recover the configuration of the whole data. Hence, we named this fast MDS method by Split-and-combine MDS (SCMDS).

Assume X_1 and X_2 are matrices in which the columns are the two coordinates of the overlapping points obtained by applying MDS to two grouped data sets. Then there exists an affine mapping that maps X_1 to X_2. Let $\bar{X}_1$ and $\bar{X}_2$ be the means of columns of X_1 and X_2, respectively. In order to obtain the affine mapping, we apply QR factorization to both $X_1 - \bar{X}_1 \mathbb{1}^T$ and $X_2 - \bar{X}_2 \mathbb{1}^T$. Then we have $X_1 - \bar{X}_1 \mathbb{1}^T = Q_1 R_1$ and $X_2 - \bar{X}_2 \mathbb{1}^T = Q_2 R_2$. It is clear that the mean of the center of column vectors of $X_1 - \bar{X}_1 \mathbb{1}^T$ has been shifted to zero. Because X_1 and X_2 come from the same data set, the difference between $X_1 - \bar{X}_1 \mathbb{1}^T$ and $X_2 - \bar{X}_2 \mathbb{1}^T$ is a rotation. Therefor, the triangular matrices R_1 and R_2 should be identical when there is no noise in X_1 and X_2. Due to randomness of the sign of columns of Q_i in QR factorization, the sign of columns of Q_i might need to be adjusted according to the corresponding diagonal elements of R_i, so that the signs of tridiagonal elements of R_1 and R_2 are the same.

After necessary modification to the sign of columns of Q_i, we conclude that

$$Q_1^T (X_1 - \bar{X}_1 \mathbb{1}^T) = Q_2^T (X_2 - \bar{X}_2 \mathbb{1}^T).$$

Furthermore, we have

$$X_1 = Q_1 Q_2^T X_2 - Q_1 Q_2^T (\bar{X}_2 \mathbb{1}^T) + \bar{X}_1 \mathbb{1}^T.$$

Here, the unitary operator is $U = Q_1 Q_2^T$ and the shifting is $b = -Q_1 Q_2^T \bar{X}_2 + \bar{X}_1$. Since the key processing of finding this affine mapping is QR decomposition, the computational cost is $O(k^3)$, where k is the number of columns of X_1 and X_2. Therefore, the cost $O(k^3)$ complexity is limited by the number of samples in each overlapping region. The proof of the computational cost of SCMDS is given as follows:

Assume that there are N points in a data set, we divide these N samples into K overlapping subgroups, where N_G is the number of points in each subgroup and N_I the number of points in each intersection region. Then we have the relationship

$$KN_G - (K-1)N_I = N$$

or

$$K = \frac{(N - N_I)}{(N_G - N_I)}.$$

For each subgroup, we apply classical MDS to compute the configuration of each group data, which cost $O(N_G^3)$. In each overlapping region, we apply QR factorization to compute the affine transformation, which cost $O(N_I^3)$. Assume that the true data dimension is p, and the lower bound of N_I is $p+1$. For the convenience, we take $N_G = \alpha p$ for some constant $\alpha > 2$. Then the total computational cost is about

$$\frac{N-p}{(\alpha-1)p} O(\alpha^3 p^3) + \frac{N - \alpha p}{(\alpha-1)p} O(p^3) \approx O(p^2 N).$$

The first term of above equation is the complexity of MDS in K groups, and the second term is the complexity of QR in the $K-1$ overlapping regions.

When $p << N$, the computational cost $O(p^2 N)$ is much smaller than $O(\sqrt{N}N)$, which is the computation time of the fast MDS method proposed by Morrison et al., 2003 [3]. The key idea of our fast MDS method is to split data into subgroups, then combine the configurations to recover the whole one. Since all the order three complexities are restricted in the small number of data entries, we can therefor speed up MDS. The concept of split-and-combine is also similar to the concept of parallel computing. Thus, SCMDS method can be easily implemented via a parallel algorithm.

B. From SCMDS to SCPCA

Because MDS is similar to principle component analysis (PCA) when the data configuration is Euclidean, we can adapt SCMDS method to obtain the fast PCA in the same constrain $p << N$.

Assume that X is a p-by-N matrix, where there are N samples with p dimension. $D = X^T X$ indicates product matrix of X, and $\mathbb{1}$ is an N-by-1 vector whos elements are all 1's. We define a symmetric matrix B by

$$\begin{aligned} B &= \left(X - \frac{1}{N} X \mathbb{1}\mathbb{1}^T\right)^T \left(X - \frac{1}{N} X \mathbb{1}\mathbb{1}^T\right) \\ &= D - \frac{1}{N} D \mathbb{1}\mathbb{1}^T - \frac{1}{N} \mathbb{1}\mathbb{1}^T D + \frac{1}{N^2} \mathbb{1}\mathbb{1}^T D \mathbb{1}\mathbb{1}^T \\ &= D - \bar{D}_r - \bar{D}_c + \bar{D}_g, \end{aligned}$$

where $\bar{D}_r = \frac{1}{N} D \mathbb{1}\mathbb{1}^T$ is the row mean matrix of D, $\bar{D}_c = \frac{1}{N} \mathbb{1}\mathbb{1}^T D$ is the column mean matrix of D and $\bar{D}_g = \frac{1}{N^2} \mathbb{1}\mathbb{1}^T D \mathbb{1}\mathbb{1}^T$ is the ground mean matrix of D. The operator from D to $D - \bar{D}_r - \bar{D}_c + \bar{D}_g$ is called double centering. If we define a matrix H by

$$H = I - \frac{1}{N} \mathbb{1}\mathbb{1}^T,$$

B can be simplified to $B = HDH$. Since matrix B is symmetric, the SVD decomposes B into $B = Z \Sigma Z^T$. Then we have

$$\sqrt{B} = Z \Sigma^{\frac{1}{2}} P^T = (X - \frac{1}{N} X \mathbb{1}\mathbb{1}^T)^T,$$

for some unitary matrix P. In practice, we set $P = I$ to obtain the MDS result $\sqrt{B}$. Therefore, the row vector of $\sqrt{B}$ is the coordinates of X with the mean of data been shifted to the original point and rotated by some unitary matrix P.

If D is a distance matrix with each element $d_{i,j} = \sqrt{(x_i - x_j)^T (x_i - x_j)}$, the double center of D^2 is equivalent to $-2B$, provided that $\sum_{i=1}^N x_i = 0$. Hence, the MDS method performs double centering on D^2, multiplies by $-\frac{1}{2}$, and then performs SVD, which gives the configurations of the data set.

The constrain $\sum_{i=1}^N x_i = 0$ in MDS is the same as the constrain in computing PCA. The score matrix P of PCA is an unitary matrix, and it is derived by

$$(X - \frac{1}{N} X \mathbb{1}\mathbb{1}^T)(X - \frac{1}{N} X \mathbb{1}\mathbb{1}^T)^T = P \Sigma P^T.$$

If we have the unitary matrix P, we can use $(X - \frac{1}{N} X \mathbb{1}\mathbb{1}^T)^T P$ to obtain $\sqrt{B}$. The result of MDS ($\sqrt{B}$) simply uses the first r orthogonal columns of this unitary matrix to represent the data. We define an orthogonal matrix Z_r by the first r columns of Z and the sub-diagonal matrix Σ_r by the up-left block of Σ, then $\sqrt{B_r} = Z_r \Sigma_r^{\frac{1}{2}}$ is the r-dimensional configuration of data set. If we have r-dimensional MDS configuration, we can obtain the first r columns of the score matrix of PCA, denoted by P_r. That is

$$P_r = (X - \frac{1}{N} X \mathbb{1}\mathbb{1}^T) Z (\Sigma_r^{\frac{1}{2}})^{-1}.$$

Thus, the key scheme of PCA is embedded in the MDS method. When the number of samples is large and the data set is high dimensional, the complexity is costly. However, if there are many samples which are linearly dependent, the actual rank of the data matrix is much smaller than the matrix size. In this case, SCMDS has advantage in computing speed. And the approach of obtaining P_r by SCMDS is called SCPCA.

C. From SCPCA to SCSVD

The concept of SVD and PCA are very similar. Since the PCA starts from decomposing the covariance matrix of data set, it can be considered as adjusting the center of mass of a row vector to zero. On the other hand, SVD operates directly on the product matrix without shifting. If the mean of the matrix rows is zero, the eigenvectors derived by SVD are equal to the eigenvectors derived by the PCA. We are looking for a method which will give a fast method to produce the SVD result without recomputing the eigenvectors of the whole data set, when the PCA result is given. The following is the mathematical analysis for this process.

Let X be a column matrix of data set. $\tilde{X} = X - \bar{X} \cdot \mathrm{i}^T$, where $\bar{X}$ is the mean of columns of X. Hence, the row mean of $\tilde{X}$ is zero. Assume that we have the PCA result of X, that is, $\tilde{X}\tilde{X}^T = P\Sigma^2 P^T$. Then we have $\tilde{X} = P\Sigma U^T$ for some orthogonal matrix U. Assume that the rank of $\tilde{X}$ is r and r is much smaller than the matrix size, we observe that $rank(X) = r$ or $rank(X) = r + 1$, depending on whether $\bar{X}$ is spanned by P. If $\bar{X}$ is spanned by P, then

$$X = \tilde{X} + \bar{X} \cdot \mathrm{i}^T = P\Sigma U^T + P \cdot c \cdot \mathrm{i}^T = P(\Sigma U^T + c \cdot \mathrm{i}^T),$$

where c is the coefficient vector of $\bar{X}$ when represented by P, *i.e.*, $\bar{X} = P \cdot c$.

If the singular value decomposition of $\Sigma U^T + c \cdot \mathrm{i}^T$ is $W\hat{\Sigma}V^T$, we have

$$X = P(W\hat{\Sigma}V^T) = (PW)\hat{\Sigma}V^T = Z\hat{\Sigma}V^T.$$

Because the matrix W is unitary, $Z = PW$ is automatically an orthogonal matrix as well. Then we have the SVD of X.

Checking the matrix size of $\Sigma U^T + c \cdot \mathrm{i}^T$, we can see that to compute the SVD of $\Sigma U^T + c \cdot \mathrm{i}^T$ is not a big task. This is because $\Sigma U^T + c \cdot \mathrm{i}^T$ is a r-by-n matrix, and under our assumption, r is much smaller than n, so we can apply the economic SVD to obtain the decomposition of $\Sigma U^T + c \cdot \mathrm{i}^T$.

On the other hand, if $\bar{X}$ is not spanned by P, the analysis becomes

$$X = \tilde{X} + \bar{X} \cdot \mathrm{i}^T = [P|p_{r+1}] \left(\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U^T \\ 0 \end{bmatrix} + c \cdot \mathrm{i}^T \right),$$

where p_{r+1} is a unit vector defined by

$$p_{r+1} = \frac{(I - PP^T)\bar{X}}{\|(I - PP^T)\bar{X}\|}.$$

Using the same concept of diagonalization in the case when $\bar{X}$ is spanned by P, we find the SVD of

$$\left(\begin{bmatrix} \Sigma & 0 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} U^T \\ 0 \end{bmatrix} + c \cdot \mathrm{i}^T \right) = W\hat{\Sigma}V^T.$$

Then $X = [P|p_{r+1}]W\hat{\Sigma}V^T = Z\hat{\Sigma}V^T$, where $Z = [P|p_{r+1}]W$ is another orthogonal matrix and hence the SVD of X is completed.

From the above analysis, we can have a fast PCA method by computing SCMDS first, then adapt the MDS result to obtain PCA. We named this approach SCPCA. Similarly, the fast SVD method which computes SCMDS first, then adapts MDS result to obtain PCA, and finally adapts PCA result to SVD, is called the SCSVD. These two new methods work when the rank of X is much smaller than the number of samples and the number of variables. To obtain the exact solution, the parameter N_I must be greater than the rank of X. In SCPCA or SCMDS method, if $N_i \leq r$, we only get the approximated solution of PCA and SVD. Under the necessary criterion, we can reduce the computational complexity from $\min\{O(p^2 n), O(pn^2)\}$ to $\min\{O(rp), O(rn)\}$. If the significant rank is not small, for example,$r \approx \sqrt{pn}$, the computational complexity becomes almost the same as the original PCA and SVD. Our method has no advantage in the latter case.

III. SVD for continuously growing data

In this section, we look for the solution when the data is updated constantly and we need to compute SVD continuously. Instead of scanning all the data again, we try to use the previous SVD result together with the new updated data to compute the next SVD.

Let A be an m-by-n matrix, where m is the number of variables and n is the number of samples. And we assume that both m and n are huge. When new data comes in, we collect these new data to form a column matrix which is denoted by B. Assume that we have the singular value decomposition of A, that is

$$A = Z\Sigma V^T,$$

where $Z \in M_m(\Re)$, $V \in M_n(\Re)$ are orthogonal and Σ is a diagonal. Since the data gets updated, the data matrix becomes

$$A_1 = [A|B].$$

To compute the singular value decomposition of A_1, we need to compute the eigenvalue and eigenvector of $A_1 A_1^T$.

We can represent the column matrix B by $B = ZC$, where C is the coefficient matrix of B with columns of Z as the basis. Since Z is orthogonal, the coefficient matrix C can be computed easily by $C = Z^T B$. Then we have

$$\begin{aligned} A_1 A_1^T &= [A|ZC]\,[A|ZC]^T \\ &= AA^T + ZC(ZC)^T \\ &= Z(\Sigma^2 + CC^T)Z^T \\ &= ZU\hat{\Sigma}^2 U^T Z^T \\ &= Z_1 \hat{\Sigma}^2 Z_1^T. \end{aligned} \tag{1}$$

Note that the matrix $\Sigma^2 + CC^T$ is positive symmetric. Using the spectrum theorem, we can decompose this matrix into $U\hat{\Sigma}^2U^T$. Because the matrix U is unitary, $Z1$ is Z rotated by U.

When the matrix size of A is huge, the computational cost of SVD is high. If the data is constantly growing, it is difficult to compute the singular value decomposition of A_1 in real time. Therefore, we look for an approximated solution with fast method.

Let $Z = [z_1, z_2, \cdots, z_m]$. If the new updated data B has only the components in $\{z_1, z_2, \cdots, z_r\}$, where $r << m$, then only r-dimensional space will be perturbed by this new data. This is proved as follows.

Theorem Let $A = Z\Sigma V^T$. Assume that $A_1 = [A|B]$, where B has no component in i-th column of Z for $i > r$. Then the singular value decomposition of A_1 has the same spectrum σ_i and singular vector z_i, v_i for $i > r$.

Proof: Let

$$A = \begin{array}{c} \begin{array}{cc} r & p-r \end{array} \\ \begin{pmatrix} Z_1 & Z_2 \end{pmatrix} \end{array} \quad \begin{array}{c} \begin{array}{cc} r & p-r \end{array} \\ \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \end{array} \quad \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix},$$

Where Z_1 and V_1 are the first r columns of Z and V. Because B has no component in Z_2, $B = Z_1C$ for some C. Then $A_1A_1^T$ can be written as

$$\begin{aligned} A_1A_1^T &= [A|B][A|B]^T \\ &= AA^T + UU^T \\ &= \begin{pmatrix} Z_1 & Z_2 \end{pmatrix} \begin{pmatrix} \Sigma_1^2 + CC^T & 0 \\ 0 & \Sigma_2^2 \end{pmatrix} \begin{pmatrix} Z_1^T \\ Z_2^T \end{pmatrix} \\ &= \begin{pmatrix} Z_1 & Z_2 \end{pmatrix} \begin{pmatrix} U\hat{\Sigma}_1U^T & 0 \\ 0 & \Sigma_2^2 \end{pmatrix} \begin{pmatrix} Z_1^T \\ Z_2^T \end{pmatrix} \\ &= \begin{pmatrix} Z_1U & Z_2 \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_1^2 & 0 \\ 0 & \Sigma_2^2 \end{pmatrix} \begin{pmatrix} U^TZ_1^T \\ Z_2^T \end{pmatrix} \\ &= \begin{pmatrix} \hat{Z}_1 & Z_2 \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_1^2 & 0 \\ 0 & \Sigma_2^2 \end{pmatrix} \begin{pmatrix} \hat{Z}_1^T \\ Z_2^T \end{pmatrix} \end{aligned} \qquad (2)$$

where U is unitary. We can see that the terms Σ_2 and Z_2 do not change if $Z_2^TB = 0$. Thus, the singular value decomposition of A_1 has the same spectrum σ_i and z_i for $i > r$. Moreover, the v_i for $i > r$ are unchanged too. ■

In many applications, we are only concerned with the first few spectrums and eigenvectors. For example, in high dimensional data visualization, we only consider the first two or three eigenvectors, that is $r = 2$ or 3. In this case, the new updated data should be have component in the space that spanned byZ_2. We are therefore interested in the performance of the approximated solution compared with the true solution when we only retained the first r components of B every time we updated the new data by the previous method. If the performance decays slowly, we can daringly compute only three or four components in many SVD-based methods and the result in low dimensional space is still quite reliable. So, we need to understand in what kind of conditions, the approximated solution is stable.

Now, we analyze the effect of the perturbation of a matrix in its eigenvalues and eigenvectors. Let matrix A be real symmetric and $A = S\Lambda S^T$, where S is unitary, such that $S^{-1} = S^T$. A matrix change ΔA produces changes in eigenvalues and eigenvectors, which are denoted by $\Delta\Lambda$ and ΔS respectively. Because S is orthogonal, $AS = S\Lambda$. Similarly, we have

$$(A + \Delta A)(S + \Delta S) = (S + \Delta S)(\Lambda + \Delta\Lambda).$$

The above equation can be represented by

$$A(\Delta S) + (\Delta A)S = S(\Delta\Lambda) + (\Delta S)\Lambda, \qquad (3)$$

when ignoring the small terms $(\Delta A)(\Delta S)$ and $(\Delta S)(\Delta\Lambda)$.

We multiply equation (3) by S^T, then we have

$$\begin{aligned} \Delta\Lambda &= S^T(\Delta A)S + S^TA(\Delta S) - S^T(\Delta S)\Lambda \\ &= S^T(\Delta A)S + \Lambda S^T(\Delta S) - S^T(\Delta S)\Lambda. \end{aligned} \qquad (4)$$

Because the diagonal terms of $\Lambda S^T(\Delta S)$ and $S^T(\Delta S)\Lambda$ are the same, the diagonal part of $S^T(\Delta A)S$ is what we are looking for. Applying this concept to matrix $\Sigma^2 + CC^T$, CC^T can be considered as $S^T(\Delta A)S$. We can conclude that if the maximal element of the absolute value of CC^T is smaller than the difference between $\sigma_i - \sigma_{i+1}$ for $i = 1, \cdots, r$, then the order of columns of S will not change. The first r columns of $S + \Delta S$ can be approximated stably by the first r columns of S. If CC^T is too large such that the new spectrum $\hat{\sigma_{r+1}} > \hat{\sigma_r}$, the approximation solution that only use first r components to update the new spectrum and singular vectors will fault by using $\hat{z}_{r+1}$ to replace $\hat{z}_r$. This conclusion will be demonstrated in the experimental result.

IV. Experimental result

In this section, we show that our fast PCA and SVD method works well for big sized matrix with small rank. The simulated matrix is created by the product of two slender matrices. The size of the first matrix is p-by-r, and the second matrix is r-by-n. Then the product of these two matrixes is of size p-by-n and its rank is smaller than r. When p and n are large and r is much smaller than p and n, the simulated matrix satisfies our SCSVD condition. We pick $p = 4000$, $n = 4000$ and $r = 50$ as our fist example. The elements of the simulated matrix is generated from the normal distribution $\mathcal{N}(0, 1)$.

The average elapsed time of SCSVD is 3.98 seconds, while the economical SVD takes 16.14 seconds, If we increase the matrix to $p = 20000$, $n = 20000$ and the same rank $r = 50$, the elapsed time of economical SVD is 1209.92 seconds, but SCSVD is only 195.85 seconds. We observe that our SCSVD method demonstrates significant improvement.

Note that when the estimated rank used in SCSVD is greater than the real rank of data matrix, there is almost no error (except rounding error) between economic SVD and SCSVD. Figure 1 shows the speed comparison between economical SVD (solid line) and SCSVD (dashed line) with square matrix size from 500 to 4000 by fixed rank 50.

We also use fixed parameter $N_I = 51$ and $N_g = 2N_I$ in each simulation test. We can see that the computational cost of SVD follows the order 3 increase, compared with linear increase of SCSVD. The error between economical SVD and economical SVD, and that between SVD and SCSVD are shown in Figure 2. Because the results between economical SVD and SVD are very similar, we use solid line to represent the value of economical SVD and circle plot to represent SCSVD. The values in both Figure 1 and Figure 2 are the mean of the results from 100 repeated simulated matrices. The errors between SVD and economic SVD, and that between SVD and SCSVD are all under the 10^{-4} level. Thus, when the estimated rank of SCSVD is greater than the true rank, the accuracy of SCSVD is pretty much the same as SVD in the case of small rank matrix.

The purpose of the second simulation experiment is to observe the approximation performance of applying SCPCA to big full rank matrix. We generate random matrix with fixed number of columns and rows, say 1000. The square matrix is created by the form, $A_{p\times r} \cdot B_{r\times n} + \alpha E_{p\times n}$, where r is the essential rank, E is the perturbation and α is a small coefficient for adjusting the influence to the previous matrix. Such matrix can be considered as a big sized matrix with small rank added by a full rank perturbation matrix. We will show that our method works well for this type of matrices.

Figure 3 shows the error vs. estimated rank, where the error is computed by the difference between the original matrix and the composition of three matrices from SCSVD. All the elements of matrices A, B and E are randomly generated from the normal distribution $\mathcal{N}(0, 1)$, where $\alpha = 0.01$ and the essential rank $r = 50$. We can see that when the estimated rank increases, the composition error decreases. Especially when the estimated rank is greater than the essential rank r, the composition error decays rapidly. Thus, it is important to make sure that the estimated rank is greater than the essential rank. In other words, when the estimated rank of SCSVD is smaller than the essential rank, our SCSVD result can be used as the approximated solution of SVD.

In the last experimental result, we will show that we can set the estimated rank $r = 3$, starting from the SCSVD result and using the previous updating method to continuously update the new SVD. We will show that the performance of the first three components decays very slowly. Thus, many SVD-based modern techniques, for example, Fisher linear discrimination, Latent semantic analysis [5], eigen-taste recommendation system [6], dimensional reduction, etc, become feasible even when dealing with huge data set.

We produce a series square random matrices A with size n-by-n for n between 1000 and 3000. Then we decompose A by SVD to obtain $A = Z\Sigma V^T$. We reset the diagonal terms of Σ to be exponential decay, so that the data can simulate the meaningful data in the real world. The maximal spectrum is set to be 10^4. Then we compose A by the new diagonal matrix V. We use SCSVD with estimated rank 3, and the parameter $N_I = \frac{n}{10}$, $N_g = 2N_I$.

We make 16 updates to the data, and each time we add 10% samples of original data. The new data is simulated from the normal distribution $\mathcal{N}(0, 1)$. We use our updating method to compute the first three new columns of Z and compare it with the true SVD result. Let $a^{(t)}$, $b^{(t)}$ be the maximal and minimal element of the absolute values of $\hat{Z}_3^{(t)T} Z_3^{(t)}$, respectively, where $\hat{Z}_3^{(t)}$ is the t-th updated Z by our updating method taking only the first three columns, and $Z^{(t)}$ is the t-th updated Z by normal SVD. If $a^{(t)}$ and $b^{(t)}$ are close to 1, the updated Z derived by our updating method is very close to the true Z. In Figure 4, we can see that both $a^{(t)}$ and $b^{t)}$ are close to 1, and they decay very slowly as the matrix size increases. In Figure 4, every point is the average value of 32 repeating simulations.

V. Conclusion and Future Work

We proposed fast PCA and SVD methods derived from the technique of SCMDS method. The new PCA and SVD have the same accuracy as the traditional PCA and SVD ones when the rank of a matrix is much smaller than its matrix size. The results of applying SCPCA and SCSVD to a full rank matrix are also quite reliable when the essential rank of the matrix is much smaller than its matrix size. In most information technology applications, the essential rank of a matrix is usually much smaller than its matrix size. In such cases, utilizing SCPCA or SCSVD in huge data applications will render good approximated results. Since the concept of split-and-combine is very similar to that of parallel computing, this SC-series methods (Split-and-combine series) can be easily implemented via parallel computing. Using our updating method for the growing data, we show that the approximated solution is very close to the actual solution, even when the estimated rank is as small as $r = 3$.

For the future work, we will focus on the cases when the data contains missing values. Our intuitive speculation is that the processing of splitting data should be somehow related to the locations where the missing values occur. We believe that it would be an interesting topic worth further exploration.

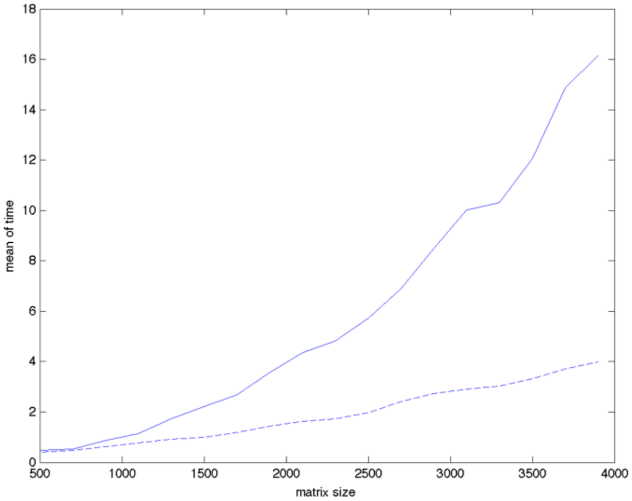


Figure 1. Comparison of the elapsed time between economical SVD (the solid line) and SCSVD (the dashed line).

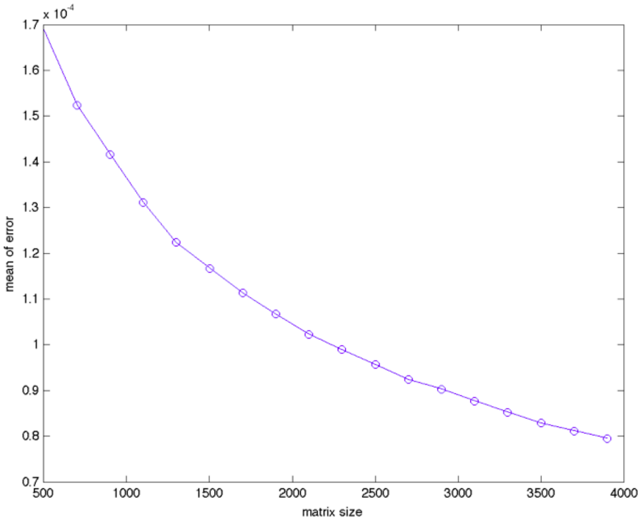


Figure 2. Comparison of the composition errors between economical SVD (the solid line) and SCSVD (the circle plot).

References

[1] D. J. Hand, "Discrimination and classification", Wiley Series in Probability and Mathematical Statistics, Chichester: Wiley, 1981

[2] M. Cox, T. Cox "Multidimensional scaling", Handbook of data visualization, Springer, 2008

[3] A. Morrison, G. Ross and M. Chalmers, "Fast multidimensional scaling through sampling, springs and interpolation",Information Visualization, Vol. 2 , Issue 1, pp. 68 - 77, 2003

[4] J. Tzeng, H. Lu and W. Li, "Multidimensional scaling for large genomic data sets", BMC Bioinformatics, 9:179, 2008

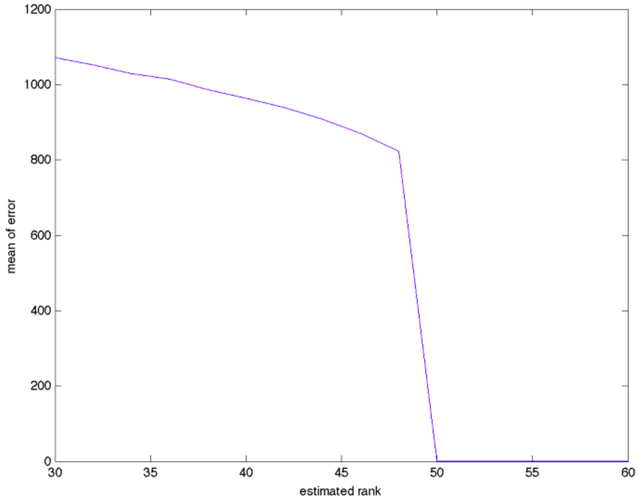


Figure 3. The effect of estimated rank to the composition error. The matrix size is 1000-by-1000 and its essential rank is 50 ($\alpha = 0.01$). When the estimated rank is greater than 50, there is almost no composition error

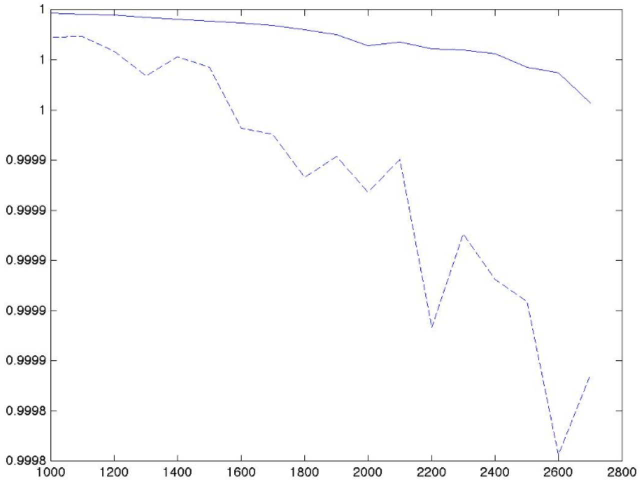


Figure 4. The orthogonality between approximated SVD and true SVD. The solid line is $a^{(t)}$ and the dashed line is $b^{(t)}$

[5] G. W. Furnas, S. Deerwester, S. T. Dumais, T. K. Landauer, R.A. Harshman, L. A. Streeter and K. E. Lochbaum, "Information retrieval using a singular value decomposition model of latent semantic structure", Annual ACM conference on Research and Development in Information Retrieval, pp. 465-480, 1988

[6] K. Goldberg, T. Roeder, D. Gupta and C. Perkins, "Eigentaste: A constant time collaborative filtering method", Information Retrieval, Springer, 2001

Physical Layer Measurements for an 802.11 Wireless Mesh Network Testbed

Stanley W. K. Ng
Faculty of Electrical and Computer Engineering
McMaster University
Hamilton, Canada
ngswk@mcmaster.ca

Ted H. Szymanski
Faculty of Electrical and Computer Engineering
McMaster University
Hamilton, Canada
teds@mcmaster.ca

***Abstract*—Physical layer measurements for an infrastructure 802.11 Multichannel MultiBand Wireless Mesh Network testbed are described. Each wireless router node design consists of a Linux processor with multiple 802.11b/g transceivers operating in the 5 GHz band for backhauling, and multiple 802.11 transceivers in the 2.4 GHz band for end-user service. Each transceiver consists of a MAC and base-band processor (BBP) in addition to a radio. A Linux-based device driver has been modified to adjust the physical layer parameters. The 802.11 standard specifies three orthogonal channels, 1, 6, and 11. The routers can be programmed to implement any static mesh binary tree topology by assigning orthogonal frequency-division multiplexing (OFDM) channels to network edges. The routers can be programmed to implement any general mesh communication topology by using a time division multiple access (TDMA) frame schedule, and assigning OFDM channels to network edges within each TDMA frame. Preliminary measurements of co-channel interference and the signal to interference and noise (SINR) ratio for the network testbed are presented, using omni-directional antenna and the 802.11b operation mode. This data can be used to optimize the performance of large infrastructure Wireless Mesh networks using 802.11 technology.**

***Index Terms*—wireless mesh network; 802.11; co-channel interference; noise; SINR;**

I. Introduction

Multihop infrastructure wireless mesh networks (WMNs) as shown in Fig. 1 represent a low-cost access network technology, which can potentially provide ’last-mile’ accessibility to much of the world. Industry estimates that by 2020 there will be several billion wireless devices, providing a range of new services. WMNs represents a promising infrastructure for supporting these wireless devices, as well providing general communications infrastructure for homes and offices. However, capacity and scalability are key challenges for such networks. Multichannel multiband meshes can use multiple radio channels in multiple frequency bands to improve system capacity and throughput. For example, channels in the 5 GHz band can implement the mesh backhauling trees between Base-Stations (BSs) in Fig. 1, and channels in the 2.4 GHz band can implement the communications between the Base-Stations and end-users. The optimized design of such WMNs requires statistics on physical layer noise and co-channel interference. However, to date there have been very few published measurements for co-channel interference and signal-to-interference-and-noise ratios encountered in practical WMN testbeds. To address this problem, an 802.11 WMN testbed has been developed and detailed noise and interference measurements are reported.

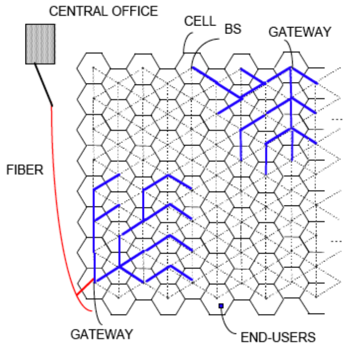
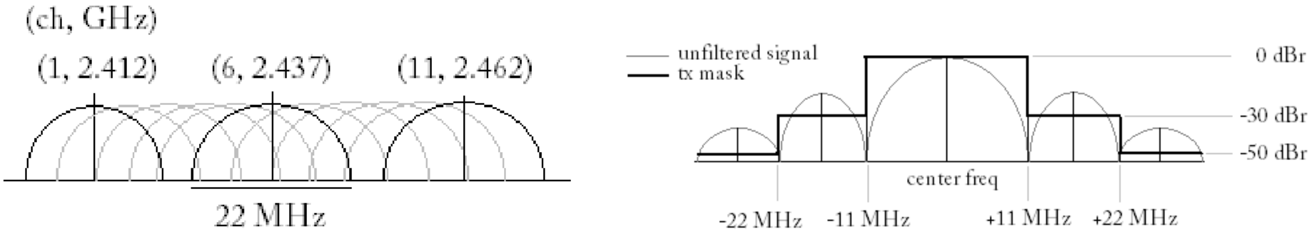


Fig. 1. A Wireless Mesh Network using a wireless cellular array.

This paper presents a small mesh testbed composed of IEEE 802.11b nodes operating in IBSS (ad-hoc) mode, called the *Next-Generation (NG) Mesh*. Each network node contains multiple wireless transceivers controlled in a Linux environment. To study co-channel interference between WiFi channels in the 2.4 GHz ISM band, preliminary physical layer (PHY) measurements for the received signal strength indicator (RSSI) and SINR are reported. The IEEE has specified standards for co-channel interference in 802.11 standard [1]. Fig. 2a illustrates the 11 channels in the 802.11 WiFi standard in the 2.4 GHz band. Each channel requires 22 MHz and channels are separated by 5 MHz. Channels 1, 6 and 11 are logically orthogonal, i.e., their spectrum is non-overlapping. The IEEE 802.11b spectral mask shown in Fig. 2b mandates a drop of at least 30 dBm at a displacement of 11 MHz (two channels) from the active channel. The standard also mandates a drop of at least 50 dBm at a displacement of 22 MHz (four channels). The spectral mask requirement ensures adequate attenuation between 802.11 channels in the ISM band. However, the spectral mask requirements apply to a single device tested in isolation, and will not apply to a real network deployment due to interference from multiple networks and other microwave devices.

In order to test the co-channel interference and SINRs in



(a) The 2.4GHz ISM Band.

(b) The IEEE 802.11b spectral mask.

Fig. 2. The IEEE channel map.

a practical 802.11 network deployment, a large file transfer was performed over one channel in the NG Mesh testbed, and the interference on the other 10 WiFi channels was measured. Using 15 dBm of transmission power over very short (5 meter) wireless links and a datarate of 11 Mbps, SINRs in that range of 20 - 30 dBm were consistently measured. In a large deployment these SINRs are typically reduced according to the distance cubed. Furthermore, the wireless link quality was largely static with no noticeable changes over any 24 hour period, except for changes in activity in remote 802.11 networks. Given the lack of published data on SINRs in realistic network testbeds, this data may help optimize network designs.

Section 2 summarizes other recent wireless networking testbeds that report physical layer metrics. Section 3 describes the design of the NG-Mesh. Section 4 presents the physical layer measurements of NG-Mesh. Section 5 closes with our conclusion and future work.

II. Related Work

Some other recently published wireless testbed designs that consider physical layer measurements are discussed next. Like NG-Mesh, all of them employ a Linux OS and multiple IEEE 802.11 wireless transceivers at each network node. In [2], a wireless network was installed in a natural reserve spanning several kilometers to collect ecological data. Each mesh node consists of multiple 802.11b/g wireless transceivers capable of multi-channel communication. The testbed was used to investigate the correlation between packet error rate (PER) and RSSI, to evaluate improvements in rate control algorithms and routing protocols. In [3], a campus-wide 802.11b/g WMN was used to provide internet access. Due to the close proximity of the mesh nodes, a high degree of co-channel interference was present. Although physical layer metrics such as the signal to noise ratio (SNR) were collected, no solution was proposed to mitigate the interference.

In [4], a mesh network composed of quickly deployable relay nodes was formed to target real-time emergency communication services. SNR measurements were used to infer link quality to guide the deployment of additional relay nodes. A case study concluded that SNR statistics reported by commercially available radios accurately indicate link reliability. In contrast to PER statistics, which report transmission errors that have already occurred, SNR measurements provide predictions of link failures before they occur. In [5], an 802.11a testbed was used to measure handoff latency of high-speed vehicles roaming between stationary road-side access points (APs). Handoff between APs occur whenever the link between the mobile station (MS) and the currently associated AP exhibits an RSSI that is lower than a pre-defined threshold.

In [6], a formula was derived for the amount of signal attenuation required for a small-scale network testbed to emulate a large-scale network, reflecting the different inter-node spacing. The transmit power and the inter-node spacings of the wireless interfaces were scaled down accordingly, allowing a small testbed (spanning a few meters) to emulate the large network for testing purposes. Variable attenuators (Broadwave 751-002-030 devices) were used. In contrast, the NG-Mesh testbed employs a modified Serialmonkey [7] Linux-based device in order to control the physical layer characteristics including the transmission frequency (WiFi channel) and transmission power.

In summary, while a number of 802.11 testbeds have been developed to date there are no detailed measurements of co-channel interference and SINRs in such network testbeds. In contrast to the previous testbeds, the NG-Mesh will be used to collect RSSI and SINR measurements on each channel in the ISM spectrum given an isolated file transfer on one specific WiFi channel. This data may be useful to optimize the performance of future WiFi WMN deployments.

III. The NG-Mesh Architecture

The NG Mesh testbed was developed from commercially available WiFi components and software. Table I summarizes the system components making up a logical node in our testbed. The *product_ID:vendor_ID* codes correspond to the Linksys WUSB54G and Wi-Spy 2.4i devices.

TABLE I
Testbed Components

Component	Specifications
PC	Q9300 2.53GHz CPU, 4GB RAM, 100GB HDD
VM	VMware 7, 512MB RAM, 1CPU, 8GB HDD
OS	64-bit RHEL 5.4, kernel 2.6.18
Wireless Interfaces	13b1:000d, RT2570/RT2525E chipset, rt2x00 Driver
Spectrum Analyzer	1dd5:2400, Chanalyzer Lite / Kismet Software

A. The Testbed Configuration

As shown in Fig. 3, one node in the NG-mesh testbed consists of four Linksys WUSB54Gv4 wireless transceivers that connect to one PC via a non-attenuating four-port USB hub. The 4 wireless transceivers provide the capability to transmit or receive on $c \leq 4$ channels simultaneously. For example, each node can be configured to permanently receive on 2 arbitrary channels and transmit on one arbitrary channel. Alternatively, each node can be programmed to transmit or receive on selected channels in selected intervals of time, thereby implementing a *Time Division Multiple Access (TDMA)* style of mesh network.

To perform our experiments, the transceivers associated with one node were placed 5 meters apart. One pair of transceivers was configured to transmit a large file over a given WiFi channel using a UDP socket. Another transceiver was configured as an interference measurement device, to receive on a different WiFi channel from which RSSI and SINR measurements could be made. The transceivers in the NG-Mesh are configured to operate in ad-hoc mode with each being assigned a unique IP address and each tunable to different radio frequencies. The configuration of all network transceivers are performed via the central PC console using Linux network configuration utilities such as *iwconfig*.

The testbed can be extended in several ways. More wireless transceivers can be connected to one node, thereby enabling one node with a small footprint to emulate a larger virtual network. Alternatively, multiple NG mesh nodes can be deployed over a large geographic region.

B. The Wireless Interfaces

The Linksys WUSB54Gv4 transceiver has a retractable 2 dBi omni-directional antenna and requires no proprietary firmware. Each Linksys transceiver contains the Ralink RT2570 (MAC/BBP) and RT2525E (transceiver) chipsets. In the 2.4 GHz band, the radio is tunable between 2.412 GHz and 2.484 GHz. In the 5 GHz band, the radio is tunable between 5.180 GHz and 5.805 GHz. The allowable channel subsets are determined by the PHY mode and geographic region. The PHY modes of IEEE 802.11a, b, and g are referenced in the driver source code but by default, the PHY mode and channel are set to IEEE 802.11b and 1 respectively. The bit-rates supported by this chipset are 1, 2, 5.5, 6, 9, 11, 12, 18, 24, 36, 48, and 54 Mbps although it was found that only a maximum bit-rate of 11 Mbps was configurable in ad-hoc mode. The maximum transmission power output of this chipset is 100 mw (20 dBm). According to specifications [8], the receiver sensitivity is nominally in the range of -65 dBm to -80 dBm.

Within the Linux community, there is an active project maintained by a team of developers known as Serialmonkey who maintains and enhances open-source Linux-based device drivers for Ralink chipsets. Their legacy drivers from the May 12, 2009 build were selected to operate the wireless interfaces of NG-Mesh. All Serialmonkey [7] and Ralink driver code bases [9] are written in the C programming language.

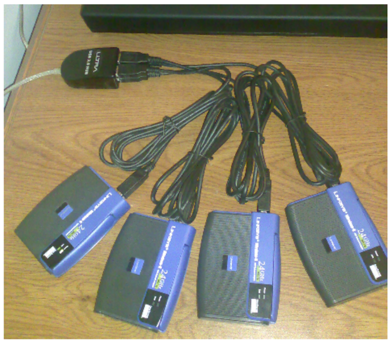


Fig. 3. Node configuration.

The Serialmonkey legacy drivers for Ralink chipsets rely on a periodically executed code segment within a function called *CMDHandler*. By implementing a channel switching function call *AsicSwitchChannel* in software in the *CMDHandler*, a maximum frequency retuning rate of one channel switch every 30 ms was achieved. This measurement indicates that a TDMA-based WiFi network can achieve channel changes at rates of 33 Hz in software. According to the IEEE standard [1], much faster channel change times of 224 μsec can be achieved in hardware.

C. The Wi-Spy Spectrum Analyzer

Our preliminary measurements were made using an inexpensive MetaGeek Wi-Spy 2.4i entry-level ISM spectrum analyzer. The software is freely downloadable from [10] [11] for Windows, Mac, or Linux and can be used to plot real-time distributions of the energy in the ISM band. The software also computes the average and peak energies over time per frequency, and can also use a separate wireless interface to detect proximate WiFi networks. The MetaGeek Chanalyzer Lite software enables this device to scan the frequency range of 2.4 - 2.492 GHz and report the energy in the range -102 - 6.5 dBm, in 375 kHz and 0.5 dBm increments respectively, roughly once per second. Under Linux, Kismet's Spectrum-Tools software enables the device to scan the frequency range of 2.4 - 2.483 GHz and report four-hundred-and-nineteen samples every 30 ms in 199 KHz steps.

IV. Experimental Results

A communication link was established by configuring one transceiver to transmit on the primary WiFi channel, and a second transceiver to receive on the same channel. Another transceiver was configured to receive on a secondary channel, to obtain co-channel interference measurements. The Wi-Spy spectrum analyzer was also configured to monitor the secondary channel. The transmit power of both communication endpoints were fixed at 15 dBm via device driver modifications while all file transmissions were performed at the 802.11b bit-rate of 11 Mbps. All of the subsequent experiments took place in a room measuring sixteen square meters within a residential neighborhood. Remote APs were detected on channels 1 (-92 dBm), 6 (-84 dBm), and 11 (-97 dBm) at the test site during testing.

In each experiment, a large 100 MB file was transmitted over the primary WiFi channel using UDP socket transfers. The background noise level was measured using one of the RT2570 interfaces operating in monitor mode and was verified using the Wi-Spy spectrum analyzer. The average noise level at the test site was reported by this device to be -99 dBm, with some occasional fluctuations at various channels, especially 1, 6, and 11, due to intermittent activity on remote networks. The monitor nodes also reported the power on each channel due to activity on remote WiFi networks. Most remote channels had powers in the range of -100 to -90 dBm, which we believe were beacon powers. These power measurements were not reported as noise by the device. However, spurious emissions with magnitudes of -80 or -70 dBm were observed on many channels, perhaps several times a minute. We believe that these spurious emissions were caused by remote user activity, though no exact measurements of remote beacons and noise were made due to the lack of access to state-of-the-art test equipment in our preliminary experiments.

Table II and Table III organize the RSSI and SINR data as reported by the Wi-Spy device during file transmission on each of the wireless channels, into a 2D matrix. The RSSI data in the tables were collected using Kismet's spectool_raw utility. A shell script was written to compute the average of the RSSI readings in each channel during file transfer activities. The SINR matrix was then computed from the RSSI matrix using: $SINR_{dBm} = Signal_{dBm} - (Noise + Interference)_{dBm}$. As confirmed earlier, the background noise level was reported by the monitor nodes to be -99 dBm most of the time. The interference power was reported per channel separately. The highlighted main diagonal in the matrices represents the active channels. Additional data on the RSSI and SINR are reported in the 2D and 3D MATLAB plots shown in Fig. 4 to Fig. 6.

In Fig. 5 and 6, each uniquely color-coded line represents a file transfer conducted at a particular WiFi channel. Given the file transfer on one channel, the signal strength was measured on each of the other channels using the Wi-Spy spectral analyzer.

Referring to the tables, the peak signal energy is always at the primary channel. Moving two channels away, i.e., a displacement of approx. 11 MHz, the signal attenuation is approx. 20 dB, yielding interference about 10 dBm above the IEEE spectral mask requirement. Moving 4 channels away causes an additional attenuation of (0 to 5) dB, yielding interference about 20 dBm above the IEEE spectral mask requirement. These results are likely due to interference from remote WiFi activity. The attenuation requirements specified by the IEEE 802.11b spectral mask shown in Fig. 2b apply to one device in isolation, and interference from other devices limits the SINRs in practice. Measurements of co-channel interference and SINRs for typical WiFi testbeds has not previously been quantified and published, so we cannot compare our results with any others.

Fig. 7a demonstrates per-channel activity in real-time along with a spectral density plot. These results were produced by the Chanalyzer Lite software under the Windows OS. Fig. 7b illustrates the power levels at different frequencies of the ISM spectrum during file transfers. The *test1* configuration uses a file transfer on wireless channel 1. The red specs represent the most frequently occurring energy readings. Therefore, during a file transmission on channel 1 the red outline indicates the mean RSSI levels to be around -70 dBm, with the RSSI dropping to about -100 dBm two channels away which conforms to the 802.11b spectral mask requirements.

Although the 802.11b bit-rate was configured as 11 Mbps, the average bit-rate was calculated from empirical measurements for verification. The average effective data rate was 3.96 Mbps, only 36.01% of the configured speed, which roughly coincides with the results of previously published data [6].

Referring to the 2D matrices and Fig. 4, 5, 6 and 7, the co-channel interference drops by about 20 dBm when the secondary channel is 2 channels away from the primary, as stated earlier. However, the interference increases when the secondary channel is 3 channels away. These observations are inconsistent with the IEEE spectral mask requirements. This pattern is repeated for every primary channel from 1-11, indicating that it cannot be explained by uncorrelated remote user activity. Referring to Fig. 2b, the spectral power density before the mast is applied is shown by the sinc() curves, and the interference power appears to increase at a distance of 3 channels. However, after the mask is applied the interference power should decrease at a distance of 3 channels. We compared our preliminary measurements with theoretical results on co-channel interference presented in [12], with similar observations. We plan to secure access to state-of-the-art measurement equipment and repeat the measurements, to precisely quantify the effects of remote WiFi activity.

V. Conclusion and Future Work

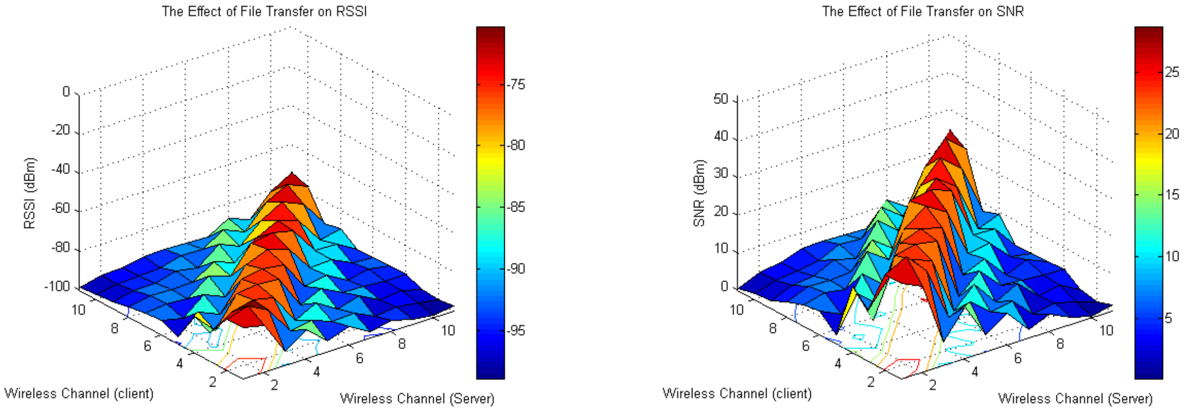
An 802.11b wireless mesh network testbed was developed using commercially available wireless transceivers and software. A single node design consists of multiple wireless transceivers which can be individually configured, i.e., the transmission / reception frequency, the transmission power, and data rate of each transceiver can be configured from a device driver in the Linux OS. Each node can be extended in several ways. Multiple wireless transceivers can be added to one node, to emulate a larger network virtually. Alternatively, multiple nodes can be deployed over a large geographic area, to implement a large mesh network as shown in Fig. 1. The optimization of a large network requires realistic statistics on co-channel interference and SINRs. Our testbed was configured to enable a large UDP file transfer on one channel over a short distance, and RSSI and SINR measurements were recorded on all other channels. Our measurements indicate that at a displacement of 2 channels or 10 MHz, the signal attenuation is -20 dBm. At a displacement of 4 channels or 20 MHz, the signal attenuation varies from -20 dBm to -25 dBm. This data indicates that interference from other WiFi networks and other microwave devices in the 2.4 GHz band is noticable in a

TABLE II
WI-SPY RSSI DATA (DBM)

Client Server	ch.1	ch.2	ch.3	ch.4	ch.5	ch.6	ch.7	ch.8	ch.9	ch.10	ch.11
ch.1	**-73.3302**	-75.9277	-92.6226	-85.1258	-94.3459	-90.2987	-96.5094	-94.1855	-96.9057	-98.4025	-97.3459
ch.2	-73.4455	**-70.2885**	-75.8846	-90.1314	-87.9647	-92.4359	-91.4103	-96.7564	-95.3974	-96.9263	-98.9359
ch.3	-85.1741	-77.8892	**-75.2816**	-77.3924	-92.6424	-87.4842	-94.0728	-90.75	-96.0696	-95.1709	-97.9272
ch.4	-81.2097	-92.0516	-76.6419	**-75.329**	-76.4194	-93.1742	-87.9548	-93.6548	-94.2032	-94.6935	-94.7806
ch.5	-96.5159	-87.3535	-91.8662	-77.5414	**-75.6401**	-76.9682	-91.5159	-88.9936	-95.1879	-92.9586	-95.2675
ch.6	-92.5466	-95.135	-86.2605	-90.8296	-77.9711	**-73.7235**	-77.0643	-89.8682	-89.1576	-94.4373	-95.4019
ch.7	-93.2748	-91.9457	-92.8978	-84.607	-88.4633	-74.754	**-73.4377**	-75.5272	-89.147	-89.5335	-95.5495
ch.8	-94.9108	-92.4936	-92.7803	-92.4936	-85.7038	-89.6338	-80.2834	**-74.6561**	-78.6688	-91.8631	-89.6752
ch.9	-96.9455	-94.6314	-94.0256	-91.5577	-93.0897	-84.7596	-89.2853	-79.3109	**-73.3205**	-78.2468	-91.4263
ch.10	-97.3462	-95.6282	-95.0897	-94.1538	-91.5032	-92.0192	-85.2853	-89.6859	-78.2756	**-72.4231**	-77.1859
ch.11	-98.4487	-95.859	-95.2179	-93.2276	-93.0577	-93.2692	-93.4551	-87.5897	-91.4038	-80.9583	**-74.5032**

TABLE III
WI-SPY SINR DATA (DBM)

Client Server	ch.1	ch.2	ch.3	ch.4	ch.5	ch.6	ch.7	ch.8	ch.9	ch.10	ch.11
ch.1	**25.6698**	23.0723	6.3774	13.8742	4.6541	8.7013	2.4906	4.8145	2.0943	0.5975	1.6541
ch.2	25.5545	**28.7115**	23.1154	8.8686	11.0353	6.5641	7.5897	2.2436	3.6026	2.0737	0.0641
ch.3	13.8259	21.1108	**23.7184**	21.6076	6.3576	11.5158	4.9272	8.25	2.9304	3.8291	1.0728
ch.4	17.7903	6.9484	22.3581	**23.671**	22.5806	5.8258	11.0452	5.3452	4.7968	4.3065	4.2194
ch.5	2.4841	11.6465	7.1338	21.4586	**23.3599**	22.0318	7.4841	10.0064	3.8121	6.0414	3.7325
ch.6	6.4534	3.865	12.7395	8.1704	21.0289	**25.2765**	21.9357	9.1318	9.8424	4.5627	3.5981
ch.7	5.7252	7.0543	6.1022	14.393	10.5367	24.246	**25.5623**	23.4728	9.853	9.4665	3.4505
ch.8	4.0892	6.5064	6.2197	6.5064	13.2962	9.3662	18.7166	**24.3439**	20.3312	7.1369	9.3248
ch.9	2.0545	4.3686	4.9744	7.4423	5.9103	14.2404	9.7147	19.6891	**25.6795**	20.7532	7.5737
ch.10	1.6538	3.3718	3.9103	4.8462	7.4968	6.9808	13.7147	9.3141	20.7244	**26.5769**	21.8141
ch.11	0.5513	3.141	3.7821	5.7724	5.9423	5.7308	5.5449	11.4103	7.5962	18.0417	**24.4968**



(a) RSSI MATLAB plot

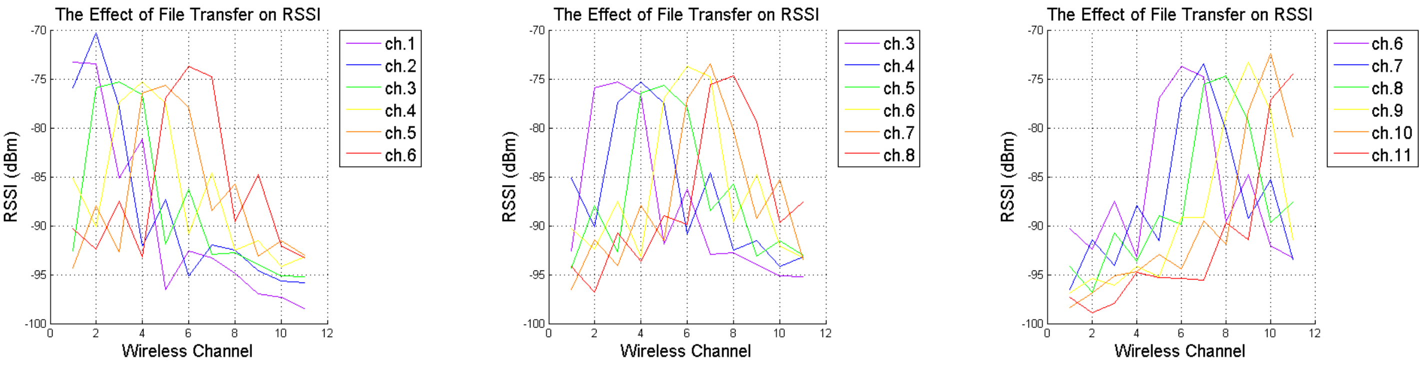
(b) SNR MATLAB plot

Fig. 4. The effect of file transfers on the ISM band.

practical mesh deployment. Our measurements were made in the early hours of the morning, and it is expected that the SINR may drop during regular business hours when more activity on remote WiFi networks will be present. We plan to quantify the interference due to remote users more thoroughly once access to state-of-the-art measuring equipment is secured. The work can be extended in several ways. The current testbed uses 802.11b transceivers with omni-directional antenna. Currently, open-source Linux-based device drivers for 802.11n devices are available but they are unstable. The same experiments can likely be performed within a year or two using the latest generation 802.11n transceivers, open-source Linux-based device drivers, with MIMO directional antenna. We expect 802.11n devices to offer a significant increase in throughput with improvements to the co-channel interference and SINR measurements.

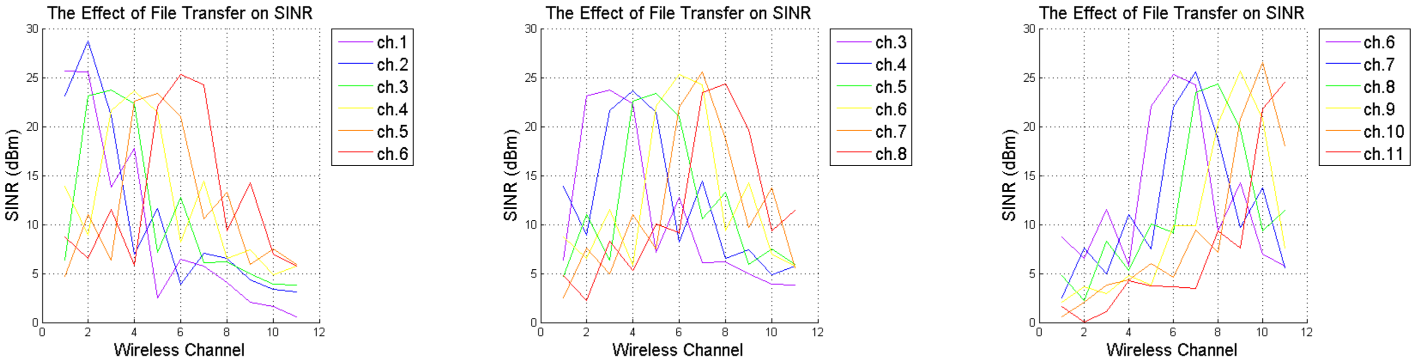
REFERENCES

[1] "IEEE Std 802.11", Second Edition, pp. 345-346, 2005.



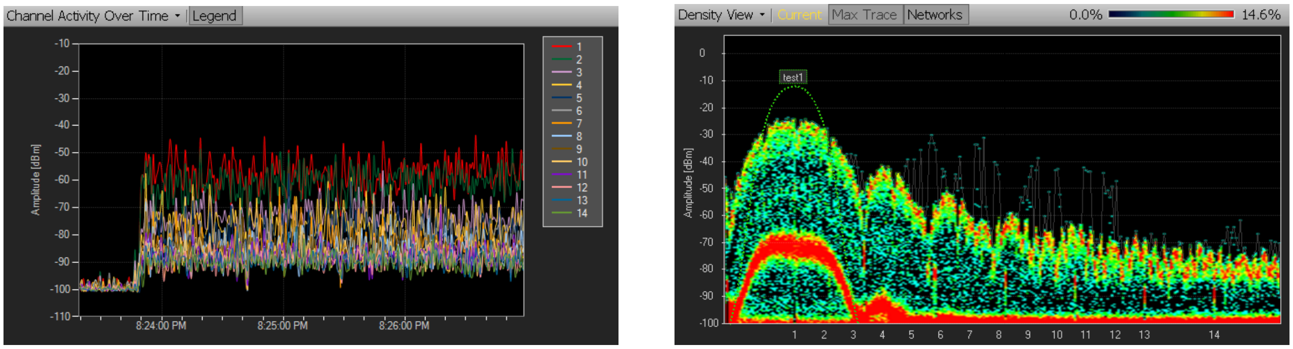
(a) File transfers on channels 1-6 (b) File transfers on channels 3-8 (c) File transfers on channels 6-11

Fig. 5. The effect of file transfer on RSSI.



(a) File transfers on channels 1-6 (b) File transfers on channels 3-8 (c) File transfers on channels 6-11

Fig. 6. The effect of file transfer on SINR.



(a) RSSI Before and During File Transmission (b) RSSI Distribution Due to File Transmission

Fig. 7. The effect of a channel 3 file transfer on the ISM band.

[2] D. Wu and P. Mohapatra, ”QuRiNet: A wide-area wireless mesh testbed for research and experimental evaluations”, *IEEE Communication Systems and Networks (COMSNETS)* , pp. 1-10, 2010.

[3] S. L. Shrestha, J. Lee, A. Lee, K. Lee, and S. Chong, ”An Open Wireless Mesh Testbed Architecture with Data Collection and Software Distribution Platform,” *IEEE Testbeds and Research Infrastructure for the Development of Networks and Communities* (TridentCom), pp. 1-10, 2007.

[4] M. R. Souryal, A. Wapf, and N. Moayeri, ”Rapidly-Deployable Mesh Network Testbed”, IEEE GlobeCom , pp. 1-6, 2009.

[5] Y. Tarte, A. Amanna, and C. Okwudiafor, ”Experimental testbed for investigating IEEE 802.11 handoff in vehicular environment”, *IEEE SoutheastCon 2010 (SoutheastCon)* , pp. 222-225, 2010.

[6] S. M. ElRakabawy, S. Frohn, and C. Lindemann, ”ScaleMesh: A Scalable Dual-Radio Wireless Mesh Testbed”, *IEEE SECON* , pp. 1-6, 2008.

[7] http://rt2x00.serialmonkey.com/wiki/index.php/Downloads 12.20.2010

[8] http://www.linksysbycisco.com/APAC/en/support/WUSB54G/download 12.20.2010

[9] http://www.ralinktech.com/support.php?s=2 12.20.2010

[10] http://www.wi-spy.ca/wispy24i.php 12.20.2010

[11] http://www.kismetwireless.net/spectools/ 12.20.2010

[12] A. Z. Al-Banna, T. R. Lee, J. L. LoCicero, and D. R. Ucci,”11 Mbps CCK - Modulated 802.11b Wi-Fi: Spectral Signature and Interference”, IEEE Int. Conf. Electro/information Technology, pp. 313-317, 2006

Insights from the Technology System Method for the Development Architecture of e-Textbooks

Ching-Cha Hsieh[1, a], Jen Wel Chen[1,2, b], Dershing Luo[1,3, c]
[1]Department of Information Management
National Taiwan University
[2]Department of Business Administration
Chinese Culture University
[3]Department of Information Management,
China University of Science and Technology
Taipei, Taiwan (R.O.C.)
[a] cchsieh@im.ntu.edu.tw
[b] jwchen@sce.pccu.edu.tw
[c] luoder2008@gmail.com

Chia-Ching Lu
Institute for Information Industry
Taipei, Taiwan (R.O.C.)
gaty@iii.org.tw

Yun Long Huang
Department of Recreation and Leisure Industry Management
National Taiwan Sport University
Taoyuan, Taiwan (R.O.C.)
ylhuang@mail.ntsu.edu.tw

***Abstract*—Textbook market is considered to be the next wave of digital-reading. However, the textbooks have their own learning purposes, user habits and applications. Whether the development of electronic textbooks (e-Textbook) may continue the current path of development of e-book technology is of concern in this stud. Besides, from the experiences of e-book development, we aim to build the next wave of e-textbook system architecture. This study analyzes key milestones in the e-book history to explore the success of Amazon Kindle and its characteristics. Through the concept of technology systems proposed by Thomas P. Hughes, we elaborate a framework for the development of e-Textbooks.**

Keywords - e-Textbooks; Development Architecture; Technology System

I. Introduction

In 2007, Amazon launched Kindle e-book reader with its online sales mechanism, digital reading technology and hardware [1,2]. It triggers the development of e-books and digital reading of the trend in global. The market research firm, In-Stat, forecasts, the global e-book device shipments in 2008 was about one million units and will be nearly thirty million units in 2013. The revenue is more than 1.1 billion U.S. dollars [3]. The largest user groups--textbooks, will become the next market of e-book. The research institution, Forrester, estimates that the textbook will be the largest market for e-book reader in five years [4].

However, there are specific attributes of textbooks: learning purposes, user habits and application developments. We concern about how to build the next generation of e-textbooks system based on current technology.

The remainder of this paper is organized as follows. Section II reviews the literature related to key development of e-books. Section III describes analytical framework of e-book development and surveys the characteristics of Amazon Kindle. Section IV introduces the concepts of technology system, then we analysis Amazon kindle's functions and services based on Hughes's view of technology system. In Section V, we propose our framework of e-Textbooks through the technology system. The conclusion and future suggestion are offered in Section VI.

II. Literature Review

The development of e-books has been already 42 years since Alan Kay proposed the concept of Dynabook [5,6,7] e-books in 1968. Because of exertion from different concepts, methods and technology by academia, industry and government, e-books presents rich and diverse definitions and appearance [8,9,10]. E-books can be denoted as digitized books, online databases, CDROM or portable storage media content, browse information on portable hardware device or the digital content. Therefore the orientation of e-books could be digital content (content), integrated device (composer), vehicle (container), access interface (channel), or mechanism (consultant) [7]. Because of different definitions, positioning and technological, e-books have been confined in small minority groups or different areas in the past decades [11].

Because Amazon launched Kindle e-book reader with its online sales mechanism in 2007, e-books and digital reading reached a new situation. The electronics giant Sony restarted its e-book development: promoting new e-book products and following e-PUB specifications [12]. At the same time, Barnes and Noble launched its own brand of e-book reader Nook in 2009. E-reader has successfully established the concept of digital reading. It is an inevitable trend that the electronic textbook – combined of learning and textbooks will be the next largest market for e-book readers..

III. ANALYTICAL FRAMEWORK OF E-BOOKS DEVELOPMENT

Textbooks are basically paper books. They seem easy to follow the development of e-books flourish. In order to understand whether e-Textbooks can follow the same trend, we review several representative electronic keys to the development process.

In 1945, Bush proposed Memex concept of a working table installed in the microfilm equipment and readers to provide people to unrestrained and easily collect, view and search information [7,13]. In 1968, Allen Kay, working with PARC, proposed the concept of personal portable notebook – Dynabook. Dynabook is like a book, it can carry possessed interaction and run teaching software and reading materials, which is designed by Smalltalk [14]. Smalltalk is an object-oriented programming language. It concretely presents the prototype of today's electronic platform, Dynabook is considered to be the beginning of a portable e-book [5,6,7]. In 1971, Hart noted the concept of digital books have great impact on exchange of knowledge [15]. He began to promote the Gutenberg project. Some books do not advocate copyright, especially classical literature, religious books and references to books. They were scanned as image files by volunteers and then compiled manually enter the nuclear school e-book text file, via the Internet access methods for free [7,16]. In 1976, one project of Oxford University – Oxford Text Archive, planned to provide students to download electronic textbook files to speed up the exchange of knowledge and preservation [7,17]. In 1985, Egan's project SuperBook, developed new software for running in PC. The software can be stored on disk, tape and CD [7,18]. From 1998 to 2003, NuvoMedia release commercial products -- Rock eBook based on the concept of Dynabook [7,19]. Users could import digital data (special format) into Rock eBook for reading. In the mean time, there were products provided by other companies whose size is similar to books. Those products could download authorized files provided by company's web site through built-in modem or via adapter on PC. Because many readers reflected that it is tired for readers to read on the LCD display for a long time [20], a new technology – to display a screen similar to the paper was developed in the 1970s. It made e-book employed by e-ink being popular [20]. On March 2004, Sony released the first e-ink reader, Sony LIBRIé. At the end of 2007, Amazon launched Kindle [20].

The main e-book technology milestones of e-books are: Content (information), Composer (software necessary to create an e-book), Container (the distribution medium and / or file format), Storage, and Access (the technology and software used to provide access to the content).

We can generalize them into four parts: Digital content, Carrier, Storage, and Access, as depicted in Figure 1. We analysis the milestones of the development of eBook on Figure 2.

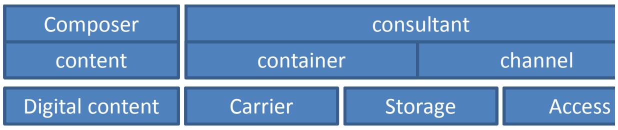


Figure 1. Inducted dimensions of e-books technology development

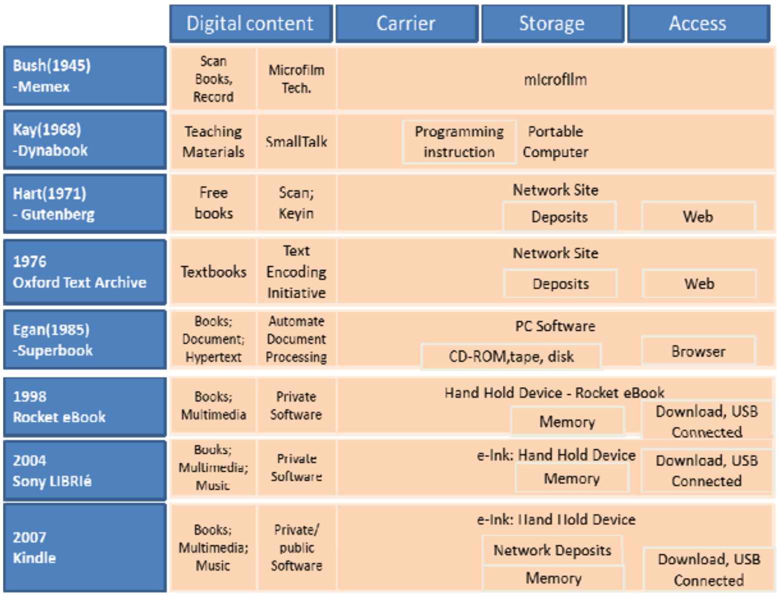


Figure 2. Development of e-Books

From the above analysis in the two figures, we can not find out the reasons why Amazon Kindle was sold more than Sony LIBRIé [21]. There should be other dimensions not having been yet evaluated.

IV. CONCEPTS OF TECHNOLOGY SYSTEM

The definitions of e-books vary in academic researches. As stated above, e-books can be digitized books, online databases, CDROM or portable storage media content, browse information on portable hardware device, or just digital content. Different definitions of the e-books and different cognition on Kindle's e-books product, much like Edison invented the light. The invention of the light acted as a representative of electricity in the era. Because impact of bring light to human, we took it as a key innovation-- effects in civilization. Because of the success of Amazon Kindle's digital reader, we took e-books as an ideal framework. We usually inference framework of e-Textbooks based on e-book. In this viewpoint, there is expected to be a single technological achievements and will made many R & D (research and development) staffs worked tirelessly to find a breakthrough innovation.

However, in 1989, Thomas P. Hughes [22] changed the view of invention for the light is the key to the development of civilization of view of electricity. Hughes studied the history of electrification in U.S. and proposed the concept of System Theory of Technology.

Based on Hughes's study, Edison not only invented the high-impedance filament technology but built up a concept--holistic conceptualization theory to consider large electrical equipment, transmission and distribution lines which did not exist at that time. Edison thought about technology of electric light bulb, systems cost of planning and installation. It made overall social atmosphere, with the government system to implement this innovative technology ideas. The team of Edison deliberately calculated the cost of scientific knowledge and experimental point of view the overall planning considerations, combined with inventors, engineers, management, finance, public relations professionals in different roles, completed with electric lighting system is one big target. Therefore the social, economic and technical can create a comprehensive system to replace gas lighting lamp lighting system, and thus contributed to the electrification of the United States. The concept of systems for innovation and technological development are of great inspiration, innovation and technology development can come from broader perspective thinking and planning thus better ensure the success of technological innovation.

Based on Hughes's technical system concept, the development of science and technology is not just technology, but also considers the management level and financial level of planning, which is very similar to the idea of John Law: 'The argument is that those who build artifacts do not concern themselves with artifacts alone, but must also consider the way in which the artifacts relate to social, economic, political, and scientific factors. All of these factors are interrelated, and all are potentially malleable [22; p112].

There are three dimensions in Thomas P. Hughes's technology system concept: 1. Inventor Entrepreneur, 2. Manager Entrepreneur, and 3. Financier Entrepreneur. But Hughes's study was based on the division of people. This paper considers John Law's claim and tries to add the fourth dimension of e-books technology development on Hughes's three concepts: technical, managerial and financial areas. In this way, we analyze the design, planning and systems architecture of Amazon Kindle. In order we can find out the reasons of its critical role in e-books.

We got following data from Amazon Kindle's web site as well as classified technology functions and services. We analysis them based on Hughes's view of technology system. The result is shown in Figure 3. Except the analysis of technology, we add two levels of management and financial, we can see clearly Amazon Kindle also provided a lot of functions and services in management and finance. Thus when we talk about development of e-books, in addition to focusing on technology for handheld devices, we should not neglect the technology-related management systems and financial supporting planning. Since a single light bulb could not be effective without a reasonable price, electricity transmission and distribution lines. Thus the following policies made by Amazon Kindle in leading position: to U.S. 9.9 dollars for digital-book, provide network management stacks, file conversion service, convenience and access to services of books, etc. The introduction of technical systems makes us gain more insights, thus we can have a clearer understanding on the Amazon Kindle's success.

V. BUSINESS SYSTEM ARCHITECTURE OF E-TEXTBOOKS UNDER TECHNOLOGY SYSTEM

Technology system provides a clear framework for understanding of technical functions. Therefore, we try to view through the technology system when thinking about e-Textbooks [19].

First of all, we believe that the development of electronic textbooks should start from digital publishing and then technology of teaching and learning activities, such as notes, priorities, study records, etc. For functions of the rapid response learning [23] and Internet, we consider the Mobile Internet Device (MID). We emphasize functions of management and financial from the view points of technical systems. For teaching, we suggest function of main control, synchronization and group functions. It enables the teacher's group students in MID devices to display teaching materials through the master control and synchronization.

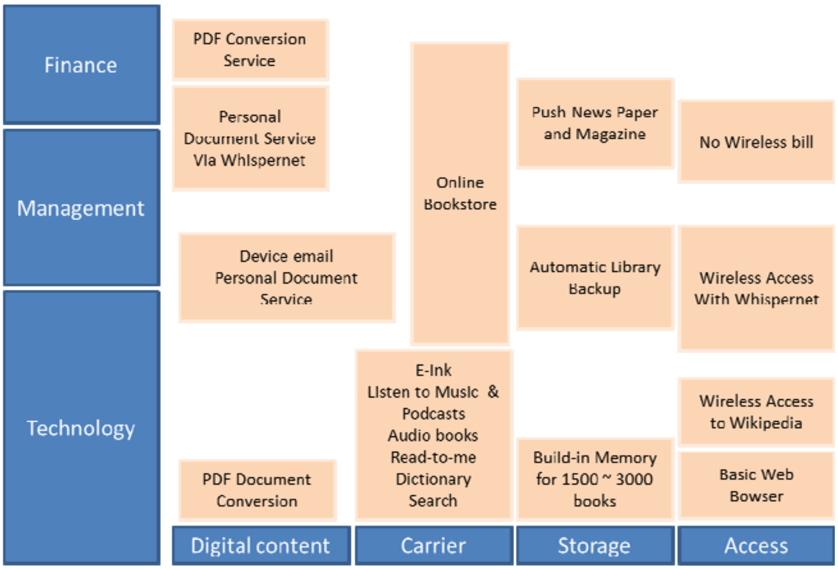


Figure 3. Technology System analysis of Amazon Kindle System function

Thus, it should provide function of cloud bookshelf; data was kept in the network and synchronized to the programmed group terminals. By the way, function of courses management is needed. For financial level, in addition to Amazon's online sales management, Push Service, the discount of purchase of electronic textbooks and teaching materials on electronic textbooks is also very important. Reading milestone certification, the cloud stacks of authorizing, printing on Demand (POD) and the authorization certificate-- students can legally print chapters, photocopying manufacturers and textbook publishers can share the profits. Therefore, the need for sales on the chapter, POD printing the profit sharing mechanism and bonus points for the readers of financial functions. Figure 4 shows the viewpoints of technology system of textbook business structure.

VI. CONCLUSIONS AND FUTURE SUGGESTIONS

This study is a government e-learning development plan application under the e-books reader technology and the electronic learning plan of the bookcase. Through the concept of technology system, we can have better insights for the holistic perspective. Not only the original e-books reader attached to a particular technology R&D team, but thinking beyond the technology development and broader way of regarding as a single point. Therefore, we can complete construction of those concepts in a holistic way.

Under the holistic conceptualization point of view, we think the future operations of the mechanisms of electronic textbook. From the technology needed to start learning technology, and gradually incorporated into management capabilities and financial technology capabilities, we develop Figure 4, the electronic textbook of rich business framework. After successfully complete the project. We try to propose a framework of electronic textbooks from technology system concept and rendering systems and applications of bookcase.

The impacts of this framework are two-fold. One is that there is always a first-mover advantage for this framework. We mean the first players involved in this framework will benefit as well as establish a firm basis. The other is the negative impacts of e-Textbook will be the issues worth further studies. From the preliminary experiences we met, there are quite different roles and functions in the classroom owing to the engagement of e-Textbooks.

Thus the original innovation-- brainstorming, while in the agitation, has been thinking of positioning the direction of rich/ innovative content and speed up the process of innovation. The concept of systems analysis, help us to understand the proposition Law-- a successful R&D is not simply a matter of developing technology, but uses a broader and comprehensive perspective of thinking and planning, to ensure the success of technological innovation.

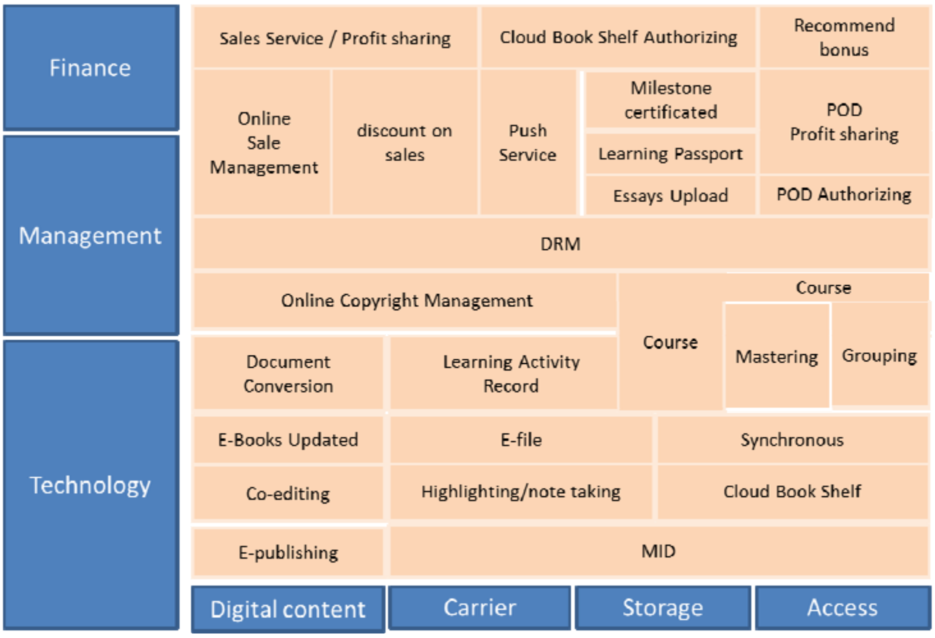


Figure 4. The framework of electronic books from technology system perspective

ACKNOWLEDGMENT

This study is conducted under the Project Digital Convergence Service Open Platform of the Institute for Information Industry which is subsidized by the Ministry of Economy Affairs of the Republic of China.

REFERENCES

[1] Kindle: Amazon's Original Wireless Reading Device (1st generation) http://en.wikipedia.org/wiki/Amazon_Kindle, Last accessed 2010/12/10

Amazon Kindle. http://www.amazon.com/dp/B000FI73MA, Last accessed 2010/12/10

[2] E-Book Semiconductor Opportunity to Reach $1.1 Billion in 2013, Says In-Stat. http://www.instat.com/newmk.asp?ID=2536,. Last accessed 2010/12/10

[3] Textbooks will be the next frontier for e-reader sellers http://www.startribune.com/business/70673177.html, Last accessed 2010/12/10

[4] Ryan, B. 1991. Dynabook revisited with Alan Kay. Byte, pp. 16.

[5] Goldberg, A. 1979. Educational uses of a dynabook. Computers & Education, 3: pp. 247-266

[6] Noorhidawati, A. 2008. A Study into Usability of Tools for Searching and Browsing E-books with Particular Reference to Back-of-the-Book Index. pp.15-28

[7] Borchers, J.O. 1999. Electronic books: Definition, genres, interaction design patterns. In ACM CHI 1999, Electronic Book Workshop. . Pittsburgh, PA.: Citeseer.

[8] Lemken, B. 1999. Ebook: the missing link between paper and screen. ACM CHI 1999, Electronic Book Workshop. Pittsburgh, PA.: Citeseer

[9] Henke, H. 2001. Electronic books and ePublishing: a practical guide for authors: Springer Verlag. pp.19

[10] Chien, C. "A Research Strategy and Framework for Prospected E-book Industry in Taiwan." The 2009 International Joint Conferences on e-CASE and e-Technology, 2009, Singapore

[11] Quint, B. 2009. "Google Book Search Expanding Outlets in Deal With Sony."." Information Today NewsBreaks & the Weekly News Digest 30.

[12] Bush, V. "As We May Think!", Atlantic Monthly, Vol.176 No.1, 1945, pp.101-108.

[13] http://www.gutenberg.org/wiki/main_page Last accessed 2010/06/15.

[14] Kay, Alan and Adele Goldberg 1977. Personal Dynamic Media. Computer, 10 (3), 31 – 41.

[15] Hart, Michael S. (23 October 2004). "Gutenberg Mission Statement by Michael Hart". Project Gutenberg. Retrieved 30 October 2010.

[16] http://ota.ox.ac.uk/ Last accessed 2010/06/15.

[17] Egan, D., Remde, J., Gomez, L., Landauer, T., Eberhardt, J., and Lochbaum, C. 1989. Formative design evaluation of superbook. ACM Transactions on Information Systems (TOIS), 7: 57.

[18] http://www.planetebook.com/mainpage.asp?webpageid=15&TBToolID=1115 Last accessed 2010/06/15.

[19] Abdullah, N. and F. Gibb 2008. Students' attitudes towards e-books in a Scottish higher education institute- part 1. Library Review, 57: pp. 593-605.

[20] Tonkin, E. 2010. eBooks: Tipping or Vanishing Point? http://www.ariadne.ac.uk/issue62/tonkin/ Last accessed 2010/06/15.

[21] Hughes, T. 1979. The electrification of America: the system builders. *Technology and Culture*, 20: pp. 124-161.

[22] Law, J. 1989. Technology and heterogeneous engineering: the case of Portuguese expansion. In: Bijker, W., Hughes, T., & Pinch, T. (Eds) The Social Construction of Technological System. MIT Press, pp. 112 – 118.

[23] McFall, R. 2005. Electronic textbooks that transform how textbooks are used. *The Electronic Library*, 23: pp. 72-81.