PROCEEDINGS

4th International Conference on INFORMATION TECHNOLOGY in Asia

CITA'05

Pervasive and Ubiquitous Computing: Computing Anytime, Anywhere for Everyone

12-15 December 2005
Hilton Hotel, Kuching, Sarawak, Malaysia
http://www.cita05.org

Organised by:

UNIVERSITI MALAYSIA SARAWAK
UNIMAS

In Collaboration with:

Information & Communications Technology Unit (ICT)
Chief Minister's Department of Sarawak

GITI
Global Information & Telecommunication Institute

Universiti Malaysia Sarawak

FOURTH INTERNATIONAL CONFERENCE ON INFORMATION TECHNOLOGY IN ASIA 2005

PROCEEDINGS OF CITA'05

Edited by

Narayanan Kulathuramaiyer
Alvin W. Yeo
Wang Yin Chai
Tan Chong Eng

Organised by

Faculty Computer Science & Information Technology, Universiti Malaysia Sarawak
Information & Communications Technology (ICT) Unit, Chief Minister's Office
Global Information & Telecommunication Institute (GITI)

Kuching, Sarawak, Malaysia

December 12-15, 2005

Editorial Preface

These are the Proceedings of the 4th International Conference on Information Technology in Asia (CITA'05), held between 12th -15th December 2005 in Kuching Malaysia. The CITA'05 conference is organised by Universiti Malaysia Sarawak in collaboration with the ICT Unit, Chief Minister's Department of Sarawak and the Global Information and Telecommunication Institute, Japan.

It is the aim of the conference to showcase applications of Information and Communications Technology (ICT) in the Asian region while fostering the exchange of ideas and research results with researchers worldwide. This focus on Ubiquitous and Pervasive Computing, is in line with our efforts to highlight the emerging trends and technologies that will help both organisations and nations to stay ahead in a dynamic and challenging environment.

We are pleased to have received a good response of 156 submissions from 17 countries. We have selected 60 papers under 8 major tracks: Community Informatics, Computing Infrastructure, Knowledge Networks and Management, Information Management, Systems Engineering, Human Computer Interaction, Computational Models and Systems and Autonomous Systems. The range of topics cover the pervasive nature of ICT applied in all aspects of our lives.

It is hoped that this conference will provide a platform to bring together researchers and practitioners to share their knowledge and experiences in preparing us towards the current trends and future issues in computing.

We would like to acknowledge and thank the many people who have contributed greatly to the conference. I wish to thank the members of the programme committee for reviewing the papers against a tight schedule, and the members of the organising committee for their hard work put in. We extend our sincere appreciation to all sponsors for their generous contributions.

We wish you all an enjoyable conference with fruitful deliberations and bid a warm "Selamat Datang" to our visitors to the Land of the Hornbills.

Associate Professor Narayanan Kulathuramaiyer
Organising Chair

International Review Committee Members

Organising Committee

Advisors
Abdul Rashid Abdullah
Khairuddin Ab Hamid

Chairman
Narayanan Kulathuramaiyer

Programme Co-chairs
Alvin W. Yeo
Azman Bujang Masli
William Patrick Nyigor

Secretary
Jane Labadin
Nadianatra Musa (Assistant)

State Liaison
Amilda Abu Bakar

Programme Committee
Wang Yin Chai (Chairperson)
Rosziati Ibrahim
Sharin Hazlin Huspi
Suriati Khartini Jali

Organising Committee

Publication Committee
Inson Din (Chairperson)
Eaqerzilla Phang
Yanti Rosminue Bujang

Technical & Logistic Committee
Halikul Lenando (Chairperson)
Sapice Jamel
Johari Abdullah
Mohamad Nazim b Jambli
Abdul Rahman b. Mat
Mohamad Nazri b Khairuddin
Noralifah bt Annuar
Mohd Imran b Bandan
Zulkifli b Ahmat
Selina bt Jawawi
Rosita Jubang
Sarina bt Ahmad
Razeki b Jelihi
Nurul Hartini bt Minhat
Roziah bt Majlan
Nur Rabizah bt Adeni

Protocol Committee
Dayang Hanani Abang Ibrahim (Chairperson)
Seleviawati Tarmizi
Abdul Rahman Mat
Mohd Johan Ahmad Khiri
Sukiah Marais
Doris Francis Harris
Zuraini Ramli
Norzian Mohammad
Piana Tapa
Mazlini Jemali
Muhsin Apong

Sponsorship Committee
Azni Haslizan Ab. Halim (Chairperson)
Fatihah Ramli

Exhibition Committee
Chai Soo See (Chairperson)
Vanessa Wee Bui Lin
Lau Sei Ping
Nurfauza Jali
Sarah Flora Samson Juan
Persatuan Teknologi Maklumat
(PERTEKMA)

Publicity Committee
Azman Bujang Masli (Chairperson)
Kartinah Zen
Azni Haslizan Ab. Halim
Roziah Majlan
Nur Rabizah bt Adeni

Workshop Committee
Bong Chih How (Chairperson)
Tan Chong Eng
Cheah Wai Shiang
Ling Yeong Tyng

Finance Committee
Hadijah Morni (Chairperson)
Hamisah Ahmad
Selina Jawawi

Multimedia Presentation & Website Design Committee
Jonathan Sidi (Chairperson)
Syahrul Nizam Junaini
Noor Alamshah Bolhassan
Muhammad Asyraf Khairuddin

Webmaster
Amelia Jati Robert Jupit
Nurfauza Jali

Table of Contents

Knowledge Networks and Management

Human Computer Interaction

Ubiquitous Information Management

Ubiquitous Systems Engineering

Computational Model & Systems

Agents and Autonomous Systems

Online Profiling and Privacy: An Exploratory Study Amongst Australian Websites.

Suku Sinnappan
Graduate School of Business,
Faculty of Commerce,
University of Wollongong,
Australia
suku@uow.edu.au

Punidha B. Sukunesan
Institute of Sathya Sai
Education,
Canberra, Australia
rpbanu@rocketmail.com

Rishika Sethi*, Gina Cetina Silva
Sydney Business School,
Faculty of Commerce,
University of Wollongong,
Australia
rs847@uow.edu.au*,
gacs71@uow.edu.au

Abstract -*The issue of online profiling and privacy has received growing attention from academics and practitioners. The unprecedented race among e-businesses to provide personalised contents has placed distrust in consumers at an alarming rate. This study combines and applies well established constructs developed by [8, 9, 11] to empirically assess the existing scenario of Australian Website's with regards to online profiling, privacy and trust. The study further highlights relevant managerial implications for managing these constructs and directions for future research are discussed, apart from suggesting an alternative approach towards online profiling i.e. self-profiling.*

Keywords: Online profiling, privacy, trust, websites, self-profiling.

1 Introduction

The advances in technology have revolutionized how e-businesses operate and gain competitive edge. Personalisation of products and services are seen as central to their customer relationship management (CRM) strategy. In order to achieve this, e-businesses have resorted to extensive collection, sharing and exchange of customer information via online profiling [1, 2, 3]. Due to the nature of the web customers leave traces of their visits, which in accumulation could be developed into a credible online profile. E-business then capitalizes on this profile to learn and formulate appropriate competitive marketing strategies. Even though this is seen as favourable amongst the retailer, it has created a very fretful scenario amongst the customers. Customers are not willing to give up personal information, and thus hardly commit to e-commerce transactions.

The data collected during e-commerce transaction is known to yield a deep and comprehensive profile of an individual's habits, preferences, finances, and ideology. Thus, privacy concerns compromise the overarching ecology of cyberspace, such that worry about privacy specifically the unmitigated profiling of individuals limits the Internet's potential to improve CRM customer's relationship management and invades customer privacy [1, 4, 5, 6]. The Centre for Business Ethics & Social Responsibility (BESR) states that, "Internet users are afraid that websites know too much about them" [7, p.1]. There has been a plethora of studies echoing the same [8, 3, 9, 10, 11s]. However, to our best knowledge there has not been any study to date examining the act of online profiling within the Australian context. Hence, the study applies well-established scales of [8, 10, 11] to highlight and discuss the salient issues concerning privacy, trust and the practice of online profiling amongst Australian websites, as well as to contrast the findings in similar research efforts by [11] and others. This study also proposes a new approach towards online profiling i.e. self-profiling [see 12] with the findings presented in three key sections. First, a review of the relevant literature within the domain of privacy and trust is presented. In the second section, details of the methodology and results of the analysis are highlighted. Thirdly, conclusions and managerial implications are discussed with directions for future research suggested.

2 Literature Review

2.1 Customer Relationship Management and Personalisation

As the development of Internet technology continues coupled with the flux of online competition where a click of a mouse is enough for an online customer to select a new provider [13], the height of e-Customer Relationship Management has forced e-businesses to develop rigorous and reliable methods to attract, retain and develop high-value customers. In some cases, this could constitute the recognition of customer profiles and dynamic presentation of products, prices or packaging deemed appropriate to that profile [14]. Website administrators and e-marketers must now continuously innovate and "engineer" branded customer experiences of their online service offering that both differentiates from competitors and strengthens customer relationships [15, 16]. One common and efficient approach is online personalisation, where contents of a website are tailored to suit the customer. This approach has been known to improve customer

retention and has been well documented within the online context. However, in order to perform this, e-businesses need information from the customers [17]. Hence, the beginning of online profiling.

2.2 Online profiling

Online profiling is the practice of aggregating information about consumers' lifestyle, product preferences and purchase history [18], gathered by tracking their movements online. The Federal Trade Commission contends that there are 3 categories of information that could be collected during online profiling i.e. anonymous information; personally unidentifiable information and personally identifiable information [see 9]. Using the information in the resulting profile, advertisers can tailor their marketing specifically to the consumer when she visits a website [2]. Further, [3] notes that online profiling could also be used for personalising web contents and products in order to enhance customer satisfaction and develop loyalty. It could help e-businesses to be selective in their efforts to discriminate profitable customers from the regulars. Online profiling can happen in several ways [see 3], most commonly when the host computer (the Website being visited) places a cookie on the end user's computer. The cookie transmits information back to the host computer. This information allows a business to track an end user's page views, the length and time of the visit and responses to advertisements. Purchases and search terms entered by the end user can also be tracked [19]. Using personalization software fed by web server logs analysis, advertisers can dynamically change content based on profile information. By combining demographic and psychographic behaviour data of other end users with analysis of a consumer's click stream data, personalization software can anticipate the consumer's actions. For instance, in the course of a consumer's visit to a website, it may appear that he may abandon a shopping cart because of the high shipping cost of a product. If other customers have demonstrated a similar tendency toward rejecting the purchase of the same product for the same reason, the Website may offer – in real time an incentive to entice other customers to buy the item [20].

The activities of profiling networks are the leading edge of a growing industry built upon the widespread tracking and monitoring of individuals' online behaviour [21]. Via profiling e-businesses see mass customisation. For a long time mass customisation has been a marketer's dream: the fantasy that machines could be built that automatically tailors products to the customer's requirements. That technology is now here. One limitation was capturing the customer's specification economically but now that the Internet is enabling customers to deliver their specifications in a structured form, everything from books to houses can be tailored online. [2] pointed that almost all e-commerce companies do online profiling and increasingly pervasive use of surreptitious monitoring systems compromises consumer trust in the Internet and undermines their efforts to protect their privacy by depriving them of control over their personal information. The customer's information is collected silently [17] and women are the prime targets due to their spending behaviour [22]. Albeit, the benefits e-businesses reap, online profiling raises consumer's distrust and is perceived as an intrusion to their privacy.

2.3 Privacy and trust

IS and marketing literature have confirmed consumer concerns over information privacy as one of the most highlighted issues in current information age [23]. The right of an individual to be anonymous while having control over personal information could define privacy [24]. Given the discussion in section 2.2 there is an increasing awareness of privacy and concern arising from the public opinion these days. One of the major concerns faced by the online customers are while submitting personal information as to what the information will be used for by an online company and whether will it be disclosed to any other third parties [9]. Privacy legislation has done much to reduce the risk from threaten data to ensure costumers' confidentiality and to reinforce customer-centric approach in E-commerce to make profiling an effective tool. Specifically, Australian online privacy legislation started with The Federal Privacy Act, which contains eleven Information Privacy Principles (IPPs), which apply to Commonwealth and ACT government agencies. It also has ten National Privacy Principles (NPPs) which apply to parts of the private sector and all health service providers The Federal Privacy Commissioner also has some regulatory functions under other enactments [see 25]. The 10 NPP's: Collection, use and disclosure, data quality, data security, openness, access and correction, identifiers, anonymity, transborder data flows, sensitive information. In terms of Internet privacy is important to recall these 10 principles to create a trusted environment for customers and business. Concerns about privacy rank high among Australians who say they are more concerned about the loss of personal privacy (56%) than health care (54%), crime (53%), and taxes (52%). When asked specifically about their online privacy, survey respondents said they were most worried about websites providing their personal information to others without their knowledge (64%) and websites collecting information about them without their knowledge [26]. [27] has estimated that 45% of consumers who do not currently make online purchases would do so if their privacy concerns were addressed, and 52% of current buyers would make more purchases with privacy protections in place. [28] reported that 30% of customers claimed that they had been stolen or cheated online, while 80% of customers said that if the companies had good security for their accounts, they would have been more confident to buy online. This lack of consumer trust can be directly translated into losses. [22, 24, 39] contend that trust has a major influence on information

disclosure. This is evident from the *Internet Consumer Trust Model* developed by [24], the *Electronic Exchange Model* by [29], the *Model of online information disclosure* [11], which sources trust to be a key antecedent to engaging in consumer transactions online. This is mainly is because trust reduces the perception of risks associated with purchasing goods and services online. Higher the trust, the better the chance of consumers engaging in electronic transactions.

3 METHODOLOGY

3.1 Sample Frame

There are numerous ways to evaluate a Website [see 30] and to gauge the components of trust [see 10]. Accordingly, we adopted a suitable methodology inline to our objectives. To gather a comfortable target population, our study required respondents with previous experience in the e-commerce area. The respondents for the survey were post-graduate students enrolled in Information Systems subjects at a large Australian university. Since the topic of Website privacy, trust and online profiling was part of their syllabus, they were considered to be an ideal target population. Prior to the actual study, the students (in groups) were asked to browse and select numerous sites and get familiarised with online profiling practices amid Australian websites during their tutorials. This process allowed respondents to trial the Websites before its subsequent evaluation. The decision to choose which Websites for the study, were derived as a convenient random selection of websites according to alphabetical order generated during the tutorials. Each group was then asked to nominate a website in alphabetical order and no websites should be nominated twice. This lead to the selection of 54 websites (Table 1) for the study, which could be attributed to the Australian scenario as it had almost all industry representative, including government and charity-based.

3.2 Survey Design and Administration

The survey for this study was conducted over two phases. In the first phase, each respondent was briefed on the instructions prior to handing out the questionnaire and were allotted 2 random websites to complete the survey. The survey gathered 108 valid responses. The survey comprised 6 key sections made up by constructs from [8, 9, 10, 11, 31] as shown in Table 2. Respondents were invited to indicate their opinion about disclosing personal information, issues on trust and perception towards Australian websites using a 7 - point scale (as in similar studies). The questions were anchored as 1 = "Very strongly disagree" and 7 = "Very strongly Agree". Further, there was a section on online privacy practices where the respondents were asked to either answer yes, no or unsure. The second phase of the survey was concerned on the issues of online profiling and was conducted immediately after the first phase hence to ensure that there was no change in the conditions of the survey. Respondents were again invited to use a 7-pont scale to indicate their views on online profiling and self-profiling.

4 DATA ANALYSIS

4.1 Survey Data in Brief

The survey saw 108 respondents whose profile reflected the typical e-commerce participants [32]. Briefly, they were mature-aged students aged between 24 and 30 years with an average of 3 years working experience, and Internet literate. Sixty percent of them were females, 94% were single with middle-income range.

Table 1: Depicting the websites used as part of the survey

Websites used as part of the study in alphabetical order
www.abc.net.au, www.acrod.org.au, www.anz.com.au, www.auspost.com.au, www.australiandeference.com.au, www.austar.com.au, www.chinaren.com.au, www.citibank.com.au, www.crosscountryrealty.com.au, www.dell.com.au, www.dlink.com.au, www.ebay.com.au, www.ernstyoung.com.au, www.getaway.com.au, www.gio.com.au, www.google.com.au , www.iinet.com.au, www.imb.com.au, www.immi.gov.au, www.jetairways.com.au, www.jobcareer.com.au, www.jobnet.com.au, www.kpmg.com.au, www.microsoft.com.au, www.msn.com.au, www.national.com.au, www.nestle.com.au, www.ninemsn.com.au, www.nissan.com.au, www.nokia.com.au, www.optus.com.au, www.orange.com.au, www.outlook.com.au, www.oz.com.au, www.pe-express.com.au, www.pwc.com/au, www.ralph.ninemsn.com.au, www.rebelsport.com.au, www.roadrunnerrecords.com.au, www.sharp.com.au, www.smartravel.com.au, www.sony.com.au, www.supportersworld.com.au, www.telstra.com.au, www.three.com.au, www.tnt.com.au, www.vicroads.vic.gov.au, www.visitmelbourne.com, www.vlinepassenger.com.au, www.wesfarmers.com.au, www.westpac.com.au, www.woolworths.com.au, www.yahoo.com.au, www.zp.com.au

Table 2: Depicting the adopted constructs from [8, 9, 10, 11, 31] used part of the survey.

No	Available Constructs from Literature Review	Adopted Key Constructs
1	Reputation [8]	Reputation and perception
2	Perceived size [8]	
3	Website trustworthiness [8]	Trustworthy
4	Attitudes towards a website [8]	
5	Willingness to visit [8]	
6	Web-shopping Risk Attitudes [8]	Risky
7	Information privacy [11]	Information privacy
8	Web and other characteristics [10]	Privacy Upheld, Online profiling and Self profiling
9	Online personalisation [9]	
10	Self-concept [31]	

The respondents' opinion about disclosure of information was given the prime importance to gauge privacy. As expected on the demographic front more than 50% of respondents (out of 108) on average were willing

to share information about their age, gender, level of education, ethnicity, nationality, religion, hobbies and interest. However, only a handful was willing to reveal their first name (26.9%) or last name (30.5%), or even martial status and spouse details (6%). About 65% of the respondents strongly opposed of sharing any information relating to income. Further, 77% of the respondents declined to share their contact details or postal information. Other high-risk information was obviously too personal to be shared i.e. credit details (87%), driving license (90%) and bank details (93%).

The survey comprised other questions relating to behavioural analysis and purchase intentions. It has been known that behavioural questions are a better predictive tool than demographics information. Hence, the inclination of respondents to share their Internet usage was queried. About 36% of the respondents were happy to share the list of websites they browsed either at work or at home. More than 40% declined to disclose the amount of time spend online either at work or home. Approximately 29% were willing to share the text exchange they had at work, as contrast to only 10% at home indicating a different level of privacy. Further, 67% strongly declined to share any chat related information exchanged at work, while only 50% declined at home.

Analysing the purchase intentions, we found that 42% of the respondents were happy to share their brand preferences. However, a mere 14% would reveal their budget or recent purchase details.

There were several constructs that were lined up to measure the 54 websites' privacy policies. About 76% of the websites were said to have clear privacy policies or act, but 52% of them were confusing and incomprehensible when got to the details. Less than half (44%) allowed their customers to access the profiled information and 31.5% or 17 websites openly confessed that customer information could be sold or shared with the public, which was found in [26]. Further, 61% of the websites (firms) were deemed to have bad reputation and 65% of them are well known among the public. One in two websites were found to share their customers' information with undisclosed third party or partner companies, and a similar ratio of websites were found to be trustworthy. Approximately, 40% of the websites were treated with cautions and met respondents' expectations surprisingly given the above shortfalls. On average, 40% of the respondents indicated that they would revisit the website, but a mere 16.7% would exchange any information or purchase anything at present moment. Respondents' web-shopping attitudes were also analysed as part of the survey. A credible number of respondents (37%) felt safe completing commercial transactions over the Internet and same number indicated that there is too much uncertainty with shopping over the Internet. Close to 43% admittedly revealed that buying over the Internet is risky.

Table 3 and 4 depicts the average score and rank for the 2 constructs (*privacy upheld* and *trustworthy*) on 6 websites. Interestingly, websites from the consultancy industry (*www.ey.com.au, www.kpmg.com.au, www.pwcglobal.com.au*) topped both the constructs. Tnt.com.au was found to up keep customer privacy the most [33], continued by the consulting group. Tnt's website was extremely commendable as it does not share any of its customer's information. *Google.com.au* (6.22) topped the list of the trustworthiest continued by *Ebay.com.au* at 6.11. The commanding market share and trust customers have in these websites are reflected in the score, in addition to their conscientious privacy policy, which carries text depicting the importance of privacy and how they thoughtfully manage it [34, 35]. While, *Vlinepassenger.com.au* was utterly disappointing on both constructs. It could be largely attributed to the non-responsible attitude shown via their privacy policy [36]. Respondents also poorly rated *Getaway.com.au* (1.73), *Australiandefence.com.au* (1.73), *Dlink.com.au* (1.73), *Dell.com.au* (1.73) on the privacy construct and so were *Chinaren.com.au* (2.78), *Oz.com.au* (2.89), *Roadrunnerrecords.com* (3.44), *Ninemsn.com.au* (3.44) on the trustworthy scale.

Table 3: Top 6 websites that appear in both of the constructs – Privacy Upheld and Trustworthy

Privacy Upheld 1=yes, 2=no	Most Trustworthy Scale 1 up to7
www tnt.com.au (1.09) www.ernstyoung com.au (1.18) www.kpmg com au (1.18) www pwc.com/au (1.18) www national com au (1.18) www nissan com.au (1.18)	www google.com.au (6.22) www.ebay com.au (6.11) www citibank.com.au (6.11) www ernstyoung com au (6.00) www.kpmg.com.au (6.00) www.pwc com/au (6.00)

Table 4: Bottom 6 websites that appear in both of the constructs – Privacy Upheld and Trustworthy

Worst Privacy Upheld 1 yes, 2=no	Least Trustworthy Scale 1 up to7
www vlinepassenger.com.au (1.91) www.visitmelbourne.com (1.91) www.getaway.com au (1.73) www.autraliandefence.com.au (1.73) www.dlink com au (1.73) www dell com.au (1.73)	www.vlinepassenger.com.au (2.00) www.visitmelbourne com (2.00) www.chinaren.com.au (2.78) www oz.com.au (2.89) www.roadrunnerrecords com (3.44) www.ninemsn.com.au (3.44)

In summary, respondents were comfortable with websites bearing convincing-responsible approach towards privacy as the below (e.g. PricewaterhouseeCoopers), which is used by almost all consulting firms i.e. (*www.ey.com.au, www.kpmg.com.au, www.pwcglobal.com.au, www.accenture.com.au*) *"...PricewaterhouseCoopers believes that privacy is an important individual right and is important to our own business and the businesses of our clients and customers. In this policy we set out the standards to which the Australian Firm is committed in ensuring the privacy of individuals. We are also bound to comply with the National Privacy Principles as set out in the Privacy Act 1988."*

Further questions regarding online profiling revealed other interesting findings, in which most supported the act online profiling and collection of customer information. A quarter of the respondents approved the practice of online profiling, and almost one in two respondents felt that online profiling will improve the quality of website.

However, only a third could attribute their 'commendable' web experience due to online profiling. Almost 36% of them saw online profiling as invading their privacy and only 20% knew how the profiled information is being used. While an alarming rate of 75% of the respondents do not have the habit of reading and understanding the adopted privacy policy at any given website. Interestingly, close to 39% or 42 respondents preferred self-profiling over online profiling to have more control over their own information. Respondents were keen to nominate professional occupancy, product or brand as a proxy for self-profiling [see 12]. Almost 37% of the respondents firmly indicated that e-business should pay an average amount of AUD$1000 before being profiled.

To further analyse the data, a reliability analysis together with an exploratory factor analysis (EFA) was conducted. A reliability analysis is conducted to ensure that the research instrument has high reliability where each construct will be measured using Cronbach alpha [37]. Expectedly, all constructs were reported to have scores above 0.7 and this confirmed the stability and consistency of responses to related items measured where an alpha value above 0.60 is acceptable [38]. An EFA is known for identifying the core structure (variables) that is latent by summarising the data set [39]. We conducted an EFA with Varimax rotation (Eigen value more than 1.0); with a factor loading excess of more than 0.55 was retained for significant results [39]. The results of the factor analysis in Table 5 revealed the latent dimensions and variables that were core to our study. The results suggest that there could 7 possible core dimensions (22 variables), which could explain 60.5% of the collected data and this needs to be verified and further explored using confirmatory factor analysis. More importantly, similar variables were highlighted in [8], which confirms the findings of the study.

find out what information is held about them, and if so how can they get this access?), a statement about what privacy laws the company complies with, what privacy seal programs they participate in, and other mechanisms available to customers for resolving privacy disputes.

Table 5: Core dimensions and variables for Constructs - Reputation and perception, Trustworthy and Risky

V	Dimensions and scores explaining 60.5% of the data						
	1(10.2)	2(9.6)	3(9.5)	4(9)	5(7.6)	6(7.34)	7(7.25)
Big	0.878						
Vlarge	0.835						
Known	0.796						
BSupp	0.766						
Return12		0.937					
Return3		0.875					
LikelyR		0.744					
NowV		0.722					
Perso			0.814				
SafeTX			0.783				
Tworthy			0.695				
Sophi			0.634				
Ishop				0.898			
Ashop				0.873			
Ushop				0.865			
Cautious					0.836		
GoodR					0.776		
Bint					0.696		
Promise						0.830	
Lose						0.560	
Risky						-0.539	
SaveTX							0.964

This statement may also describe what remedies are offered should a privacy policy breach occur, a description of the kinds of data collected, including what kinds of data may be linked to cookies, a description of how collected data is used, and whether individuals can opt-in or opt-out of any of these uses, information about whether data may be shared with other companies, and if so, under what conditions and whether or not consumers can opt-in or opt-out of this, information about the site's data retention policy, if any, and, information about how consumers can take advantage of opt-in or opt-out opportunities [41]. It could be a possibility that a lot of Australian companies, which are not able to follow a privacy policy due to lack of resources, managerial or financial factors, continue to lose customers' trust and relationship. In the long run this would restrict and squeeze them out of the digital race. Surprisingly, customers acknowledge online profiling, despite their distrust over e-businesses that share their information to others purely to obtain better web services.

This study provides an initial research effort in the theory development and application of online profiling and privacy and as such suffers from two obvious limitations. Firstly, even though respondents of this study were made of university students who were comfortable and familiar with website evaluation. They did not represent the general Australian online population, which would have induced some degree of biasness to the data. Secondly, the numbers of websites used to represent the Australian scenario were rather insufficient. Although the websites used for the study were selected to represent the Australian scenario, more websites would have been more appropriate.

5 Conclusions, Managerial Implications and Limitations

The aim of the study was to apply well establish constructs used in [8, 10, 11] to explore issues relating to online profiling, privacy and trust amongst Australian websites. The results of the study indicate similar findings with [9, 26, 40], and report a high correlation between privacy and trust as reported by [8].

The study also reports that companies have to invest more effort in adopting appropriate privacy policies to respond to customers' demands in terms of privacy concerns. On average most websites need to re-organise their privacy policy to point key terms as: The name and contact information for the company or organization, a statement about the kind of access provided (can people

6 Future research directions

As a result of the exploratory nature of the study, a number of directions for future research arise from this paper. First and foremost, further refinement of the technique is required which should use larger randomised sample sizes made up of the general Internet population to avoid bias and to provide more statistical power. Beneficial investigation could be conducted concerning the concept of self-profiling [12] and its effects on personalisation of web, particularly as consumers are inclined to use professional occupancy, products or brands as a proxy. Future research should also explore the effects of privacy and trust alongside Website quality, customer satisfaction, perceived value of the Website to consumers and loyalty to the Website. Advanced statistical techniques such as structural equation modelling (SEM) could be further employed to investigate these inter-relationships between constructs. Furthermore, the role of Internet experience should also be examined for moderating effects on these prescribed relationships. More research is required across other various industry sectors within the Australian and international context to refine the constructs. Other avenues of research could include the efforts of World Wide Web Consortium to introduce P3P and APPEL to facilitate profiling practices [42].

References

[1] A.B. Albarran, Electronic commerce. In A. B. Albarran & D. H. Goff (Eds.), *Understanding the Web: Social, political, and economic dimensions of the Internet*, pp. 73-94, Ames: Iowa State University Press, 2000.

[2] A. Shen, Request for Participation and Comment from the Electronic Privacy Information Center (EPIC), 1999. *Online Profiling Project*, Available Online: http://www.cdt.org/privacy/FTC/ profiling/shen.pdf, Accessed on 3 March 2005

[3] K.P Wiedmann, H. Buxel and G. Walsh, Customer profiling in e-commerce: Methodological aspects and challenges, *Journal of Database Marketing*, Vol 9, pp. 170-184, 2002.

[4] D.L. Hoffman, T.P. Novak and M. Peralta, Building consumer trust online. *Communications of the ACM*, Vol 42, No. 4, pp. 80-85, 1999.

[5] B.K. Kayand N.J. Medoff, *Perspective*, Mountain View, CA: Mayfield, 2001.

[6] P. Hallman, Personalization vs. Privacy: An introduction to personalization and online privacy and their impacts on legislation and the online business, KTH-Royal Institute of Technology, June 2001. Available Online: http://www248.indek.kth.se/grundutbildning/grundkurser/4d1176/privacy.pdf, Accessed on 1 March 2005

[7] Centre for Business Ethics & Social Responsibility (BESR), December 2000, Values for Management: Consumer Privacy on the Internet, Available Online, http://www.besr.org/journal/besr_newsletter_2.html, Accessed on 5 March 2005

[8] S.L. Jarvenpaa and N. Tractinsky, Consumer trust in an Internet store: A cross-cultural validation. *Journal of Computer Mediated Communication*, Vol 5, No.2, pp. 1-36. Available http://www.ascusc.org/jcmc/vol5/issue2/jarvenpaa.html

[9] R.K. Chellappa and R. Sin, Personalization versus Privacy: An Empirical Examination of The Online Consumer's Dilemma, *Information Technology and Management*, Vol.6, No.2-3, 2005. Available Online: http://asura.usc.edu/~ram/rcf-papers/perprivitm.pdf#search='profiling%20privacy%20online%20personalization', Accessed on 13 March 2005

[10] Shankar, Venkatesh, Glen L. Urban and Fareena Sultan, "Online Trust: A Stakeholder Perspective, Concepts, Implications and Future Directions," *Journal of Strategic Information Systems*, Vol 11, No. 3-4, pp. 325-344, December 2002.

[10] Miriam J. Metzger, Privacy, Trust, and Disclosure: Exploring Barriers to Electronic Commerce, *Journal of Computer-Mediated Communications*, Vol 9, No. 4, 2004.

[11] IndusAge.com.au "*Corporates should get proactive in IT education*" http://www.indusage.com.au/article.asp?id=14 assessed 10th august 2005

[12] M. Singh, "E-Services and Their Impact on B2C E-commerce". *Managing Service Quality*, Vol 12, No. 6, pp. 434-446, 2002.

[13] Julian Bright, Prepaid CRM: Know thy customer. *Total Teleco*, pp 34-35, February 2005.

[14] L.P. Carbone and S.H Haeckel, "Engineering Customer Experiences. *Marketing Management*, Vol 3, No. 3, pp. 8-19, 1994.

[15] S. Dutta, and A. Segev, "Business Transformation on the Internet," in S. J. Barnes and B. Hunt (Eds.) *Electronic Commerce and Virtual Business*, Butterworth-Heinemann, Oxford, 2001.

[16] L. Hatlestad," Privacy matters", *Red Herring, No. 90*, pp. 48-56, January 16, 2001.

[17] Maxenews, *Profiling Your Customers And Using Your Customer Data*, February 2004. Available Online:

http://www.emailcenteruk.com/maxenews/feb04/profiling_customers/htm, Accessed on 11 March 2005

[18] Tamara Dinev and Paul Hart, Internet privacy concerns and their antecedents -measurement validity and a regression model. *Behaviour & Information Technology*, Vol 23, No. 6, November/December 2004. pp. 413-422

[19] D. Pierce, Public Workshop on Online Profiling Session II: Implications of online profiling technologies for consumer privacy Online Profiling Project - Comment, P994809 Docket No. 990811219-9219-01, October 18, 1999. Available Online: http://www.cdt.org/privacy/FTC/profiling/pierce2.htm, Accessed on 3 March 2005

[20] Kimberly C. Braun Claffy, Hans-Werner. Polyzos, C. George, "Parameterizable methodology for internet traffic flow profiling", *IEEE Journal on Selected Areas in Communications. V 13 n 8 Oct 1995. p 1481-1494*

[21] A. Bartow, Woman as Targets: The Gender-Based Implications of Online Consumer Profiling, 1999. Available Online: http://www.cdt.org/privacy/FTC/profiling/bartow.htm, Accessed on 3 March 2005

[22] A.D. Miyazaki and A. Fernandez, "Internet Privacy and Security: An Examination of Online Retailer Disclosures", *Journal of Public Policy & Marketing*, Vol 19, No.1, pp. 54-61, 2000.

[23] Clarke, Roger "Information Privacy On the Internet Cyberspace Invades Personal Space" May 2, 1998 http://www.anu.edu.au/people/Roger.Clarke DV/IPrivacy.html assessed on 19th March 2005

[24] Federal Privacy commissioner. Australia. " Federal privacy law." http://www.privacy.gov.au/copyright/index.html assessed on 20th March 2005.

[25] Commonwealth of Australia. "Australian Competition and Consumer Commission." 2005 http://www.accc.gov.au/content/index.phtml/itemId/142 assessed on 12th March 2005.

[26] Forrester Research, Inc. "Helping Business thrive on Technology change" 1997-2005 http://www.forrester.com/my/1,,1-0,FF.html assessed on 20th March 2005.

[27] L. Nicolle, E-Commerce: Boom or Bust, United Kingdom, 2005. Available Online: Proquest, Accessed on 11 March 2005

[28] V. Swaminathan, E. Lepkowska-White and B.P. Rao, "Browsers or buyers in cyberspace? An investigation of factors influencing electronic exchange" *Journal of Computer Mediated Communication*, Vol 5, No. 2, pp. 1-23. 1999. Available: http://www.ascusc.org/jcmc/vol5/issue2/swaminathan.htm

[29] D. Cunliffe, "Developing Usable Websites A Review And Model," Internet Research, Vol 10, pp. 295-308, 2000.

[30] M.J. Sirgy, "Self-concept in Consumer Behaviour: A Critical Review", *Journal of Consumer Research*, 9, 287-300, 1982.

[31] Privacy and American Business. "Personalized Marketing and Privacy on the Net: What Consumers Want" November 12, 1999. http://www.pewinternet.org/reports/pdfs/PIP Trust Privacy Report.pdf. assessed on 7th March 2005

[32] Tnt.com "Website Privacy Policy" http://www.tnt.com/country/en_au/cq privacy.Par.0001.file.tmp/TNT_Website_Privacy Policy.pdf assessed 10th August 2005.

[33] Google.com "Google Privacy Centre: Google Privacy Policy" http://www.google.com.au/privacy.html assessed on 10th August 2005.

[34] eBay.com "eBay privacy policy" http://pages.ebay.com.au/help/policies/privacy-policy.html?ssPageName f:f:AU assessed 10th August 2005.

[35] Vlinepassenger.com "Privacy Statement" http://www.vlinepassenger.com.au/privacy.html assessed 10th August 2005

[36] J. Cronbach, l, "Coefficient alpha and the internal structure tests", *Psychometrika*, Vol 16, pp. 297-334, 1951.

[37] J. Nunnally, *Psychometric Theory*, New York: McGraw Hill, 1978.

[38] F.J. Hair, E.R. Anderson, L.R. Tatham and C.W. Black, *Multivariate Data Analysis*, (5th ed.), New Jersey: Prentice Hall International, 1998.

[39] H.H. Teo, L.Wan, and W. Li, "Volunteering personal information on the Internet: Effects of reputation, privacy initiatives, and reward on online consumer behavior", *Proceedings of the Hawaii International Conference on System Sciences Proceedings of the 37th Annual Hawaii International Conference on System Sciences*. v 37 2004. p 2841-2850

[40] Naipeng Ding and A. Sidney, "Recognizing modern CRM in retail business", *Proceedings of the International Conference on Information and Knowledge Engineering, IKE'04 Proceedings of the International*

Conference on Information and Knowledge Engineering, IKE'04, pp. 451-460, 2004.

[41] http://www.W3C.org "Platform for Privacy Preferences (P3P) Project" http://www.w3.org/P3P/ assessed 10th August 2005.

A Framework for Healthcare Data Warehousing

Sellappan P.
Department of Information Technology
Malaysia University of Science & Technology
Kelana Square, 47301 Petaling Jaya, Malaysia
sell@must.edu.my

Yih Huey N.
Department of Information Technology
Malaysia University of Science & Technology
Kelana Square, 47301 Petaling Jaya, Malaysia
yhng@must.edu.my

Abstract - *Currently, information collected by the various healthcare providers (hospitals and clinics) in Malaysia is not integrated. Each collects information for its own internal use. So, the information is not aggregated and analyzed at the state or national level to provide useful information for decision making (e.g., by the Ministry of Health). Extracting and aggregating information from heterogeneous and distributed systems have always been a challenge for the healthcare industry. This paper proposes a framework for integrating healthcare information into a single data warehouse. It is simple, cost-effective and yet solves the information integration problem. It consists of a schema mapping process which is partially automated and an Extraction, Transformation and Loading (ETL) process which is fully automated. To demonstrate its viability, a data warehouse containing materialized data is built from several distributed and heterogeneous databases such as SQL Server and Oracle. It is implemented using the Microsoft .NET Framework.*

Keywords: Data Warehouse, Healthcare Information Integration, Extraction Transformation and Loading (ETL), Metamodel, Schema Mapping

1 Introduction

Currently, healthcare providers (i.e., hospitals and clinics) in Malaysia use conventional transaction processing systems to record their daily clinical and financial data. However, as peoples' lifestyles become more complex and technologies become more advanced, users demand more quality or valued-added services. Healthcare providers are expected to increase the value of their transaction processing systems - they need to turn data into actionable information [21].

Healthcare providers generate voluminous data, but they are not leveraged for generating valuable information. Extracting, aggregating and analyzing healthcare information from distributed and heterogeneous data sources (databases) have always been a challenge for the healthcare industry. Aggregating and integrating healthcare information is important for understanding the health of a community because quality services depend on the availability of accurate, relevant and timely information. This paper proposes a healthcare data warehousing that will efficiently integrate healthcare information from distributed and heterogeneous data sources (databases) into a single data warehouse. Such a data warehouse can be used to implement healthcare decision support systems that can enhance the quality of healthcare services provided. It can for example support OnLine Analytical Processing (OLAP) and data mining to generate more focused healthcare information [2].

Most healthcare providers do not use data warehouse technology because it is costly. The framework proposed in this paper solves the data integration problem. It is simple and requires less resources compared to the data integration technologies available in the market. Basically, it provides features to (a) semi-automate the schema mapping process and (b) fully-automate the Extraction, Transformation and Loading (ETL) process. These features allow the integration of healthcare information from heterogeneous and distributed databases.

2 Framework Architecture Design

The framework is designed based on the client-server architecture (Figure 1). It comprises of two main components: Agent and Web Server. The Agent is installed on each client computer. Data extracted from the clients are uploaded to the Web Server for integration via the Internet.

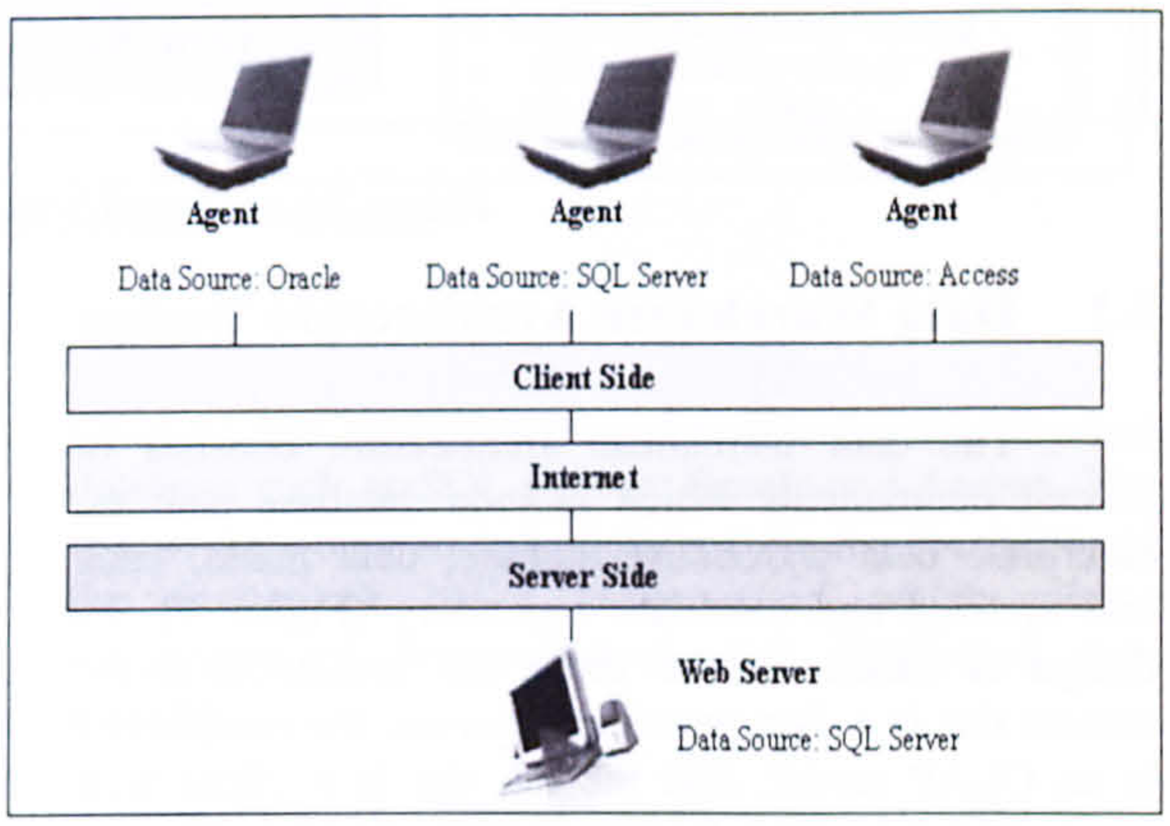

Figure 1 High Level View of the Framework Architecture

The Web Server consists of a data warehouse Manager and an Extraction, Transformation and Loading (ETL) engine (Figure 2). The Manager manages the data warehouse and the metadata. The ETL engine manages the data integration process. A Web application is

developed using the ASP.NET to facilitate the administrative tasks. It allows the stakeholders (e.g., healthcare providers, administrators, Malaysia Medical Association, Ministry of Health, researchers) to access the integrated healthcare information. The healthcare providers can upload their XML data files extracted automatically by the Agent via the Web application. (This can be extended in future to provide stakeholders with analysis functions to assist them in their decision making.)

The Agent performs the schema mapping and data extraction (Figure 2). The schema mapping is semi-automated; the engine first extracts the clients' database metadata and the data warehouse metadata. The user then performs the mapping manually using the client database structure and the data warehouse structure. The mapping generates a schema mapping metadata for data integration. The data extraction process on the other hand is fully automated. The ETL engine extracts the data from the client data repository using the schema mapping metadata and then converts them to XML data files. The files are uploaded to the Web server for data integration.

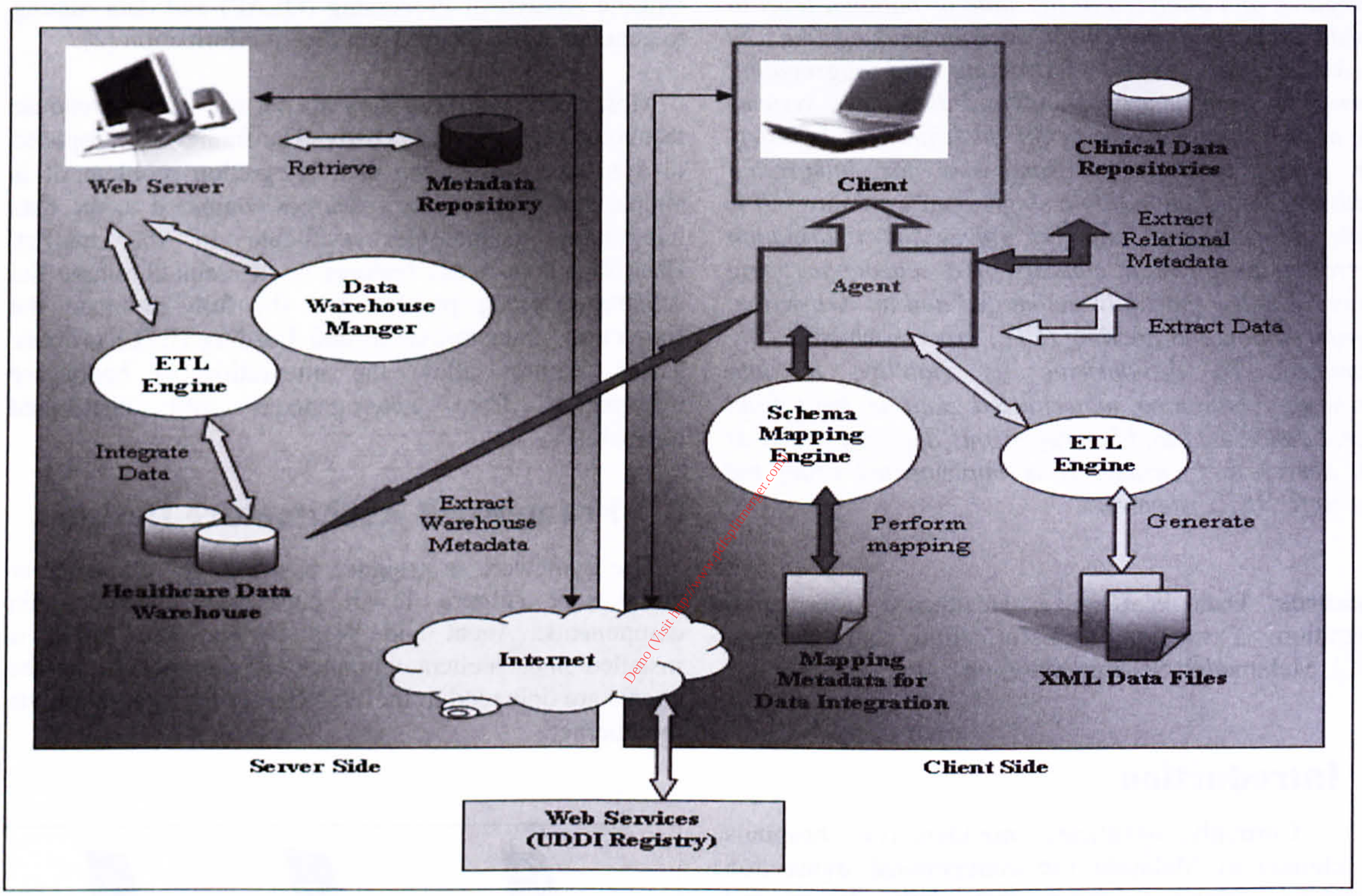

Figure 2 Low Level View of the Framework Architecture

2.1 Data Warehouse Architecture Design

The data warehouse architecture consists of several components which include database sources, enterprise data warehouse storage, data marts, back end systems and metadata repository (Figure 3). Its design is based on the three tier architecture: the bottom tier is a data warehouse server, the middle tier is an OLAP server, and the top tier is a client with query and reporting facilities. This research focuses only on the bottom tier. The middle and the top tiers are areas of future work.

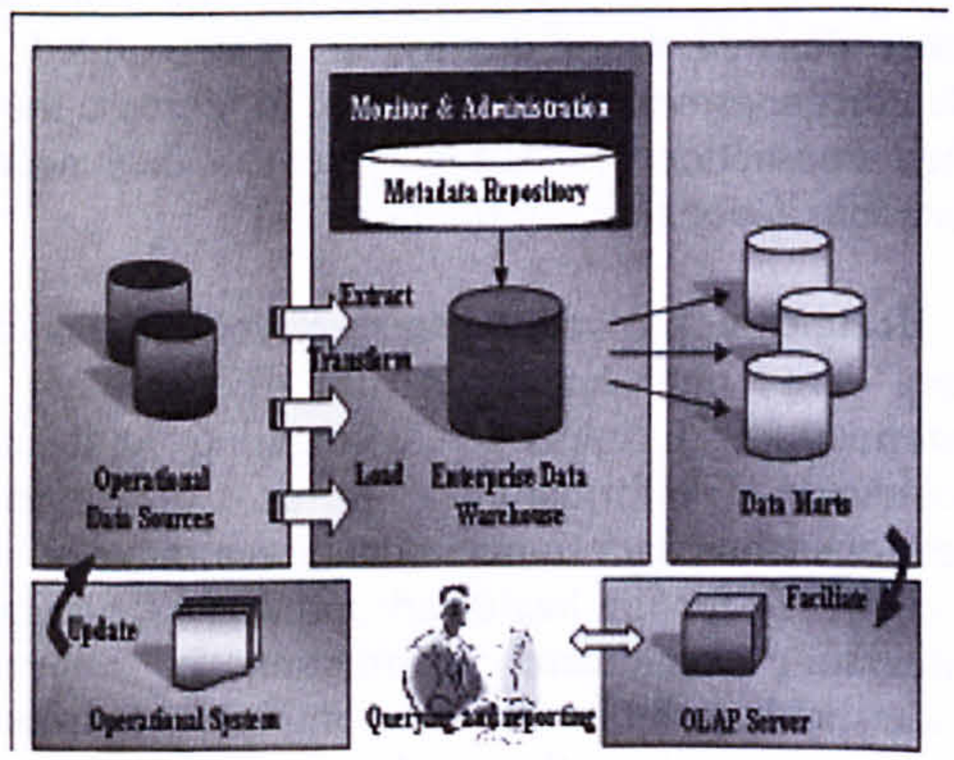

Figure 3 Data Warehouse Architecture

2.2 Conceptual Data Warehouse Schema Design

The conceptual data warehouse schema can be modeled using either dimensional model or tabular model. Both models have their own advantages. However, in this research the tabular model is used for modeling the data warehouse schema. This is because the tabular model enables both the details and summarized data to be captured which will provide higher values to the various stakeholders in the future as they can be optimized for numeric and textual analysis.

Although the tabular model is used to structure the data warehouse schema, the dimensional model can also be derived easily when the needs arise such as to support OLAP functions by extracting the respective tables and attributes. However, the existing tabular model can still support analysis function by using Relational OLAP (ROLAP) tools.

2.3 Data Warehouse Metamodel Design

The metamodel is an important element in the data warehouse environment as it represents the structures and semantics of the different data models within the framework [24]. It is used to integrate the heterogeneous data sources. Figure 4 shows the design of the metamodel.

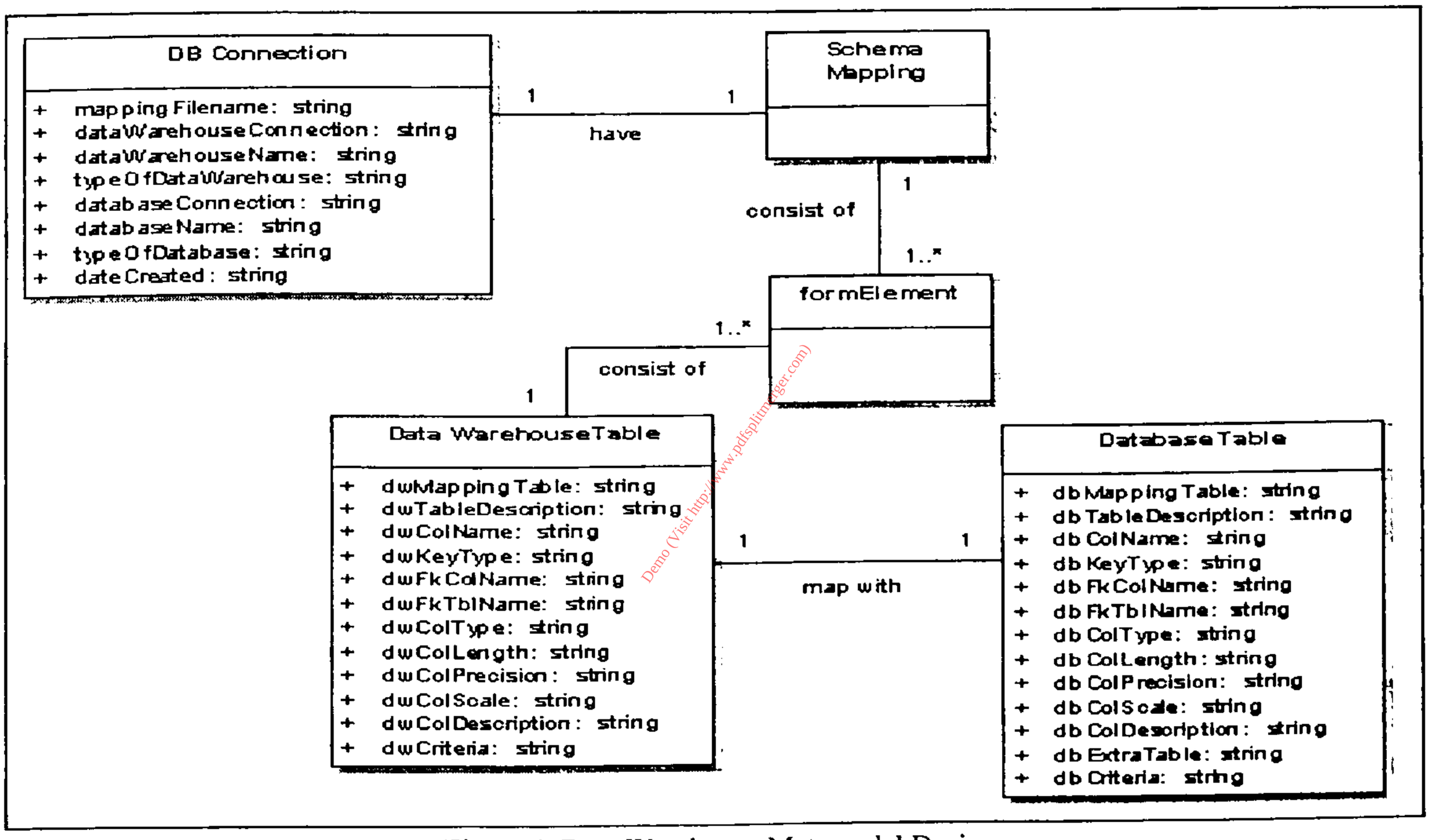

Figure 4 Data Warehouse Metamodel Design

2.4 Conceptual Operational Database Schema Design

Three different types of conceptual operational database schemas are designed and each of the schemas is modeled using different structures and semantics. The semantic differences include homonyms semantic (i.e., different sources use the same name for different constructs) and synonyms semantic (i.e., different names for the same construct). The schema structures represent different business flows although they may have the same business nature (i.e., clinic or hospital). Furthermore, different types of data sources are used to develop the physical database such as SQL Server, Oracle and Access. The reason for doing so is to demonstrate the feasibility of the framework being implemented which allows integration of heterogeneous and distributed data sources.

3 Implementation

The framework developed provides a semi-automated schema mapping engine which adheres to the specified metamodel (Figure 4) and a fully automated Extraction, Transformation and Loading (ETL) engine. erability, flexibility and convenience to users.

3.1 Schema Mapping

The framework proposed here automatically extracts and loads the data source metadata and the data warehouse metadata. It displays a list of tables and their respective attributes in a tree view format to facilitate the mapping process. Users interact with the schema mapping engine through the Agent to identify and select attributes for mapping attributes from the data source to the corresponding attributes in data warehouse tables as shown in Figure 5. Users perform the mapping by comparing the data source structure and the data warehouse structure. Users need to do this manually because only they will know about their organization's business flow. The organizations may have different business flows even if the nature of their business is the same. Thus, only the users can determine the correctness of the schema mapping. This step is crucial because only correct mapping will lead to correct data integration.

The schema mapping metadata is automatically generated according to the specified metamodel. It is structured to XML format. Figure 6 shows an example of schema metadata specifying mapping information.

The mapping is only needed for the first time when data from a new data source is to be integrated to the data warehouse. Subsequently, the metadata is loaded from the system whenever users update the structure of their data sources or the structure of the data warehouse.

3.2 Extraction, Transformation and Loading Process

The schema mapping metadata created from the schema mapping process is used for expressing the Extraction, Transformation and Loading (ETL) process. Algorithms are developed to automate the ETL process. Data extraction involves extracting data from multiple heterogeneous data sources such as SQL Server, Oracle and Access. This may require resolving semantic conflicts and handling multiple data types and data structures. Data is extracted based on specified criteria in the schema mapping metadata and structured according to the data warehouse structure. This eases data integration and loading. Data transformation involves detecting and removing errors, inconsistent data, checking for null values, data concatenation, type conversion and eliminating duplicate records. All these steps improve the quality of data before they are loaded into the data warehouse. Data loading includes integrity constraints and surrogate key management, and integration.

The ETL engine in the Agent is responsible for automating the data extraction and data transformation processes whereas the ETL engine in the Web server is responsible for automating the data transformation and data loading processes. The materialized data integration is achieved by loading the actual data into the data warehouse.

Integrity Constraints Management

To resolve integrity constraints imposed by the data warehouse during the uploading of data into the respective tables, all the independent tables are processed first, followed by the dependent tables. An independent table is one which does not have any foreign keys. A dependent table is one which has one or many foreign keys. The order of uploading tables is important so as to take care of integrity constraints. The rule applied is to upload tables with the least number of foreign keys. The respective tables referenced by these foreign keys must be uploaded first before the tables containing these foreign keys are uploaded.

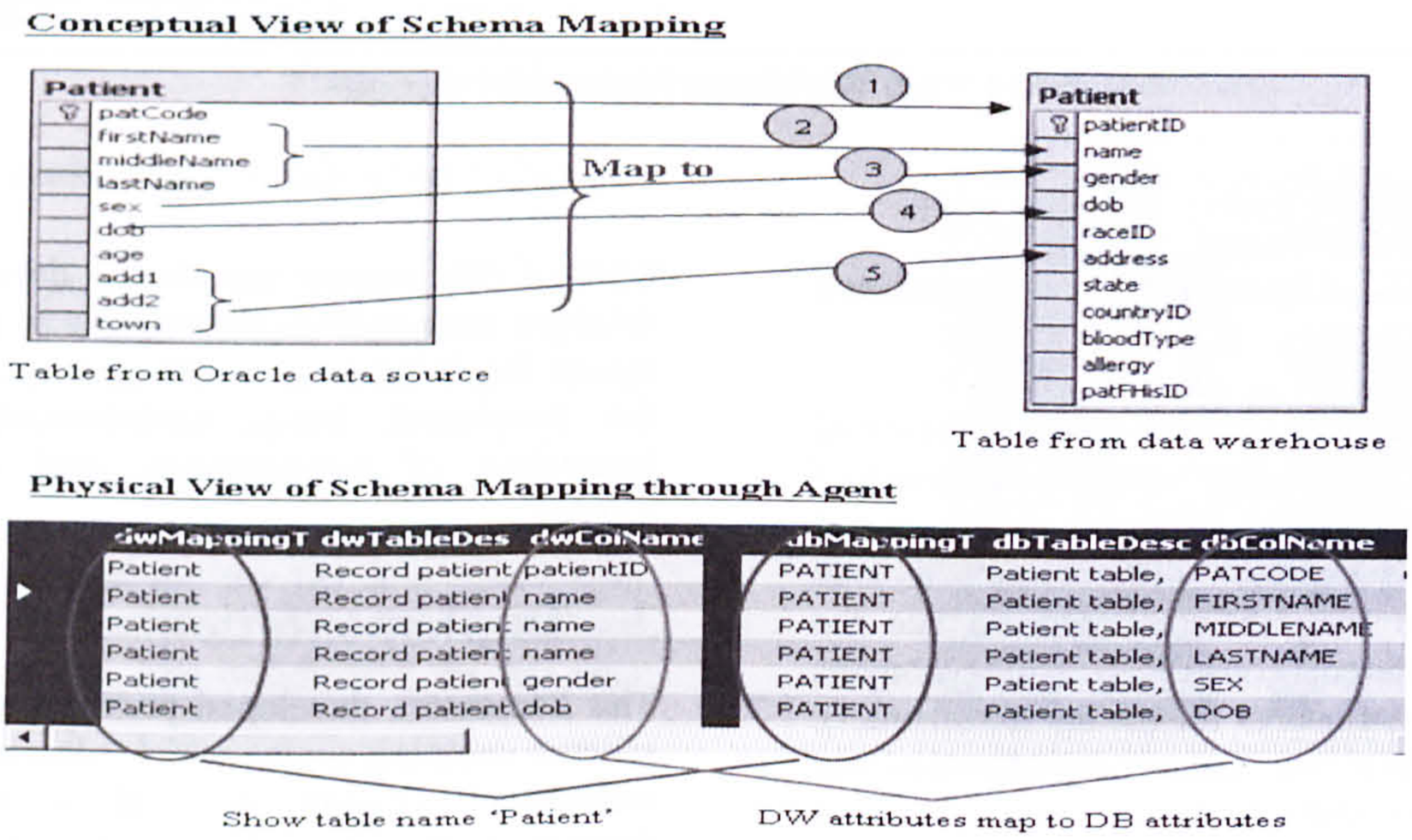

Figure 5 Example of how schema mapping can be performed

db1-mappSschema.mapp

```
<?xml version="1.0" encoding="utf-8"?>
<ArrayOfFormElement
   xmlns:xsd="http://www.w3.org/2001/XMLSchema"
   xmlns:xsi="http://www.w3.org/2001/XMLSchema-
   instance">
 <formElement>
  <dbMappingTable>tblStaff</dbMappingTable>
  <dbTableDescription />
  <dbColName>staffID</dbColName>
  <dbKeyType>PK</dbKeyType>
  <dbFkColName>-</dbFkColName>
  <dbFkTblName>-</dbFkTblName>
  <dbColType />
  <dbColLength />
  <dbColPrecision />
  <dbColScale />
  <dbColDescription />
  <dbExtraTable>tblStaffType</dbExtraTable>
  <dbCriteria>tblStaff.staffTypeID =
     tblStaffType.staffTypeID AND
     tblStaffType.description = Doctor'</dbCriteria>
```

Continue...

```
  <dwMappingTable>Physician</dwMappingTable>
   <dwTableDescription>Record physician personal
      data</dwTableDescription>
   <dwColName>physicianID</dwColName>
   <dwKeyType>PK</dwKeyType>
   <dwFkColName>-</dwFkColName>
   <dwFkTblName>-</dwFkTblName>
   <dwColType>bigint</dwColType>
   <dwColLength>8</dwColLength>
   <dwColPrecision>19</dwColPrecision>
   <dwColScale>0</dwColScale>
   <dwColDescription>Unique
      identifier</dwColDescription>
   <dwCriteria />
 </formElement>
 <formElement>
        :
        : (define next element)
 </formElement>
</ArrayOfFormElement>
```

Figure 6 Sample of Schema Metadata Specifying Mapping Information

Surrogate Key Management

Primary key conflicts arise because different data sources use different formats such as numeric, text or both. A uniform surrogate key is assigned to handle this problem. This avoids having inconsistent or poor quality data. Index tables are created to assign the surrogate key so that it stores both the primary key that is generated automatically by the system and the actual key used to reference to the data. Each data source has a set of index tables to reference the individual data warehouse tables and all these index tables are maintained throughout the life of the data warehouse.

4 Features of the HIIF

The schema mapping framework used in this research supports mapping of heterogeneous data sources which include:

- *Mapping Tables with One-to-One Relationship* – Tables with one-to-one relationship are the most common type. The attributes of a table in the data source are directly mapped to the attributes in the corresponding table in the data warehouse.

- *Mapping Tables with One-to-Many Relationship* - The schema mapping allows mapping of tables with one-to-many relationship. For example, for the tables Diagnosis, Patient and Medicine, one patient may have several diagnoses and one diagnosis may require several medicines.

- *Mapping Tables with Integrity Constraints* - The schema mapping allows mapping of tables with integrity constraints. For example, the foreign key attribute of a table in the data source can be mapped to the corresponding attribute of the table in the data warehouse. Specifying additional conditions may or may not be required depending on how the tables in the data source are structured in relation to the corresponding tables in the data warehouse.

The framework provides flexibility to the users to perform the mapping which include:

- *Flexible Mapping Sequence* - When users perform the schema mapping, they do not have to map according to the sequence of the column specified by the ORDINAL_POSITION of the data warehouse. They are free to map any attribute from any table and in any sequence. Additionally, users can easily delete any attribute mapped or re-map a deleted attribute at any time. Thus it provides flexibility to uses to perform the mapping.

- *Concatenation of Data* - Data extracted from different attributes can be concatenated to form a single attribute. For eample, data extracted from attributes firstName, middleName and lastName can be concatenated to a single attribute called say, Name.

- *Easy Handling of Structures and Semantics Conflicts* - Integration of different data sources with different data structures and semantics can lead to structure and semantic (e.g., homonyms and synonyms) conflicts. Homonym conflicts occur when different data sources use the same name for different attributes while synonyms

conflicts occur when different names are used for the same attribute. Users however need not worry about the structure and semantic conflicts occurring among the data sources and the data warehouse as these can be easily handled by specifying conditions while mapping the attributes. The framework itself handles these semantic conflicts during the integration process using the schema mapping metadata which captures the schema mapping details defined by the user.

- *Allows Easy Update of Schema Mapping Metadata* - The user can easily update the schema mapping metadata generated to accommodate changes made to the structures in the data sources or the data warehouse. Also, users need not convert their existing data models to other formats to use this schema mapping engine unlike the one proposed by Tan and Zaslavsky, [24] where users have to convert their data model to XML format before they can use the schema mapping engine. This schema mapping is thus more flexible.

- *Allows Users to Specify Additional Conditions for Easy Mapping* - The schema mapping allows users to specify conditions while mapping attributes of the data sources to the attributes of the data warehouse. As different organizations may use different data structures, specifying conditions is important as it allows different data source structures to map to the data warehouse structure. This makes the integration of heterogeneous data sources possible. The condition specified is a simple WHERE clause used in an SQL query.

5 Framework Generalization

Although the framework presented here is for the healthcare domain, it can easily be adapted to other application domains such as education, finance and e-government. This is because the approach used is generic. However, some user interfaces will have to be changed when porting the framework to another application domain.

6 Further Research

Currently, the framework supports only three types of databases: Oracle, SQL Server and Access. It can be easily extended to support other databases such as MySQL, dbBase and Interbase. The data warehouse used in this framework is based on the SQL Server and the algorithms for automating data uploading are based on the data warehouse schema design. Any changes to the data warehouse schema design (e.g., adding new tables or attributes) will require updating the data uploading functions as these are implemented by calling insert stored procedures. Further work can be done to automate the data loading process so that it supports other data warehouse schema design.

The framework can be extended to include validation of schema mapping. When mapping attributes in a data source to the corresponding attributes in the data warehouse, most of the related metadata information such as the data types, table and column descriptions, and integrity constraints (primary key or foreign key) are loaded automatically. Users may however want to enter additional tables and conditions. This can create problems such as schema mapping becoming more error prone. For example, users may enter table names, conditions or SQL syntax that are invalid. This can affect the data extraction process since the schema mapping metadata is used for expressing the Extraction, Transformation and Loading (ETL) process. Validation of schema mapping can ensure that the mapping is done correctly.

The framework provides some data cleaning to improve the data quality. The data cleaning work includes detecting and removing errors and inconsistencies, checking for null values, eliminating duplicate records and resolving structure and semantic conflicts. This is done before loading the transformed data into the data warehouse. Currently, the framework only deals with attribute semantic conflicts but not with the semantics conflicts in the actual data. The latter occurs when data are extracted from different data sources with different formats are integrated (e.g., date based on UK format (DD/MM/YYYY) and date based on US format (MM/DD/YYYY)). A similar situation arises when integrating currency and metric data.

Other enhancements can include updating records of the data warehouse based on changes made to the data source, developing an update scheduler, developing OnLine Analytical Processing (OLAP) with data mining to support decision making process, and developing algorithms to auto-generate the dimensional tables to facilitate OLAP and data mining functions. Issues such as scalability, maintainability and security can also be explored.

References

[1] Barker, R. (1998). Managing *data warehouse.* Retrieved June 20, 2004, from Veritas Software Corporation and EOUG Web site: http://eval.veritas.com/webfiles/docs/manage_data warehouse.pdf

[2] Berndt, D., (2001). Consumer decision support systems: a health care case study. *Proceedings of the 34th Hawaii International Conference on System Sciences,* Maui, Hawaii, 1-10.

[3] Bito, Y., Kero, R., Matsuo, H., Shintani, Y., Silver, M. (2001). Interactively visualizing data