UTeM

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

IMPROVING COMPLETE ACCESS TO DATA WITHIN COLLABORATIVE SYSTEMS IN HEALTHCARE DOMAIN

FATHIN NABILLA BINTI MD LEZA

MASTER OF SCIENCE IN INFORMATION AND COMMUNICATION TECHNOLOGY

2016

© Universiti Teknikal Malaysia Melaka

Faculty of Information and Communication Technology

IMPROVING COMPLETE ACCESS TO DATA WITHIN COLLABORATIVE SYSTEMS IN HEALTHCARE DOMAIN

FATHIN NABILLA BINTI MD LEZA

Master of Science in Information and Communication Technology

2016

© Universiti Teknikal Malaysia Melaka

IMPROVING COMPLETE ACCESS TO DATA WITHIN COLLABORATIVE SYSTEMS IN HEALTHCARE DOMAIN

FATHIN NABILLA BINTI MD LEZA

A thesis submitted
in fulfillment of the requirements for the degree of Master of Science
in Information and Communication Technology

Faculty of Information and Communication Technology

UNIVERSITI TEKNIKAL MALAYSIA MELAKA

2016

© Universiti Teknikal Malaysia Melaka

DECLARATION

I declare that this thesis entitle “Improving Complete Access to Data within Collaborative Systems in Healthcare Domain” is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature

Name Fathin Nabilla binti Md Leza

Date

© Universiti Teknikal Malaysia Melaka

APPROVAL

I hereby declare that I have read this thesis and in my opinion, this thesis is sufficient in terms of scope and quality for the award of Master of Science in Information and Communication Technology.

Signature

Name Dr. Nurul Akmar Emran

Date

© Universiti Teknikal Malaysia Melaka

DEDICATION

Every challenging work needs self-efforts as well as
guidance of elders especially those who were very closed to our hearts
My humble effort I dedicate to my loving;

Father and Mother,
Whose affection, love, encouragement and du‟a of day and night
make me able receive such success and honor
Alhamdulillah

To my beloved mother and father, thank you.

© Universiti Teknikal Malaysia Melaka

ABSTRACT

The problem of accessing complete data especially involving different sources had been a major concern in many fields. Due to the absence of data model that become the barrier to access and manage complete datasets, data providers within the collaborative environment may affect users to have fast and seamless accessibility. There had been many data accessibility improvements methods such as: data replication, data sharing, cloud computing and much more. This dissertation had presented the critical review of these methods based on their criteria of accessibility. However, studies to measure complete access to data were limited. A data accessibility model has been proposed within a collaborative environment named as Collaborative Integrated Database System (COLLIDS), in order to overcome the issue to complete access to data. COLLIDS aims to improve complete access to data in multi-data provider‟s context in a case study. The highlighted feature in COLLIDS which is the completeness analyzer provides relative completeness analysis among data provider. Population-based Completeness (PBC) was applied in COLLIDS completeness analyzer to measure the completeness for the dimension of interest, among the population of data providers. It has been hypothesized that there will have an increment in the ratio of completeness of data accessible by data providers. Therefore, COLLIDS have been evaluated in terms of increment of data completeness for all data providers. Further analysis also has been conducted to compute the completeness cases of data providers. In this dissertation, healthcare domain has been chosen as the case study in order to explore the problem of accessing complete data seamlessly. Sample data have been collected from 106 healthcare providers called as „panel clinics‟. PBC have been used in this dissertation to measure the reference population which is the union of patient datasets collected from all the clinics. Two groups of datasets have been measured; „As-Is Completeness‟ and „ Completeness Increment‟ in normality test and Wilcoxon-Sign Rank test, in order to know the significance differences between both groups. The outcome of this dissertation will have to contribute towards understanding for the practical analysis and the evaluation of COLLIDS data model, in measuring complete access to data within multiple data providers‟ environments.

© Universiti Teknikal Malaysia Melaka

ABSTRAK

Masalah mengakses data yang lengkap terutama melibatkan sumber-sumber yang berbeza adalah satu kebimbangan utama dalam pelbagai bidang. Oleh kerana ketiadaan data model telah menjadi halangan untuk mengakses dan menguruskan sebuah dataset yang lengkap, khususnya kepada pembekal data dalam lingkungan kolaboratif, sejurus boleh menjejaskan pengguna untuk mempunyai akses cepat dan lancar. Terdapat banyak penambahbaikan akses kepada data kaedah seperti; replikasi data, perkongsian data, pengkomputeran awan dan sebagainya. Disertasi ini telah membentangkan kajian kritikal kaedah berdasarkan kriteria kaedah-kaedah akses kebolehcapaian data. Walau bagaimanapun, kajian untuk mengukur akses penuh kepada data adalah terhad. Model akses data telah dicadangkan dalam lingkungan kolaboratif yang dinamakan sebagai Sistem Pangkalan Data Bersepadu Kolaboratif (COLLIDS), untuk mengatasi isu akses data yang lengkap. COLLIDS bertujuan untuk meningkatkan akses penuh kepada data dalam konteks pembekal data yang berbilang dalam sebuah kajian kes. Ciri yang diketengahkan dalam COLLIDS adalah penganalisis lengkap yang menyediakan analisis lengkap data berhubung kalangan para pembekal data. Populasi Lengkap (PBC) telah digunakan dalam penganalisis lengkap COLLIDS untuk mengukur lengkap data bagi dimensi yang tertentu antara para pembekal data. Ia telah dihipotesiskan bahawa terdapat kenaikan dalam jangkaan nisbah lengkap data yang diakses oleh para pembekal data. Oleh itu, COLLIDS dinilai dari segi kenaikan data lengkap untuk semua pembekal data. Analisis lanjut juga turut dijalankan untuk menentukan kes lengkap untuk pembekal data COLLIDS. Dalam kajian ini, bidang kesihatan telah dipilih sebagai kajian kes untuk meneroka masalah akses data yang lengkap dengan lancar. Sampel data telah dikumpulkan daripada 106 pusat kesihatan dipanggil sebagai 'klinik panel'. PBC kemudiannya digunakan untuk mengukur populasi rujukan yang merupakan kesatuan antara semua dataset pesakit yang dikumpul daripada semua klinik. Dua kumpulan dataset telah diukur; iaitu 'sedia ada lengkap' dan 'kenaikan lengkap' dalam ujian normal dan juga ujian Wilcoxon-Sign Rank, demi mengetahui perbezaan dan juga signifikasi antara kedua-dua kumpulan. Hasil daripada kajian ini akan menyumbang ke arah pemahaman analisis praktikal dan penilaian model COLLIDS, dalam mengukur akses penuh kepada data dalam lingkungan pembekal data yang kolaboratif.

© Universiti Teknikal Malaysia Melaka

ACKNOWLEDGEMENT

First and foremost, I would like to thank Allah for the will and spirit He has given me, to further complete my study. Thank you to my family especially my beloved father Md Leza Salleh and mother Suraya Shafiee for full support they had given me throughout my study.

Finally I would like to thank my supervisor, Dr. Nurul Akmar Emran who has provided excellent guidance to me throughout my research especially for her support and facilities along this research study. I also would like to thank you to Dr. Yogan Jayakumar, Dr. Samad Hassan Basari and Dr. Gede Pramudya Ananta for guiding me in statistical analysis for this thesis. Also, special thank you to Miss Nuridawati Mustafa that have helped me in giving support and courage when I was having hard time. Not forgotten, to other Faculty of Information Communication and Technology (FTMK) lecturers and my friends who also have been supporting me during this study. Lastly, thank you too, to all my UIC DDPZ students for always making me smiles during my cloudy days. Thank you so much.

© Universiti Teknikal Malaysia Melaka

TABLE OF CONTENTS

© Universiti Teknikal Malaysia Melaka

© Universiti Teknikal Malaysia Melaka

LIST OF TABLES

© Universiti Teknikal Malaysia Melaka

LIST OF FIGURES

© Universiti Teknikal Malaysia Melaka

© Universiti Teknikal Malaysia Melaka

LIST OF ABBREVIATIONS

ABE	-	Attribute Advanced Encryption
AES	-	Standard Based Encryption
AHIMA	-	American Health Information Management Association
CDEC	-	California Data Exchange Center
CDSS	-	Collaborative Data Sharing System
COLLIDS	-	Collaborative Integrated Database System
CTL	-	Clover Transformation Language
CWA	-	Close World Assumption
DBMS	-	Database Management System
DP	-	Data Provider
DQAF	-	Data Quality Assessment Framework
DWR-	-	Department of Water Resources
EHR	-	Electronic Health Records
ETL	-	Extract, Transform, Load
GIS	-	Geographical Information System
HIS	-	Hospital Information System
HRM	-	Hierarchical Replication Model
ICANN	-	Internet Corporation for Assigned Names and Numbers
ICD10	-	International Statistical Classification of Diseases and Related Health

© Universiti Teknikal Malaysia Melaka

		Problems 10th Revision
IDE	-	Integrated Development Environment
IODE	-	The International Oceanographic Data and Information Exchange
IRWM	-	Integrated Regional Water Management
IWRIS	-	Integrated Water Resources Information System
LAN	-	Local Area Network
NBC	-	Null-based Completeness
ODP	-	Ocean Data Portal
OGC	-	Open Geospatial Consortium Web Services
OTP	-	One-Time Password
PBC	-	Population-based Completeness
QC	-	Query Completeness
QR	-	Quick-Response
RDF	-	Resource Description Framework
SBC	-	Schema-based Completeness
SQL	-	Standard Query Language
TBC	-	Tuple-based Completeness
TC	-	Table Completeness
USGS	-	United States Geological Survey
UTeM	-	Universiti Teknikal Malaysia Melaka
WHO	-	World Health Organization

© Universiti Teknikal Malaysia Melaka

LIST OF APPENDICES

© Universiti Teknikal Malaysia Melaka

LIST OF RELATED PUBLICATION

1. M.Leza, F.N. and Emran, N.A., 2014. Data Accessibility Model Using QR Code for Lifetime Healthcare Records. World Applied Sciences Journal (Innovation Challenges in Multidiciplinary Research & Practice), vol 30, pp.395–402.
2. M.Leza, F.N., Emran, N.A., and MKA. Ghani, 2014. Review of Data Accessibility Methods In Healthcare. Science International, vol 26 , issue 5, pp.1733–1737
3. M.Leza, F.N., Emran, N.A., and MKA. Ghani, 2013. Improving Data Accessibility Using QR Code in Healthcare Domain. In: e-Proceeding of Software Engineering Postgraduates Workshop (SEPoW) Innovative Software Engineering for Creative and Co-Organizer. pp.119–123.

© Universiti Teknikal Malaysia Melaka

CHAPTER 1

INTRODUCTION

In this chapter, research background and summary of research work, research goals and the significance of the research work are presented. Section 1.1 explains the research background that provides the motivation of the work. Section 1.2 presents research aim and objectives followed by research contributions in Section 1.3. Research scope is also provided in Section 1.4. Finally, Section 1.5 presents the thesis outline.

1.1 Research Motivation

In many database operating environment, the fundamental operation is data accessibility. Data accessibility is an operation triggered by users in order to retrieve desired information. Data accessibility, accessibility (in the context of access to data), or data access have similar definitions in many contexts/aspects.

According to Clio (2008), data accessibility is the ability to access the data regardless of natural or man-made disasters, hardware failures or in any circumstances. In terms of data quality, according to Pipino et al. (2002), data accessibility is claimed as the range to which data is available or easily and quickly retrievable. The term of “data accessibility” also can be referred as other terms; “accessibility” or “data access” which brings similar definitions in different domains.

© Universiti Teknikal Malaysia Melaka

The definitions of data accessibility so far signifies its‟ importance in data retrieval. Ideally, the retrieved data is complete and the process of retrieval is fast and seamless in many domains. The inability to retrieve complete data is not only frustrating but also may jeopardize crucial operation such as decision making. For instance, the use of data in data warehousing to support decision making for one company can be affected due to incomplete data, as reported by Eckerson (2002) and Ballou and Tayi (1999).

Incomplete data also can affect data analysis. For example, according to white paper report from SPSS (2009), incomplete data can be a serious problem as it can lead to misleading results of analysis (i.e. cannot run factor analysis, and results may not be statistically significant). In the end, the result obtained due to this problem is not accurate as the analysis ended up not analyzing what supposed to be measured (SPSS, 2009).

Incomplete data due to restricted access can result in to time-consuming procedures. For example, in healthcare domain, incomplete data may cause repeatable procedures where patients are required to undergo the same process (i.e. registration process, blood test and urine test). In another example, incomplete data causes many organizations (i.e. telecommunication firms, insurance company, regional bank, information services firm and global chemical company) to suffers huge losses in terms of profit as mentioned by Eckerson (2002). Accessing complete data in fast and seamless[1] manner is challenging especially when involving multiple data providers. This is due to several factors that become the challenges to gain access to complete data. This problem will be described in next sub-section.

[1] Seamless is defined as smooth and continuous which is without difficulty, with no apparent gaps or spaces between one part and the next. Retrieved from Oxford Dictionaries at http://www.oxforddictionaries.com/definition/english/seamless

© Universiti Teknikal Malaysia Melaka

1.1.1 Factors Affecting Data Accessibility

Each data provider enforces different security requirements, and access control procedures. As a result, to retrieve data from these data providers one must fulfill different data access requirements and different security levels (less stringent or more stringent). For example, sensitive and confidential data require more stringent security requirements (i.e. bypassing firewalls, acquire exclusive access privileges) because exposure of these data can affect organization operational and financial due to data leakage. According to McCarty (2015) there have been reports that confidential data has been exposed 330 times, in *Internet Corporation for Assigned Names and Numbers* (ICANN) (an Internet Domain Name organization). This security breach unfortunately has impacted 96 applicants whose confidential data have been compromised (McCarthy, 2015). Similar incidents have been reported by ICANN several times which causes delay of the organization"s landmark program launching.

Secondly, fast and seamless access to complete data is hindered by the nature of data management structure. The best scenario to observe this problem is in healthcare environment. In healthcare, a patient commonly seeks treatment from more than one healthcare provider. As a result, his /her treatment records are stored in multiple healthcare providers. Getting complete treatment records require manual records integration which is not only impractical, but also time consuming. Due to this problem, a patient is usually required to experience painful tiring and time-consuming medical procedure in order to get complete crucial treatment records such as immunizations, medical screening, ante-natal and post-natal events (Abd Ghani et al., 2008). Therefore, data sharing is not in practice even though these healthcare providers are supposed to work collaboratively. Practically, data providers (healthcare providers) are still operating in silos.

© Universiti Teknikal Malaysia Melaka

Thirdly, access to data that are stored in database systems usually requires network connectivity. Some network connectivity such as ad-hoc networks (especially mobile connections) are unstable (Rani et al., 2013). The speed and reliability of mobile connections (Rani et al., 2013) can vary from network to network and might lead to unstable connections that hinder data accessibility. Within multiple data-providers environment, fast and seamless access to complete data can be affected due to network problems inherent in data provider‟s network systems. Since data providers‟ network connectivity may vary, complete data retrieval depends on successful access to data. Therefore, stable network connectivity is a requirement to retrieve complete data seamlessly.

Fourthly, heterogeneous access methods complicate fast and seamless access to complete data. Within multiple data providers‟ environment, it is unlikely to find data providers who use the same access method that is familiar to all users. Even though web-based method is a common access method used by many data providers nowadays, each data provider maintain its own database in isolation. Consequently in order to acquire complete data, users must gather and integrate the data manually. Some data providers who uses biometric systems requires the users to fulfill biometric authentication before access to data can be granted. In addition, access methods which rely on the use of devices such as code reader (i.e. barcode, QR Code), and smart card reader requires the users to setup the devices prior to data access. Unless users are willing to adapt with access methods heterogeneity, seamless access to complete data can be achieved.

Finally, the fifth challenge in retrieving complete data seamlessly is to deal with the quality of data (Olson, 2003). As data are residing at multiple data providers‟ databases, the quality of it is subject to data quality imitative taken by the data providers. Common data problems are inconsistency (i.e. format heterogeneity and conflicting values, data

© Universiti Teknikal Malaysia Melaka

duplication), accuracy and missing values (Eckerson, 2002). For example, format heterogeneity normally caused when different data providers have different database structures that have different data formats. Data that are of different formats are considered as „dirty‟ and are not fit for use. In order to use the data, a data cleaning process is required. The process of data cleaning that most commonly used today is ETL (Extract, Transform, Load) (Rahm and Do, 2000). Through ETL process, which provided with tools helps in cleaning the data into desired data formats thus reducing the errors. This data cleaning process delays the effort to acquire complete data and therefore fast and seamless access to data is hindered.

The challenges presented in this section are the obstacles in retrieving complete datasets in fast and seamless manner, especially in multi-data provider‟s context. Due to the challenges, several proposals can be found in the literature that attempt to deal with the problems. For example, data integration methods (i.e. data sharing (Smith, 1994; Ives et al., 2008a; Tang et al., 2011), data replication (Hara, 2001; Hara and Madria, 2006; Lu et al., 2010), and data federation (Lans, 2010; Danyaro et al., 2014)) are proposed to improve data completeness where different datasets are integrated into single unified view of all data (Lans, 2010). Nevertheless, these methods do not aim to provide fast and seamless access to complete data. Consequently, issues regarding the practicality of these methods remain uncovered. In addition the way completeness is measured is missing from the work. Hence, the limitations in the current studies of data accessibility and data completeness are the motivation of this research. Specifically, the following are the research problems (RPs) that have been identified:

© Universiti Teknikal Malaysia Melaka

- RP1: There is limited comprehensive studies on how complete data can be accessed, particularly in collaborative systems in healthcare domain
- RP2: Lack of data accessibility models that put emphasis on accessing complete data using multiple access methods. Therefore, understanding of the practicality and implementation is limited.
- RP3: Studies that aim to improve data completeness do not cover ways to measure completeness within multi-data provider environment.

The motivation of this research is therefore to answer the following fundamental research questions (RQs) that address the research problem stated above:

- RQ1: What are data accessibility methods available and why these methods are proposed?
- RQ2: How to design data accessibility model that can improve completeness?
- RQ3: How to evaluate the proposed data accessibility model?

In the next section will provide the aim and objectives of the research that are formulated to answer the above questions.

© Universiti Teknikal Malaysia Melaka