DATA QUALITY IMPROVEMENT OF STUDENT DATA SUBMISSION TO MYMOHES

MAJINA BINTI SULAIMAN

A dissertation submitted
in partial fulfilment of the requirements for the degree of
Master of Advanced Information Technology

Faculty of Computer Science and Information Technology
UNIVERSITI MALAYSIA SARAWAK
2012
Dedication

Untuk ayah dan ibu tercinta
yang sentiasa memberiku dorongan untuk berjaya
dan menjadi manusia yang berakal
serta dihiasi dengan ilmu agama
Declaration

I certify that all works in this dissertation are carried out between February 2010 and April 2012 and they have not been submitted for any academic awards at other colleges, institutes or universities. The work presented here is carried out under the supervision of Dr Noor Alamshah bin Bolhassan. All other works in the dissertation are my own except those where noted.

Signed,

Majina binti Sulaiman

, May, 2012

Acknowledgment

Alhamdulillah, I would like to thank all of those people who helped to make this dissertation possible. First, I wish to thank my advisor, Dr Noor Alamshah bin Bolhassan for all his guidance, encouragement, support, and patience. Also, I would like to thank Dr Mohd Ibrahim Safawi bin Mohd Zain for his very helpful insights and comments. Special thanks to my family members, my “littlebug” Siti Rahimah Abdul Halim and my best friend Zuraina Abdul Jahar for their support and motivation which gave me courage in completing this dissertation. Finally, I would also like to thank to my colleagues who have involved with MyMoheS project in UNIMAS for their cooperation and technical support.
# Table of Contents

Dedication................................................................................................................... i  
Acknowledgment......................................................................................................... ii  
Table of Contents....................................................................................................... iii  
List of Figures............................................................................................................. v  
List of Tables............................................................................................................. vi  
List of Acronyms........................................................................................................ vii  
Abstract....................................................................................................................... viii  
Abstrak........................................................................................................................ ix  

**Chapter 1: Introduction** ......................................................................................... 1  
1.1 Introduction ......................................................................................................... 1  
1.2 Background of Study.......................................................................................... 1  
1.2.1 MyMoheS ....................................................................................................... 1  
1.2.2 Data Quality ................................................................................................ 3  
1.3 Importance of Data Quality to UNIMAS ....................................................... 6  
1.3.1 Performance Measurement ......................................................................... 6  
1.3.2 Enables Competitive Advantage and Business Effectiveness .................. 7  
1.3.3 Support Better Internal Data Analysis ......................................................... 7  
1.3.4 Support Data Integration ............................................................................. 8  
1.4 Objectives ......................................................................................................... 9  
1.5 Statement of Problem ....................................................................................... 9  
1.5.1 Data Erroneous ............................................................................................ 10  
1.5.2 Data Source Dependencies .......................................................................... 11  
1.5.3 Frequency of Data Update .......................................................................... 11  
1.5.4 Data Collection and Expertise Limitation ................................................. 11  
1.6 Scope of work ................................................................................................... 12  
1.7 Chapter summary ............................................................................................. 13  

**Chapter 2: Literature Review** .............................................................................. 14  
2.1 Introduction ....................................................................................................... 14  
2.2 Data Quality Framework................................................................................... 15  
2.2.1 Complete Data Quality Methodology (CDQM) ........................................ 15  
2.2.2 Total Data Quality Management (TDQM) ................................................. 17  
2.2.3 Data Governance Framework ..................................................................... 18  
2.3 Data Cleansing Tools......................................................................................... 20  
2.3.1 Tool Number One: Trillium Software ....................................................... 21  
2.3.2 Tool Number Two: DataFlux ...................................................................... 24  
2.4 Chapter Summary ............................................................................................ 27  

**Chapter 3: Methodology** .................................................................................... 29  
3.1 Introduction ....................................................................................................... 29  
3.2 UNIMAS Data Quality Strategy for MyMoheS ............................................. 29  
3.3 MyMoheS Data Quality Framework ................................................................ 31  
3.3.1 Define Roles and Responsibilities .............................................................. 33  
3.3.2 Define Type of Data Erroneous ................................................................. 35  
3.3.3 Define Dimensions of Data Quality ......................................................... 36  
3.3.4 Data Assessment Work Flow ..................................................................... 38  
3.3.5 Iteration and Quality Improvement Plan ................................................... 44
List of Figures

Figure 1: The TDQM Cycle ................................................................................................... 17
Figure 2: Data Governance Life Cycle ................................................................................ 19
Figure 3: Data Quality Tools Magic Quadrant 2011 ............................................................... 21
Figure 4: DataFlux Data Management Methodology .......................................................... 27
Figure 5: Data Quality Strategy Outline ............................................................................ 30
Figure 6: MyMoheS Data Quality Framework for UNIMAS .............................................. 33
Figure 7: Sample of Data Erroneous .................................................................................. 35
Figure 8: Data Mapping for Student Status ....................................................................... 40
Figure 9: Student Data Standardization and Integration ...................................................... 43
Figure 10: Quality Improvement Plan Cycle ...................................................................... 45
Figure 11: Database Connection of Central Database and Node Database ......................... 48
Figure 12: Query Processing for Data Extraction ............................................................... 49
Figure 13: Create Table Using METADATA ...................................................................... 49
Figure 14: Data Cleansing Criteria ..................................................................................... 50
Figure 15: Scheduled Tasks for Data Cleansing ................................................................. 51
Figure 16: Log and Report Analysis ................................................................................... 53
Figure 17: Output Data Analysis ....................................................................................... 54
Figure 18: Level of Data Quality ....................................................................................... 59
Figure 19: Data Cleansing and Validation Process ............................................................. 61
Figure 20: List of Invalid Data Fields with Error Detectors ............................................... 63
Figure 21: First Data Assessment Graph Analysis ............................................................... 64
Figure 22: First Data Assessment Based on Data Quality Metrics ..................................... 64
Figure 23: Second Data Assessment Graph Analysis .......................................................... 65
Figure 24: Second Data Assessment Based on Data Quality Metrics ................................. 66
Figure 25: Third Data Assessment on Graph Analysis ........................................................ 67
Figure 26: Third Data Assessment Based on Data Quality Metrics ..................................... 67
Figure 27: Graph of Results Comparison ......................................................................... 69
Figure 28: UNIMAS Data Model ...................................................................................... 77
List of Tables

Table 1: Eight Common Keys of Data Quality Dimensions ................................................ 5
Table 2: List of Roles and Responsibilities ..................................................................... 34
Table 3: Sample of Gap Analysis Data .......................................................................... 39
Table 4: Sample of Validation Criteria .......................................................................... 41
Table 5: Sample of Rejection Codes ............................................................................... 42
Table 6: Data Quality for Completeness and Accuracy .................................................... 57
Table 7: Validation Criteria for Data Field [postcode] .................................................... 62
Table 8: Sample of Validation Criteria for Data Transformation ................................... 68
Table 9: Comparisons on Data Assessments .................................................................. 68
List of Acronyms

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>APA</td>
<td>Academic Performance Audit</td>
</tr>
<tr>
<td>BSC</td>
<td>Balanced Scored Card</td>
</tr>
<tr>
<td>CDQM</td>
<td>Complete Data Quality Methodology</td>
</tr>
<tr>
<td>DBA</td>
<td>Database Administrator</td>
</tr>
<tr>
<td>DGI</td>
<td>Data Governance Institute</td>
</tr>
<tr>
<td>ETL</td>
<td>Extract, Transfer and Load</td>
</tr>
<tr>
<td>HCM</td>
<td>Human Capital Management System</td>
</tr>
<tr>
<td>IFAS</td>
<td>Integrated Finance and Accounting System</td>
</tr>
<tr>
<td>ILS</td>
<td>Integrated Library System</td>
</tr>
<tr>
<td>InMinds</td>
<td>Intelligent Mining and Decision Support System</td>
</tr>
<tr>
<td>IHL</td>
<td>Institutional Higher Learning</td>
</tr>
<tr>
<td>IPTA</td>
<td>Institut Pengajian Tinggi Awam</td>
</tr>
<tr>
<td>IRIS</td>
<td>Integrated Research Information System</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>JKBPD</td>
<td>Jawatankuasa Bertindak Pangkalan Data</td>
</tr>
<tr>
<td>JPA</td>
<td>Public Service Department of Malaysia</td>
</tr>
<tr>
<td>MDM</td>
<td>Master Data Management</td>
</tr>
<tr>
<td>MoHE</td>
<td>Ministry of Higher Education</td>
</tr>
<tr>
<td>MyRA</td>
<td>Malaysia Research Assessment</td>
</tr>
<tr>
<td>ODBC</td>
<td>Open Database Connectivity</td>
</tr>
<tr>
<td>QAD</td>
<td>Quality Assurance Division</td>
</tr>
<tr>
<td>Setara</td>
<td>Rating System for Malaysian Higher Education Institutions</td>
</tr>
<tr>
<td>SPP</td>
<td>Sistem Pengurusan Pelajar</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>TDM</td>
<td>Total Data Management</td>
</tr>
<tr>
<td>TDQM</td>
<td>Total Data Quality Management</td>
</tr>
<tr>
<td>UNIMAS</td>
<td>Universiti Malaysia of Sarawak</td>
</tr>
</tbody>
</table>
Abstract

MyMoheS system is a data collecting system that has been established by the Ministry of Higher Education (MoHE) for the purpose of publication in order to generate information and knowledge that will indicate the true developmental progress in the national higher learning sector. Thus, each institutional higher learning (IHL) in Malaysia strives to enhance the credibility of the supplied data and to ensure data quality of the institution remains at a good level. Hence, each supplied data is reviewed and evaluated on the level of data quality so that the data will exclude inconsistent data, blank field data, unspecified status and incomplete information before submission by the IHL. To ensure a quality data transmission process, a data quality framework is proposed to help UNIMAS identify and improve current data quality with implementation of data quality metric and validation criteria that meet the needs of data requirement. During data analysis, three types of data assessments was carried out using different validation criteria based on data incompleteness, data inaccuracy and data transformation. The results of these data assessments show the quality of data can be enhanced by combination of correct data incompleteness check, data inaccuracy check and data transformation. Nevertheless, to guarantee data quality in the long run, a work process for monitoring the audit and review process and the updating of data by an appointed officer has been proposed.
Abstrak

Pengumpulan data Sistem MyMoheS telah diwujudkan oleh Kementerian Pengajian Tinggi (KPT) untuk tujuan penerbitan bagi menjana maklumat serta pengetahuan yang dapat menunjukkan petunjuk sebenar kemajuan sektor pengajian tinggi negara. Oleh yang demikian, setiap Institusi Pengajian Tinggi Awam (IPTA) di Malaysia berusaha untuk mempertingkatkan kredibiliti data yang dibekalkan dan memastikan kualiti data institusi sentiasa berada pada tahap yang baik. Oleh yang demikian, setiap data yang dibekalkan hendaklah disemak dan diteliti tahap kualiti datanya supaya tiada masalah data tidak konsisten, medan data yang loompong, data masih berstatus tidak dinyatakan dan maklumat tidak lengkap sebelum dihantar oleh IPTA. Bagi memastikan proses penghantaran data yang berkualiti ini, satu rangka kerja kualiti data dicadangkan bagi membantu UNIMAS untuk mengenalpasti dan memperbaiki masalah kualiti data dengan penetapan metrik data kualiti dan pengesahan kriteria yang memenuhi keperluan data sebenar. Dalam proses analisis data, tiga jenis penilaian data telah dijalankan dengan menggunakan pengesahan kriteria yang berbeza berdasarkan kepada semakan data tidak lengkap, ketidaktepatan data dan transformasi data. Keputusan penilaian ini menunjukkan kualiti data boleh dipertingkatkan dengan penggunaan semakan yang betul bagi data tidak lengkap, ketidaktepatan data dan transformasi data. Justeru itu, bagi menjamin kualiti data untuk jangka masa panjang, satu pelan proses kerja dicadangkan untuk pemantauan audit dan proses penyemakan serta pengemaskinian data oleh pegawai yang dilantik.
Chapter 1: Introduction

1.1 Introduction

In the last decade, organizations started to use computer application systems to store data electronically and process the data to support daily operations and to make decisions. When the data growing increased, the demands on data integration, analysis and exchange as well as data quality issues also increased. Quality of data is affected by the way the data is entered, stored and managed. Poor data quality such as incorrect or incomplete information has given serious consequences to both the business and organisation itself which leading to inaccurate facts and analysis. Maintaining data quality requires going through the data periodically with a proper data improvement process plan. Strong, Lee, and Wang (1996) cautioned that information-system professionals should not focus on how to improve data accuracy only but also consider the data accessibility and data relevance in context with the data consumer’s tasks.

The work described in this chapter focuses on improving quality of UNIMAS data by defining the background of study, the current process and problem statement of data preparation and importance of data quality to UNIMAS in other perspectives.

1.2 Background of Study

1.2.1 MyMoheS

MyMoheS is an information system that was initiated and developed by the Ministry of Higher Education (MoHE) to streamline and integrate data from all IHL to support
exchange and sharing of information data among IHL, MoHE, Public Service Department of Malaysia (JPA), Treasury Malaysia and other government agencies in need of information uniformity. MyMoheS helps the IHL and Research & Development Division of MoHE to prepare planning, policy formation and development of management information systems on higher education more efficiently.

The main objectives of MyMoheS are (i) to consolidate data from all IHL into a central database, (ii) to enable the information to be accessed quickly and easily, (iii) to build a versatile data warehouse that can be used for management and monitoring; and (iv) to provide a facility for periodic and on-demand basis reporting. Therefore, the success of MyMoheS is highly dependent on data accuracy and data quality that provided by each IHL.

Historically, MyMoheS started its operation during semester 2 academic sessions 2007/2008 with two data intakes in a year. In 2009, this system was upgraded and a star rating method was also introduced by MoHE for the purpose of viewing the level of IHL readiness in management and data supplied according to time submission, data completeness and quality data. MyMoheS is consists of four (4) modules namely student module, staff module, research and consultation module and institutional information module.

For the MyMoheS data preparation in UNIMAS, only few IT officers were involved in 2007. At that time, the project coordinator was the UNIMAS Corporate Division. Later in August 2008, Quality Assurance Division (QAD) was established to function as the monitoring body of all work processes related to quality. QAD formed a task force that
responsible for data entry and data preparation for MyMoheS. This task force is known as Jawatankuasa Bertindak Pangkalan Data UNIMAS (JKBPD) has been appointed by Vice Chancellor of UNIMAS which consists of System Developers, Database Administrator (DBA), Data Entry Officers and System Owners. The task force committee is responsible for collection, coordination and to prepare UNIMAS data for the use of stakeholders.

Objectives of this task force committee are:

i) to provide a central database for multi-data source from all division in UNIMAS;
ii) to manage the process of updating and entering data into operational database;
iii) to ensure valid current information of central database;
iv) to support and supply data for projects such as MyMoheS, MyRA, SETARA, APA and
v) to provide the need for an efficient and system effectiveness to meet the new data requirement in future.

1.2.2 Data Quality

From a research perspective, Carlo Batini & Monica Scannapieca in 2006, gave a brief history of data quality by addressing data quality in different areas; namely (i) statistics which first investigated data quality problems using a mathematical theory to conceive duplicates in statistical data set in the late 1960's, then followed by (ii) management in the early 1980's which focused on how to control data manufacturing systems to detect and eliminate data quality problems and (iii) finally in the beginning of the 1990's, computer science starts to analyse the problems of defining, measuring and improving the quality of electronic data stored in databases, data warehouses and legacy systems.
The quality guru, J.M. Duran (1964) explained that data is of high quality when it is “fit for their intended uses in operations, planning and decision making”. In other words, if the data allows a person to do an effective job in operations, planning and making a correct decision, it is deemed to be quality data.

The quality practitioner, Crosby wrote his first book entitled “Quality is Free” (Crosby, 1980) and defined data quality as “the expense of doing things wrong”. It means a data quality can save a company more money than it costs to prevent defects. Crosby added that the quality can be achievable, measurable and profitable when commitment and understanding of all levels in an organization is built.

According to Health Information and Quality Authority, Dublin (2011), data quality has been defined as:

“the totality of features and characteristics of a data set, that bear on its ability to satisfy the needs that result from the intended use of the data”.

In order to consider data as a good quality, the correct data must be available to support decision makers anytime and anywhere they require; and in a format in which they can rely on.

Another aspect or feature of data quality is data quality dimension, which is used to define, measure and manage the quality of data and information. David Loshin (2006) categorized data quality dimension into data model or conceptual view, data values and
data representation. Table 1 defines the eight (8) common keys of data quality dimension.

Table 1: Eight Common Keys of Data Quality Dimensions

<table>
<thead>
<tr>
<th>Dimensions</th>
<th>Definition</th>
<th>Sourced by</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility</td>
<td>Accessibility of data refers to how easily it can be accessed; the awareness of data users in what data is being collected and knowing where it is located</td>
<td>Health Information and Quality Authority (2011)</td>
</tr>
<tr>
<td>Accuracy</td>
<td>Accuracy of data refers to how closely the data correctly captures what it was designed to capture. Verification of accuracy involves comparing the collected data to an external reference source that is known to be valid.</td>
<td>Health Information and Quality Authority (2011)</td>
</tr>
<tr>
<td>Completeness</td>
<td>Completeness of data refers to the extent to which the data collected matches the data set that was developed to describe a specific entity. Monitoring for incomplete lists of eligible records or missing data items will identify data quality problems.</td>
<td>Health Information and Quality Authority (2011)</td>
</tr>
<tr>
<td>Uniqueness</td>
<td>The dimension of uniqueness is characterized by stating that no entity exists more than once within the data set. When there is an expectation of uniqueness, data instances should not be created if there is an existing record for that entity.</td>
<td>David Loshin (2006)</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Timeliness refers primarily to how current or up-to-date the data is at the time of release, by measuring the gap between the end of the reference period to which the data pertains and the date on which the data becomes available to users.</td>
<td>Health Information and Quality Authority (2011)</td>
</tr>
<tr>
<td>Integrity</td>
<td>The inability to maintain a single representation for each entity across systems poses numerous vulnerabilities and risks.</td>
<td>David Loshin (2006)</td>
</tr>
<tr>
<td>Validity</td>
<td>Validity of data refers to data that has been collected in accordance with any rules or definitions that are applicable for that data. This will enable benchmarking between organizations and over time.</td>
<td>Health Information and Quality Authority (2011)</td>
</tr>
<tr>
<td>Consistency</td>
<td>Consistency refers to data values in one data set being consistent with values in another data set. In other words, two data</td>
<td>David Loshin (2006)</td>
</tr>
</tbody>
</table>
values drawn from separate data sets must not conflict with each other; for example the sum of the number of students registered for each programme should not exceed the number of students for the entire university.

1.3 Importance of Data Quality to UNIMAS

Data is the most essential component of today’s business process. Without good data quality, it will be difficult to get accurate report metrics. Due to bad quality of data, users will get frustrated and lose their trust in the data. Below are the reasons why data quality is important to UNIMAS.

1.3.1 Performance Measurement

Kathryn E. Newcomer (1997) defined performance measurement as

“the selection and use of quantitative measures of capacities, processes, and outcomes to develop information about critical aspects of activities, including their effect on the public”.

It is an important step to review the performance of an organization when formulating the direction of the strategic planning or activities and measure the strengths and weaknesses of the organization. As part of “Plan-Do-Check-Act” cycle, measurement plays a key role in quality and productivity improvement activities to meet customer requirements, to provide standards for establishing comparisons and visibility i.e. a “scoreboard” for monitoring performances. But the most important aspect is to highlight
quality problems and determine areas of priority and understand the impact on business performance, on sustaining current performance and reducing any drop in performance.

1.3.2 Enables Competitive Advantage and Business Effectiveness

Larry English (1999) said quality information enables competitive advantage. The much improved quality of data will help UNIMAS attain increased performance by improving the efficiency and effectiveness of key business processes which can result in a key competitive advantage.

Redman (1995) also cites the successful management of data and information will provide a means of competitive advantage. Despite this, data failures are embarrassing and show the weaknesses of management of an organization. For example, an analysis on courses popularity among IHL helps the university to measure the demands on particular courses applied for by students among IHL.

1.3.3 Support Better Internal Data Analysis

Other than MyMoheS, data quality can also support better internal data analysis in UNIMAS. According to QAD, UNIMAS BSC aims to adopt a model contrived by two Harvard University professors; Norton and Kaplan which encompasses four vital perspectives namely financial, customers, internal process, learning and growth in any organization that covers both short and long term measures. The design of UNIMAS BSC is to determine whether current performance “meets expectations” in achieving
UNIMAS vision, mission and goals. It sets benchmark standards to review UNIMAS achievements in a 5-year-cycle period.

Other than UNIMAS BSC, Intelligent Mining and Decision Support System (InMinds) is another analysis tool that provides statistical analysis, data mining, online analytical processing and reporting to the top and mid-management of UNIMAS for better decision making. This project was launched in 2010, which covers Academic Area including Analysis of Student Achievement in Learning Outcome and Program Outcome Assessments, Analysis of Lecturer Workload, Statistic of Student Intake and etc. These tools definitely offer an ideal solution for the UNIMAS management to analyse current and future business performance across the university.

1.3.4 Support Data Integration

The data integration involves a process of consolidating data that resides in different data sources and provides a unified view of data to users. Data quality solutions in data integration help users to discover how multi data are wrongly formatted, incomplete, redundant, and completely missing. To achieve good data quality, data preparation process must follow a standard of data structure and reference data.

Other than integrating quality data, data integration allows users to share data repository for data analysis as explained in Section 1.3.3. This can ensure the organization uses good data quality to facilitate effective analysis for the right decisions at the right time.
1.4 Objectives

In order to provide a platform as data supplier to MyMoheS, the objectives of this study are:

i) To identify weaknesses of the current data preparation process which is time consuming and overly complex. Currently, the data is extracted from multiple sources by several System Developers using separate query statements. Practically, this process is not the best practice for data preparation and data consolidation. Thus, a model or framework for unified definition of data quality across different domains should be formed.

ii) To detect errors, improve data quality and transmission; as well as to enhance the level of data quality and integrity through resources optimization and decrease data erroneous. The improvement process and data cleansing initiatives give the ability for the business to sustain data quality throughout the organization.

iii) To improve the business process; manage data accuracy and timeliness upon MyMoheS data submission.

1.5 Statement of problem

MoHE must face the fact that each respective IHL is managing different databases with different platforms and data schema. Due to this, data errors are usually present for the following reasons:-
1.5.1 Data Erroneous

According to Ranjit & Kawaljeet (2010), data quality issues can happen at data sources, data profiling stage, data modelling and data staging ETL (Extract, Transformation and Loading) processes.

The common types of data errors which are found during data submission to MyMoheS are;

i) Missing data values are often typing mistakes which happened at operational source or through ETL processing. For example, the records are either missing a numeric or alpha field is blank.

ii) Bad data values referring to numbers and values which may be valid for a given field but are not so in the context of a particular event or range. For example, a numeric field has alpha value or data value like “02/05/1075” for a birth date.

iii) Referential data errors are incorrect values that are hardly detectable without thorough inspection. The data can be incorrectly extracted or entered as there is no referential value to identify the error. This is where most student data gets mixed up. For example; invalid student postcode address or invalid telephone number format.

iv) Errors that set by business rules, for example; the school type is a compulsory field for Malaysians but this rule is not applied during data extraction.

v) Obscure understanding of the field’s perusal by MyMoheS secretariat like “Tidak Nyatakan” and “Tidak Berkenaan”.

10
1.5.2 Data Source Dependencies

The integration of multi-data sources needs data source dependencies. The main problems that relates to the integration of these data sources are the identification and solution of conflicts between data schema and the data itself. To standardize the data submission by IHL, MoHE has provided reference tables which consist of general codes such as code of state, code of country and etc. Hence, the data accuracy highly depends on MyMoheS data source and later it will ease the process of data integration and consolidation.

1.5.3 Frequency of Data Update

Frequent changes on MyMoheS reference data will reflect the data extraction process and increase the number of data erroneous. Thus, it is not an easy task for IHL to maintain current data structure and values when the new data reference needs to be updated based on MyMoheS new requirements. This might involve changing of current data structure, user interface or data mapping.

1.5.4 Data Collection and Expertise Limitation

In a process of preparing and collecting data, the data must always be kept updated in the databases. The source data might be distributed across multiple legacy systems and it can exist in inconsistent or incomplete formats due to data entry problems. In UNIMAS, student data of Undergraduates and Postgraduates are extracted from different databases using separate SQL statements. This data is prepared separately as they are handled by separate entities. The output of data extraction processes are saved
into separate text file format. After that, these text files will be sent to Database Administrator (DBA) who responsible for handling data submission to MyMoheS and managing databases in UNIMAS.

Data errors as mentioned earlier in Section 1.5.1 can only be seen upon data transmission to MyMoheS. If the data consist of missing values or invalid values, it will fail in the transmission process. The System Owners need to make the data correction by entering a valid value to replace invalid data. These processes of extraction, consolidation, transmission and correction will be repeated until the data transmission is successful. This may take several days to complete. The limitation of expertise involved in these processes also given an impact to the data preparation process. Most System Developers are not only focused on the data preparation process but are also involved in other in-house system developments. Therefore, more time and effort are needed by the System Developer to complete the tasks.

1.6 Scope of work

In particular, this study provides an opportunity to map current business problems within the UNIMAS database, to allow a better understanding of the data and attempt to clarify and improve current data quality. The study undertaken attempts to synthesize on existing definitions and experiences from information management and quality assurance.

This dissertation concerns the issues that define data quality on student data and the mechanisms that can be developed to achieve and sustain good quality of the data for
UNIMAS. Thus, the work focuses on the solution of preparing high quality data for MyMoheS submission and improves the current student data preparation process.

1.7 Chapter summary

Good quality data is the data that meets the requirements of data users to support service delivery, quality improvement, performance reporting and planning. MyMoheS has played a major role in integrating and consolidating data from all IHL. The success of MyMoheS depends on the accuracy of coding and data entry, the timeliness of data, the comprehensiveness of data collection and the degree of which all relevant records are captured.

In the next chapter, further discussion and focus will be on data quality methodology, techniques and comparison of two top leaders in data quality tools development which can make data quality improvement process much easier.
Chapter 2: Literature Review

2.1 Introduction

Nowadays, data quality has emerged as a discipline through specific on-going research programs in various areas. In practice, data quality is a concern for professionals involved with a wide range of information systems ranging from data warehousing and business intelligence, to customer relationship management and supply chain management.

A review of the data quality literature by Wang, Storey, & Firth (1995) summarized that the research efforts are focused on (i) operation and assurance costs, (ii) research and development, and (iii) the production of data products. The Massachusetts Institute of Technology (MIT) program led by Professor Richard Wang has initiated Total Quality Management (TQM) programs which produced large number of publications and outlined the systematic prevention of errors through the adoption of total quality management principles and many of current research utilizes these principles.

To further discuss on data quality, this chapter provides the literature review of data quality framework. Section 2.3 will describe a comparison between two top leaders of data quality tools.