



School of Computing

A RULE BASED APPROACH TO DATA CERTIFICATION

- APPLYING DQXML FOR SYSTEM INDEPENDENT DATA CERTIFICATION -

Fakir Hossain

A dissertation in fulfilment of the requirements for the award of MSc

Supervisor: Dr. Markus Helfert

Submission Date: 9 September 2013

Declaration

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of MSc is entirely my own work, that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge breach any law of copyright, and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.



Fakir Hossain

September 2013

Candidate Number: 55228956

Abstract

Many researchers and practitioners have been attracted to improve data quality due to its monumental importance as a key success factor. Mathematical and statistical models have been deployed to information systems to introduce constrain and transaction based mechanisms to prevent data quality related problems. Entire management of the process and roles involved in data generation has also been scrutinized. Vast amount of knowledge base progressed in this area are mostly limited from practical perspective. Quality related meta data is absent from most information systems. Neither process mapping nor data modelling provides sufficient provision to measure quality or certification of data in the information systems. Furthermore, on-going monitoring of data for quality conformance through a separate process is expensive and time consuming. Recognising this limitation and aiming to provide a practical-orient comprehensive approach, I propose a process centric quality focused solution incorporating data product quality, conformance monitoring and certification. I base my work on DQXML developed by Ismael Caballero and deploy rigour of design science to construct InfoGuard. InfoGuard consists of DQXML incorporating quality meta data and an independent data quality monitor that provides certification of data through a rule based process centric framework for on-going data quality monitoring.

Table of Contents

1. Introduction	- 1 -
1.1 Research Objective	- 1 -
1.2 Research Questions	- 2 -
1.3 Motivation.....	- 2 -
1.4 Relevance.....	- 3 -
1.5 Thesis Structure	- 4 -
2 Overview of Data Quality and Related Work.....	- 5 -
2.1 Data Quality	- 5 -
2.2 Measuring the Quality of Data.....	- 7 -
2.3 Information Manufacturing System	- 13 -
2.4 Total Data Quality Management (TDQM)	- 14 -
2.5 Overview of Related Work	- 15 -
2.6 Defining Data Certification.....	- 17 -
3 Research Methodology	- 18 -
3.1 Methodology Requirement	- 18 -
3.2 Selecting a Methodology	- 19 -
3.3 Design Science Research Methodology.....	- 20 -
3.4 Action Research.....	- 22 -
3.5 Action Design Research.....	- 24 -
3.6 Organisational Context	- 27 -

4	Design and Construction of the Artefact.....	- 28 -
4.1	Root Cause of Data Quality Problems	- 28 -
4.2	Data Quality Blocks.....	- 31 -
4.3	Extending IMS Design – Introducing Global DQ Blocks	- 32 -
4.4	Designing Data Quality Certificate.....	- 37 -
5	Evaluation.....	- 39 -
5.1	Evaluation Strategy	- 39 -
5.1.1	What is Evaluated	- 39 -
5.1.2	How is it Evaluated.....	- 40 -
5.1.3	When Evaluation Takes Place.....	- 41 -
5.2	Evaluation Method.....	- 41 -
5.3	Evaluation Scenarios.....	- 42 -
5.3.1	Scenario 1: Up to Date Student Address.....	- 42 -
5.3.2	Scenario 2: Missing Student Email Addresses.....	- 43 -
5.3.3	Scenario 3: Incorrect Booking Dates	- 44 -
5.3.4	Scenario 4: Missing Booking Notification to Hotel.....	- 46 -
5.4	Implementing InfoGuard.....	- 46 -
5.4.1	Scenario 1: Up to Date Student Address.....	- 46 -
5.4.2	Scenario 2: Missing Student Email Addresses.....	- 48 -
5.4.3	Scenario 3: Incorrect Booking Dates	- 50 -
5.4.4	Scenario 4: Missing Booking Notification to Hotel.....	- 51 -
5.5	Findings.....	- 53 -
5.5.1	Scenario 1: Up to Date Student Address.....	- 53 -

5.5.2	Scenario 2: Missing Student Email Addresses.....	- 56 -
5.5.3	Scenario 3: Incorrect Booking Dates	- 58 -
5.5.4	Scenario 4: Missing Booking Notification to Hotel.....	- 61 -
5.6	Data Quality Certificate	- 64 -
6	Summary and Conclusion.....	- 65 -
6.1	Summary	- 65 -
6.2	Comparison with other DQ Management Strategy	- 67 -
6.3	Limitation and Future Work	- 68 -
	List of References.....	- 69 -

List of Tables

Table 1: DQ Dimensions (Wang & Strong, 1996).....	- 7 -
Table 2: Elements of DQ Certificate.....	- 38 -
Table 3: InfoGuard Evaluation	- 68 -

List of Figures

Figure 1: Total Data Quality Management (Wang, 1998)	- 14 -
Figure 2: Design Science Methodology (Takeda, et al., 1990) (Vaishnavi & Kuechler, 2008) (Henver, et al., 2004).	- 20 -
Figure 3: Design Research Steps (Henver, et al., 2004) (Peffer, et al., 2007).....	- 21 -
Figure 4: Action Design Research (Sein, et al., 2011).....	- 25 -
Figure 5: Cause of DQ problems	- 29 -
Figure 6: IMS Adoption to Business Changes	- 30 -
Figure 7: IMS Design Incorporating Global IQ Blocks	- 33 -
Figure 8: DQXML for Booking Reservation Date Condition	- 34 -
Figure 9: DQXML for Booking Email Business Rules	- 36 -
Figure 10: DQ Monitoring and Certification Model.....	- 37 -
Figure 11: Notice for Unsatisfactory Attendance	- 43 -
Figure 12: Process for Posting Grades	- 44 -
Figure 13: Process for Making a Booking	- 46 -
Figure 14: Quality Rules for Scenario 1	- 47 -
Figure 15: Quality Rules Settings in InfoGuard for Scenario 1	- 48 -
Figure 16: Quality Rules for Scenario 2	- 49 -
Figure 17: Quality Rules Settings in InfoGuard for Scenario 2.....	- 50 -
Figure 18: Quality Rules Settings in InfoGuard for Scenario 3.....	- 51 -
Figure 19: Quality Rules Settings in InfoGuard for Scenario 4.....	- 52 -
Figure 20: Ability to Detect DQ Problems for Scenario 1	- 53 -
Figure 21: DQ Improvement over Time for Scenario 1.....	- 55 -
Figure 22: Ability to Reduce Lead Time for Scenario 1	- 55 -

Figure 23: Ability to Detect DQ Problems for Scenario 2.....	56 -
Figure 24: DQ Improvement over Time for Scenario 2.....	57 -
Figure 25: Ability to Reduce Lead Time for Scenario 2.....	58 -
Figure 26: Ability to Detect DQ Problems for Scenario 3.....	59 -
Figure 27: DQ Improvement over Time for Scenario 3.....	60 -
Figure 28: Ability to Reduce Lead Time for Scenario 3.....	60 -
Figure 29: Ability to Detect DQ Problems for Scenario 4.....	61 -
Figure 30: DQ Improvement over Time for Scenario 4.....	62 -
Figure 31: Ability to Reduce Lead Time for Scenario 4.....	63 -
Figure 32: Data Quality Certificate at the End of 12 Weeks	64 -

List of Abbreviations

DQ	Data Quality
DQXML	Data Quality Extensible Markup Language
DS	Design Science
IMS	Information Manufacturing System
IP	Information Product
IQ	Information Quality
IS	Information System
IT	Information Technology
TDQM	Total Data Quality Management
XML	Extensible Markup Language

1. Introduction

Quality data has become increasingly one of the main factors in securing organisational success and business performance (Ge, et al., 2011). Recognizing the importance of data quality (DQ), practitioners and researchers have for many years considered ways to improve DQ. Researchers have worked on mathematical and statistical models to prevent DQ problems. One of the most prominent approaches is database constraints and business rules (Keeton, et al., 2009). Moreover, the management issues related to the process of data generation has attracted many researchers. These managerial oriented approaches are mostly following concepts of quality management (Ge, et al., 2011). However, DQ problems still remain persistent (Blake & Mangiameli, 2011). Increased DQ problems in general places doubt in the confidence of the data consumer. Lack of data certification often prevents user from making a decision due to lack of confidence or causes inappropriate decision based on misperception of the accuracy of data. This research takes a rule based approach to construct an independent DQ monitor to achieve on-going DQ certification.

1.1 Research Objective

Despite many efforts at improving DQ, problems still persistent causing devastating impacts on organisations (Furber & Hepp, 2011) (Blake & Mangiameli, 2011) and individuals (Madnick, et al., 2009). DQ problems are quantified and improvement plans are made only after organisations have been adversely affected (Caballero, et al., 2006). Most of the DQ improvement methodologies are designed to operate at a stage when DQ problems are already discovered after severe consequences. However, to minimize impact of DQ problems, approaches need to be adopted to detect DQ problems at earliest possible time. In order to address this problem, on-going data certification is necessary to flag DQ problems at

an early stage. Objective of this research is not to add to the number of approaches that have been introduced to improve the level of DQ. This research aims to deal with lack of availability of the DQ level to the data consumer by providing on-going DQ certification. There is also a need for the data certification to be system independent for it to be credible. Data certification approach must also be practical and adoptable. Another objective of this research is to see if on-going DQ certification can detect IQ problems and reduce lead time for DQ problem detection. I will also observe if DQ certification has any impact on DQ level.

1.2 Research Questions

Based on the above identified problem, the following questions have been adopted for further exploration. The Total Data Quality Management (TDQM) cycle, described further in section 2.3, was adopted to examine various issues in relation to this research (Madnick, et al., 2009) (Ge, 2009).

What is data certification? (Define)

Can Data Quality Extensible Markup Language (DQXML) be used to certify data? (Measure)

Is data certification effective on an on-going basis? (Analyse)

Can data certification reduce DQ problem detection lead time and improve DQ? (Improve)

1.3 Motivation

Due to wider adoption of Information Technology (IT), DQ problems are on the rise costing trillions in revenue (Furber & Hepp, 2011). All types of organisations (Klein, et al., 1997) are affected by poor DQ. For example manufacturing (Redman, 1996) (Wang, et al., 2001),

banks (Wang & Strong, 1996) and public sector (Strong, et al., 1997) are to name a few. Due to the importance of DQ, many researchers have proposed various approaches to deal with DQ (Furber & Hepp, 2011). Despite two decades of research in DQ, DQ problems remain persistent (Blake & Mangiameli, 2011). Typically an organisation realizes DQ problems after they had been adversely affected and only then DQ problems are detected, quantified and improvement plans can be made (Caballero, et al., 2006). Most of the research in DQ area focuses on either assessment techniques for DQ or ways to improve DQ by preventing the problems occurring in the first place. However, as DQ problems occur (which continues to be the case) lack of on-going data certification leaves DQ problems undetected. In this area, research focuses on detecting certain types of problem such as duplicated records (Hernandez, 1998), record linkage (Fellegi I, 1969) or cleansing tasks (Galahards, et al., 2001) (Herbert, et al., 2004). Some other works related to effectively monitoring database provides for effective computing techniques (Buneman & Clemons, 1979) (Huang, et al., 2010).

One of the shortfalls of current approaches is that they mainly focus on examining the end product of an information system (IS). However, data is created and manipulated through the various steps along the IS. In my research, I examine the data flow, construct DQ rules and relate it to the business process that creates this data. By this view I am able to certify data in the context of the process that generates the data.

1.4 Relevance

The cost of poor DQ, in a commercial context, is not only the cost of the direct loss resulting from it but also the cost of correcting the DQ problems. Indeed, poor DQ can impact business at an operational, decisional and strategic level (Redman, 1996). Severe impact of poor DQ at personal, social and commercial context is well recognized (Madnick, et al.,

2009). Researchers have examined the impact of DQ in specific application areas of information manufacturing system (IMS) such as patient records (Mikkelsen & Aasly, 2005) and Enterprise Resource Management (Xu, et al., 2002). The impact of DQ on organisational performance and cost benefit analysis has attracted many researchers (Lee & Strong, 2004) (Sheng & Mykytyn, 2002). Redman examined the effect of poor DQ on organisational effectiveness (Redman, 1998). The impact of poor DQ on decision making process has been examined in various studies (Chengular-Smith, et al., 1999) (Fisher, et al., 2003) (Jung, et al., 2005) (Ragnathan, 1999) (Keeton, et al., 2009). Slone (Slone, 2006) has demonstrated the relationship between DQ and predictability of an organisational outcome. The impact of DQ has also shaped the way IT is managed within organisations where policy and process have to change to adapt to DQ challenges (Lee, 2004) (Lee, et al., 2006) .

1.5 Thesis Structure

The remaining chapter of this thesis is organized as follows. Chapter 2 outlines the concepts of DQ and describes DQ dimensions. It also discusses related work in this area and identifies a gap in the literature, which underpins the contribution of this research. Furthermore, this chapter provides the theoretical framework for the study. Chapter 3 describes the research methodology employed in this study. After setting the basics of methodology requirements, justification is provided for the selected methodology. This chapter concludes by setting the research design. In Chapter 4, I design and construct the solution. I set out the theoretical basis for the proposed solution and layout the detail architecture for the artefact. In chapter 5, I report the evaluation of the artefact. To start off, I set out the evaluation strategy. I then describe the evaluation scenarios and present the findings in the study. In the final chapter, chapter 6, I summarize and conclude on the findings. I also set out the limitation of the research and provide direction for further research in this area.

2 Overview of Data Quality and Related Work

This section provides an overview of DQ concepts. Some of the most widely used definitions of DQ are highlighted and DQ dimensions are also detailed in this section. An overview of the literature and related work in the area to date is also discussed.

2.1 Data Quality

Defining *exactly* what is meant by DQ has been the focus of much research in this field. A great deal of literature has taken a product approach toward DQ (Wang, et al., 1998). Similar to a product, DQ has been defined as “fit for use by the data consumer” (Wang & Strong, 1996). Khan has defined DQ as “Information that meets specification or requirements” (Khan, et al., 1998). English defined DQ as information that “Consistently meets end user’s expectation” (English, 1999). Redman defined DQ as information that is “Fit for use, free of error, meet desired features” (Redman, 2001). DQ is “a measure of how fit information is for a purpose” (Keeton, et al., 2009). DQ often refers to technical issues while Information Quality (IQ) refers to nontechnical issues (Madnick, et al., 2009). However, as adopted in many DQ literatures, in this report, I will not make a distinction between DQ and IQ and use DQ to refer to the full range of issue, both technical and non-technical (Ballou, et al., 1998) (Madnick, et al., 2009) (Ge, 2009). In this research, I adopt a view that user perspectives are very important in defining DQ.

Taking it further from the definition, traditionally DQ has been broken down into various quality dimensions that represent a single aspect of the quality. Various approaches taken into defining the quality dimensions have been categorized into the following three approaches by Wang & Strong (Wang & Strong, 1996): Intuitive, Theoretical and Empirical.

An intuitive approach undertaken by Ballou (Ballou & Pazer, 1985) considered the following four data quality dimensions: Accuracy, Timeliness, Completeness and Consistency. In theoretical approach, theory or models are proposed in order to derive and justify DQ dimensions. DQ dimensions were derived observing the deficiencies between the view presented by the information management system and the view obtained by observing the real world (Wand & Wang, 1996). After analysing these deficiencies, four intrinsic (system oriented) dimensions were derived: Completeness, Unambiguous, Meaningful, and Correct. In empirical approach, users are consulted to determine the quality dimensions that are important to them when assessing information quality.

Lee (Lee, et al., 2002) provided another categorization for DQ dimensions contributions. Lee categorized them in empirical where user or consumer perspective was taken into account in deriving DQ dimensions (Wang & Strong, 1996) (Zmud, 1978). Lee's second category of authors based their dimensions based on literature review (DeLone & McLean, 1992) (Goodhue, 1995) (Jarke & Vassiliou, 1997). Another set of researchers focused only on dimension being able to measure objectively (Ballou & Pazer, 1985) (Wand & Wang, 1996).

Most approaches addressing DQ tend to narrowly focus just on "Accuracy" (Wang & Strong, 1996). However, at this stage, nearly 200 dimensions of DQ have been identified and there is disagreement in their nature (are these concepts, goals or criteria?), their definitions or measures (Wang, et al., 1993) (Lee, et al., 2002) (Wang & Strong, 1996) (Zmud, 1978) (DeLone & McLean, 1992) (Goodhue, 1995) (Jarke & Vassiliou, 1997) (Ballou & Pazer, 1985) (Wand & Wang, 1996). Many studies have confirmed that DQ is a multi-dimensional concept (Pipino, et al., 2002). Despite the multidimensional nature of DQ, it is nevertheless a single phenomenon. DQ dimensions are inherently dependent on each other (Lee, et al., 2002). For example, to get more accurate information, more time might be required. Accessibility and security are also dependent on each other.

Intrinsic dimensions	Contextual Dimensions	Representational Dimensions	Accessibility Dimensions
<ul style="list-style-type: none"> • Believability • Accuracy • Objectivity • Reputation 	<ul style="list-style-type: none"> • Value-added • Relevance • Completeness • Timeliness • Appropriate amount of data 	<ul style="list-style-type: none"> • Interpretability • Ease of understanding • Representational • Consistency • Concise representation 	<ul style="list-style-type: none"> • Accessibility • Access security

Table 1: DQ Dimensions (Wang & Strong, 1996)

The most adopted study in the DQ area was undertaken by Wang and Strong (Wang & Strong, 1996) where data consumers were involved in every phase of the study. This approach has been adopted in much of the recent research (Ballou, et al., 1998) (Batini, et al., 2009) (Caballero, et al., 2006) (Ge, 2009) (Lee, et al., 2002) and is also the standard benchmark for DQ dimensions for this current research. These dimensions are summarized in Table 1.

2.2 Measuring the Quality of Data

Reasonable means must be available to measure various aspects of DQ dimensions to be able to assess and certify quality of information management systems. As some of the dimensions are user dependent, for instance believability, it is important to take both subjective and objective approach, as adopted in (Pipino, et al., 2002). For the objective measurements,

three functional forms, simple ration, min or max operation and weighted average were used. Some of the measuring mechanism of DQ dimensions is given below.

Accuracy

Accuracy or free-of-error represents how much of the data does not contain error. This measure is conducted by simple ration between data containing error and total number of data. As the convention is to represent the measurement between 0 and 1, where 1 is totally accurate and 0 is totally inaccurate, the accuracy dimension is measured by using the following formula (Pipino, et al., 2002).

$$Accuracy = 1 - \frac{\text{Number of data containing error}}{\text{Total Number of data}}$$

Completeness

Generally there are three types of completeness issues addressed in literatures. They are schema completeness, column completeness, and population completeness (Pipino, et al., 2002). Schema completeness relates to whether all relevant attributes of an entity is present in the schema. Column completeness is to do with when records are missing values in various columns. Finally population completeness is to do with when records themselves are missing. In either case, completeness can be presented as the ration between incomplete items and total number of items.

$$Completeness = 1 - \frac{\text{Number of incomplete items}}{\text{Total Number of items}}$$

Consistence

There are various types of consistency issues with data quality. Consistency might be within values of a specific entity (for example, Co is written as Co., Co or even spelled out as County). Consistency also relates to other entities where if not complied with could result in redundant records. Consistence presentation of data is also of significant importance. In either case measuring consistency could be represented in the following equation (Pipino, et al., 2002).

$$Consistency = 1 - \frac{\text{Nuner of violation of specific Consistency Type}}{\text{Total Number of Consistency Check}}$$

Believability

Believability dimension is defined to be the extent to which data is regarded as true and credible (Pipino, et al., 2002). This dimension is measured by comparing assessment of creditability by the user, commonly accepted standard and user's previous experience. To be conservative, min of all three factors are considered to the measurement of Believability.

$$Believability = \min (\text{credibility by user, commonly accepted standard, user's previous experience})$$

Objectivity

Objectivity is the extent to which data is impartial and objective (Wang & Strong, 1996). This dimension could also be measured by using similar equation to believability dimension. This could be done by comparing assessment of objectivity by the user, commonly accepted standard and user's previous experience. To be conservative, min of all three factors are considered to the measurement of objectivity.

$$\textit{Objectivity} = \min (\textit{credibility by user, commonly accepted standard, user's previous experience})$$

Appropriate Amount of Data

This dimension is defined as “data quantity being neither too much not too little” (Pipino, et al., 2002). The measurement takes into consideration amount of data units needed and amount of data units provided.

Appropriate Amount of Data

$$= \min \left(1 - \frac{\textit{Data Unit Needed}}{\textit{Data Unit Provided}}, 1 - \frac{\textit{Data Unit Provided}}{\textit{Data Unit Needed}} \right)$$

Timeliness

Timeliness dimension refers how up-to-date data is with respect to the task it is used for (Pipino, et al., 2002). Primarily two factors, currency and volatility govern the timeliness dimension (Ballou, et al., 1998) . In the context of information manufacturing system, currency refers to the age of data units used for producing information. This age is dependent on three factors. This depends on when the information is delivered to the customer, the time the data was obtained to produce the information and age of the data when received (Ballou, et al., 1998).

$$Currency = (\text{Delivery Time} - \text{Input Time}) + \text{Age}$$

Volatility is the shelf life of the information, i.e., how long the information remains valid. This is quite context dependent as some information might remain valid for ever (e.g. someone's date of birth) and others might be too old within minutes (e.g. stock cotes). Considering both factors above, the following formula was suggested (Ballou, et al., 1998).

$$Timeliness = \left\{ \max \left(1 - \frac{\text{currency}}{\text{Volatility}}, 0 \right) \right\}^s$$

Where

s is the sensitivity factor applied by the user.

Accessibility

Accessibility is measuring how easy it is for a user to attain required information. Accessibility could have various aspects to it when it comes to measurement. For instance, to at the timely aspect of user's easiness, following metric was suggested in (Pipino, et al., 2002).

$$Accessibility = \left\{ \max \left(1 - \frac{t_1}{t_2}, 0 \right) \right\}^s$$

Where

t_1 is the time interval from request by the user to delivery to the user.

t_2 is the time interval from request by the user to the point at which data is no longer useful.

s is the sensitivity factor applied by the user.

One of the important things to note is that quality dimensions are related to one another. For example, in order to make data more accurate, more time would be required. Therefore, emphasis on accuracy will have a negative impact on timeliness. Therefore, data quality cannot be measured for each dimension without evaluation dimensions in quality matrix, taking into account in other related dimensions, to achieve best overall quality of data. Such comprehensive models for data quality matrix have been suggested in various literatures (Pipino, et al., 2002) (Ballou & Pazer, 1985) (Ballou, et al., 1998). In fact most of the measuring equation has been developed in the context of analysing such data quality matrixes.

2.3 Information Manufacturing System

In contrast to traditional approaches, DQ researchers proposed a novel perspective on IS and regarded IS as IMS (Wang, 1998). Wang suggested an analogy between product manufacturing systems to information manufacturing. In this thesis, I define IMS as a system that operates on raw data as the input to create information products (IP) as the output.

In a typical IMS, there are three components at the design layer of IMS: Process Modelling, Data Modelling and Business Rules Design (Kovacic, 2004). As requirements to construct an IMS are gathered, a set of business rules and business process are documented either in natural languages or graphic based notations for ease of understanding. There are a number of well-established graphic based conventions to document processes enabling limited automation from design to implementation of IMS. A data model is then constructed to support the required data for the IMS based on the process and business rules.

At the implementation level, one or more software applications are built to implement the processes and business rules. One or more physical databases are constructed to support the data model from the design layer. Applications and databases work together to comprise the IMS.

2.4 Total Data Quality Management (TDQM)

Adopting a product perspective on information permitted the use of many of the concepts and processes that were well established in physical product manufacturing domain (Ballou, et al., 1998). Specifically: Plan, Do, Check and Act are the key component of the widely accepted Deming Cycle for product quality enhancement (Wang, 1998). Based on this, the TDQM cycle, illustrated in Figure 1, was proposed which is iterative of nature. This consists of four components: Define, Measure, Analyse and Improve (Ballou, et al., 1998).

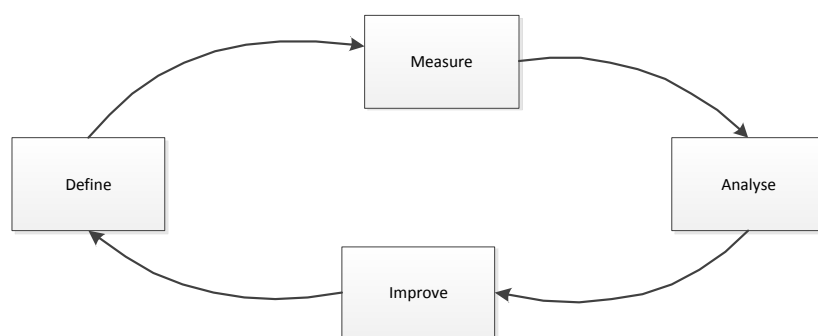


Figure 1: Total Data Quality Management (Wang, 1998)

DQ dimensions are identified and the corresponding DQ requirements are described by the “Define” component. DQ matrixes are constructed by the “Measure” component. The root causes for the DQ problems are identified and the impact of the poor quality data is calculated by the “Analysis” component. Finally, techniques for improving DQ are provided by the “Improve” component. These are then applied in line with the DQ requirements set out at the “Define” component (Wang, 1998).

2.5 Overview of Related Work

Some of the earliest work, by mathematician and statisticians, to address DQ problem focused at the database layer of IMS. Researchers have looked into dealing with DQ problems arising out of data integration for warehousing purpose (Fan, et al., 2001) (Goh, et al., 1999) (Madnick & Zhu, 2006), decision analysis (Aggarwal & Yu, 2009) (Keeton, et al., 2009) and schema matching (Rahm & Bernstein, 2001) (Doan & Halevy, 2005). The Entity-Relationship Model has been extended to include DQ related characteristics (Wang, et al., 1993) (Storey & Wang, 1998).

The database layer has been improved to prevent data inconsistency and corruption by introducing data constraints (static and dynamic), transaction management and other measures (Brock, 2000) (McCune & Henschen, 1989). Database constraints were introduced to deal with data arriving at its door step without fulfilling the required quality criteria (Decker, 2009). However, most database management systems are only able to deal with static constraints (McCune & Henschen, 1989) (Brock, 2000) (Vianu, 1983). This ability is further limited when multiple databases are involved. Quality tolerance implantation is also very limited via database constraints. For example, if an email is required or not required for all students, it is easy to implement. However, if I was to implement that 80% of the total student records should (not must) have email address recorded in IS, it is difficult to implement via database constraints.

From the business and management perspective, the primary focus has been to model DQ assessment and processes with the intent to increase the DQ of stored data. A great deal of emphasis is placed on processes that involve people and qualifying data as it is captured (Keeton, et al., 2009) (Wang, 1998) (Ballou, et al., 1998) (Shankaranarayanan, et al., 2000).

Successful DQ measurement is one of the key requirements for any DQ management (Batini, et al., 2009). However, there is enormous disagreement and lack of unification as to deriving an agreed protocol to achieve this goal (Caballero, et al., 2006). As some of the dimensions are user dependent, for instance believability, it is important to take both subjective and objective approach, as adopted in (Pipino, et al., 2002). For the objective measurements, three functional forms, simple ratio, min or max operation and weighted average were used. Quality dimensions are related to one another. For example, in order to make data more accurate, more time is required. Therefore, emphasis on accuracy will have a negative impact on timeliness. Therefore, DQ cannot be measured for each dimension in isolation. DQ evaluation requires a quality evaluation matrix, taking into account other related dimensions, to achieve best overall quality of data. Such models for DQ matrix have been suggested in various literatures (Pipino, et al., 2002) (Ballou & Pazer, 1985) (Ballou, et al., 1998). In fact most of the measuring equation has been developed in the context of analysing such DQ matrixes.

The area of DQ monitoring focuses on detecting certain types of problem such as duplicated records (Hernandez, 1998) or cleansing tasks (Galahards, et al., 2001) (Herbert, et al., 2004). Some other works related to effectively monitoring database provides for effective computing technics (Buneman & Clemons, 1979) (Huang, et al., 2010). However, on-going DQ certification can detect DQ problem to minimize the impact.

One of the pioneer works in data certification focused on designing a brokerage to support quality aware data exchange in the context of a co-operative system (Scannapieco, et al., 2004). However, to the best of my knowledge, the model has not yet been demonstrated in the context of a real use case. Efforts have also been made to develop models that can provide quality meta data along with query results. (Missier, et al., 2006). However the model heavily relies on the data provider to associate metadata to the quality matrixes.

Furthermore incomplete metadata can produce inaccurate indication as to the quality provided along with data. Development and population of the quality matrixes can be quite expensive with limited reusability. This research aims to provide an architect for a data quality certification solution that is system independent, credible, practical, adoptable and provides on-going data certification to the data consumer. The solution further aims to detect DQ problem and reduce DQ problem detection lead time. I will also observe the impact on DQ of on-going data certification.

2.6 Defining Data Certification

There has been various definition explored throughout DQ research about data certification. Data certification has been defined as applying a structured quality process to ensure that data meets or exceeds the standards established by its intended consumers (Brunson, 2005). Data certification has also been broken down into the context of data life cycle. For instance, to certify an application input data means to evaluate the main characteristics of data, such as: completeness, correctitude, credibility, accuracy, security of access and relevance. On the other hand, certifying the results of a process involves verifying supplementary correlations compared to the characteristics of initial data, the existence of a defined structure, and the correctitude of processing in all used cases. Usefulness and importance of data has to be considered when the process of data certification is designed (Ivan, et al., 2000). However, in the context of the scope of this research, I only look into data already existing in the system. For the purpose of this research, I adopt a working definition for data certification. I define data certification as follows:

“Data certification is a compliance report on various dimensions of data quality tolerance as defined by the source system rules.”

3 Research Methodology

This section provides rationale for the research methodology deemed suitable for this research.

I conclude this section by presenting my research design.

3.1 Methodology Requirement

Selecting an appropriate research methodology is crucial to address a research question in a consistent and rigorous manner. It is a process that must be careful, systematic to establish facts or principles (Kumar, 1996). In selecting the research methodology, consideration was made as to whether it was the most suited to the purpose of study, if the research framework was sufficiently flexible, if it was supportive of every aspect of the questions posed in the research and if it was of high professional standard.

Socio technique nature of IS systems requires researchers to investigate various aspect of the environment where IS system operates (Herver, et al., 2004). Researchers in IS environment faces far more complex situation than a natural scientist as theory and practice cannot be separated. Steady interaction between practice and theory is essential for the research process (Checkland, 1985). In this research, IS users are interacted regularly to design on-going data certification. Subjective nature of DQ is fully reliant on IS users. Hence this research required selecting a methodology that caters for integrated and flexible philosophy of inquiry.

The core objective of this research is to develop an independent data certifier. This required an iterative approach to refine the tool after each outcome evaluation. This aspect of the research influenced selecting an appropriate methodology where iterative refinement was a core part of the methodology (Barab & Squire, 2004). Finally, the selected methodology had

to be of high professional credibility. Any limitation within the methodology also had to be addressed (Collins & Hussey, 2002).

3.2 Selecting a Methodology

Methodology is the process of deriving a scientific outcome, not the outcome itself (Manion, et al., 2000). It is far more than a collection of methods; it is the rational and philosophical basis that underlines the study. Methods on the other hand are tools for data gathering for the purpose of forming basis for the scientific inquiry. Methods are tools for the study and the methodologies are the principals how such tools are interpreted (Collins & Hussey, 2002).

Traditionally research is broken into two main categories: qualitative and quantitative research. In quantitative research the purpose is to develop and evaluate causal theories (Denzin & Lincoln, 2005) where qualitative research describes data and characteristics about the population or phenomenon being studied (Martyn, 2008).

In IS research, technological artefacts (computerised systems) are studied in a social (IS users, organisations) settings (Alavi & Carlson, 1992). Methods typically used in the field of engineering are most suitable to undertake IS research from technological perspective. On the other hand methods used in social science are most appropriate to study the non-technical aspect of IS research. Conducting the research from either social or technical perspective presents a choice of either qualitative or quantitative research perspective. This research requires observing current DQ theories, developing new theory, implementing the theory in practice and conducting observation of the effect. Both qualitative and quantitative approach on its own presented a challenge to adopt or enhance artefacts. Either or combined approach might be insufficient as part of the research aim is to develop a tool for certifying DQ rather than mere description or explanation about DQ. When it comes to human creation, designing

artefacts to attain a specific goal, design science research methods are more appropriate (March & Smith, 1995). My research key feature is to create an artefact in the form of data certification. Design science (DS) oriented methodology deemed to be suitable compared to others to address the research questions discussed in section 1.2.

3.3 Design Science Research Methodology

DS aims to extend human and organisation capacities by creating new and innovative artefacts. DS research in IS is comprised of three elements: principals of DS research, practice rules and process for doing and presenting the research (Peppers, et al., 2007). As shown in Figure 2, DS is applied to a specific problem within the existing body of knowledge in the current environment. A specific set of steps are carried out and output of each step is used in subsequent steps to ensure rigour within the research. Finally, the findings are fed back to the body of knowledge to solve the specific problem being addressed.

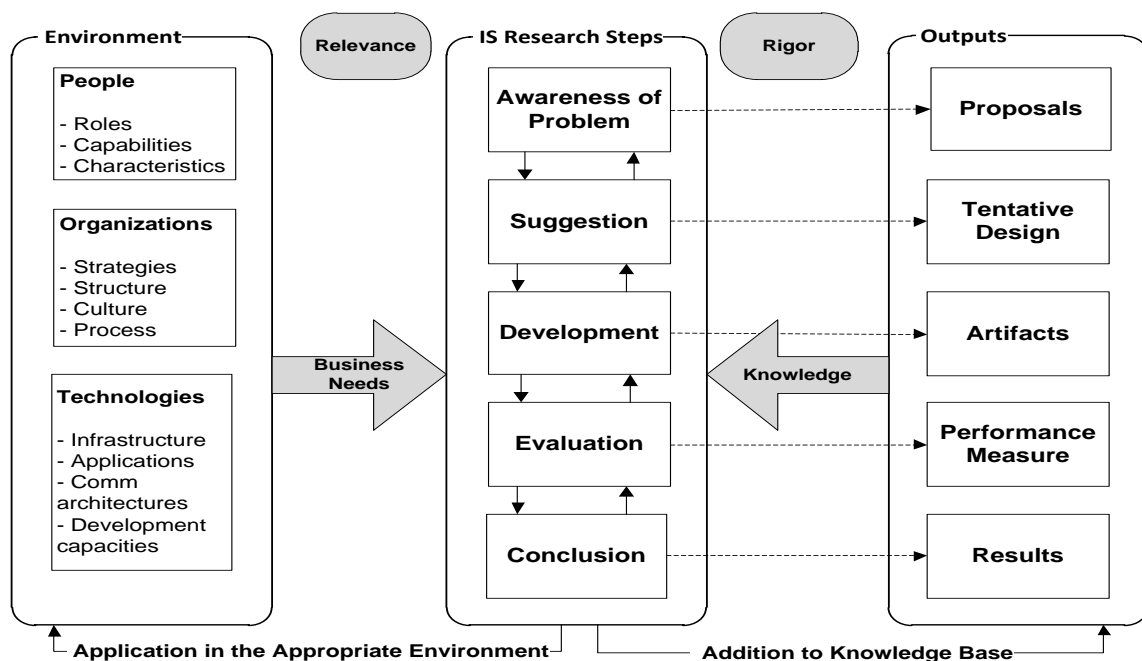


Figure 2: Design Science Methodology (Takeda, et al., 1990) (Vaishnavi & Kuechler, 2008) (Henver, et al., 2004).

Henver et al (Henver, et al., 2004) has provided justification and outlined a detail process steps for carrying out DS research. Peffers articulated six steps in conducting IS research (Peffers, et al., 2007). The main emphasis is provided on creation of artefacts to address a specific problem. Both design and evaluation of the artefact is to be conducted with verifiable rigour. The artefact must be based on prior theories and finally should be communicated to relevant audiences. These six steps in my research are summarized in Figure 3.

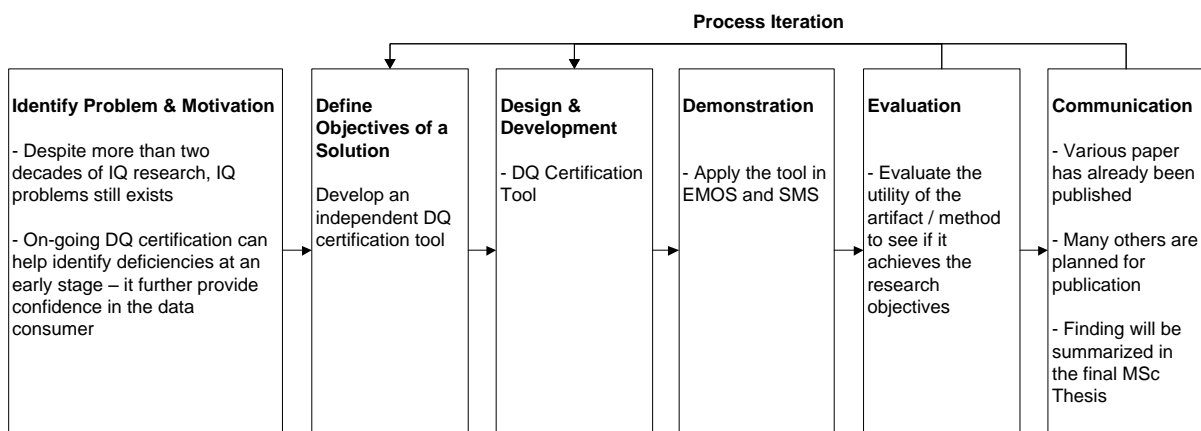


Figure 3: Design Research Steps (Henver, et al., 2004) (Peffers, et al., 2007)

Comprehending reality through explanation or predicting behaviour is primary objective of behavioural science. This is done through development and verification of theories. However, objective of DS is to design or change reality by artefact construction. DS is widely accepted in the engineering field and in recent years this approach has successfully been adopted by few IS researchers (Schelp & Winter, 2006). Henver (Henver, et al., 2004) also supports that DS is the most suitable research methodology for studying both social and technical nature of developing IS artefacts. In this research development of the artefact to provide on-going data certification is at the core. IS artefacts are product of DS which can be

implemented in organizations to solve specific problems and add to the knowledge base of DQ as a research field (Braun, et al., 2005) (Henver, et al., 2004).

However, in most DS efforts, identification of the problem comes before designing the artefact which is then followed by the evaluation. This is the “*build and then evaluate approach*” (Sein, et al., 2011). Designing the artefact goes beyond basing on technical knowledge alone. It requires interaction between various stakeholders even at a design phase. Sein argues that there is a need for a research method that explicitly recognizes artefact emerging from design, use and on-going refinement in the context and not designing in isolation. Action research (AR) caters for theory generation with researcher’s intervention to solve specific problem. It links theory with practice with doing typically in an interactive process based on working hypothesis refined over period through cyclical interactions. So, I turn my focus into AR to determine its suitability for my research.

3.4 Action Research

Action research is “learning by doing”. Gilmore (Gilmore, et al., 1986) defined action research as the following:

“Action research...aims to contribute both to the practical concerns of people in an immediate problematic situation and to further the goals of social science simultaneously. Thus, there is a dual commitment in action research to study a system and concurrently to collaborate with members of the system in changing it in what is together regarded as a desirable direction. Accomplishing this twin goal requires the active collaboration of researcher and client, and thus it stresses the importance of co-learning as a primary aspect of the research process”.

Winter (Winter, 1989) provides some key characteristics of action research. In an action research, researcher reflects upon own work through participating with others. All participants are co-researchers working collectively to reach a solution. There is always risk arising out of open discussion. This plural structure of action research requires that the final report is on-going collaboration rather than final conclusion. Finally, theory and practice are transforming each other in an iterative way.

Artefact to solve a specific problem is rarely empirically studied (Wynekoop & Russo, 1993). Case studies have been used for introduction of artefact use, change resistance, social defence and user's interest. Most of the approaches only address the construction phase of artefact. Several researchers have supported action research to be appropriate for artefact construction as it is iterative, it permits going into details and it permits long term observation (Juha-Pekk, 1998).

Action research is a variant of a case study and a field experiment (Galliers, 1992). Similar to case study, action research evaluates a state of particular circumstances. However, in action research a researcher participates and acts at the same time. This dual role permits researcher to improve the situation in the organization and to contribute to scientific knowledge by creating generalizable concepts and theories (Juha-Pekk, 1998).

Action research is very appropriate for this research due to its potential to evaluate and refine ideas, principles, techniques, and methods, as well as to address a specific real-world problem. Ability to understand the situation first-hand provides the advantage to action research (Baskerville & Wood-Harper, 1996). Susman (Susman, 1983) provided following five steps approach for carrying out an action research: Diagnosis, Action Planning, Action Taking, Evaluating and Specifying Learning.

3.5 Action Design Research

Generally DS and AR are treated as separate research methodology. There are number of researches where both DS and AR form part of the same research process. First, artefact is designed through DS and then evaluation of the artefact through implementation in an organisational context is done through AR (Iivar, 2007) (Foley, 2011). Sein argues for a new research method, Action Design Research (ADR) that recognises that artefact emerges from interaction with the organisation context (Sein, et al., 2011). He argues that ADR deals with two disparate issues: (1) address a specific problem in a specific setting by intervening and evaluating (2) construction and evaluation of the artefact address problems that are typified by the encountered situation. The ADR method contains steps and principals that address these issues. ADR method steps and principals are presented in Figure 4. I briefly explain each of the principal of ADR and report how I align each step of my research with the principals.

Principal 1: Problem Formulation: this principal views field problems as knowledge creation opportunity. ADR generates knowledge based on a specific problem that can be applied to a class of problems. This make the research activity problem inspired (Sein, et al., 2011). As described in chapter 1, this research is inspired by a specific problem in the organisational context where lack of information about DQ level feeds lack of confidence to data consuers. A solution is required for on-going DQ certification.

Principal 2: Theory-Ingrained Artefact: this principal emphasis that the artefact must be based on theories. Prior theories are used to structure the problems, identify possible solutions and to guide design. Artefact in my research is based on previous theory and

success in the DQ field. As described in Chapter 4, DQ measurement formula and evaluation matrix is based on existing theories. I also based the quality rule definitions based on DQXML (Caballero, et al., 2006).

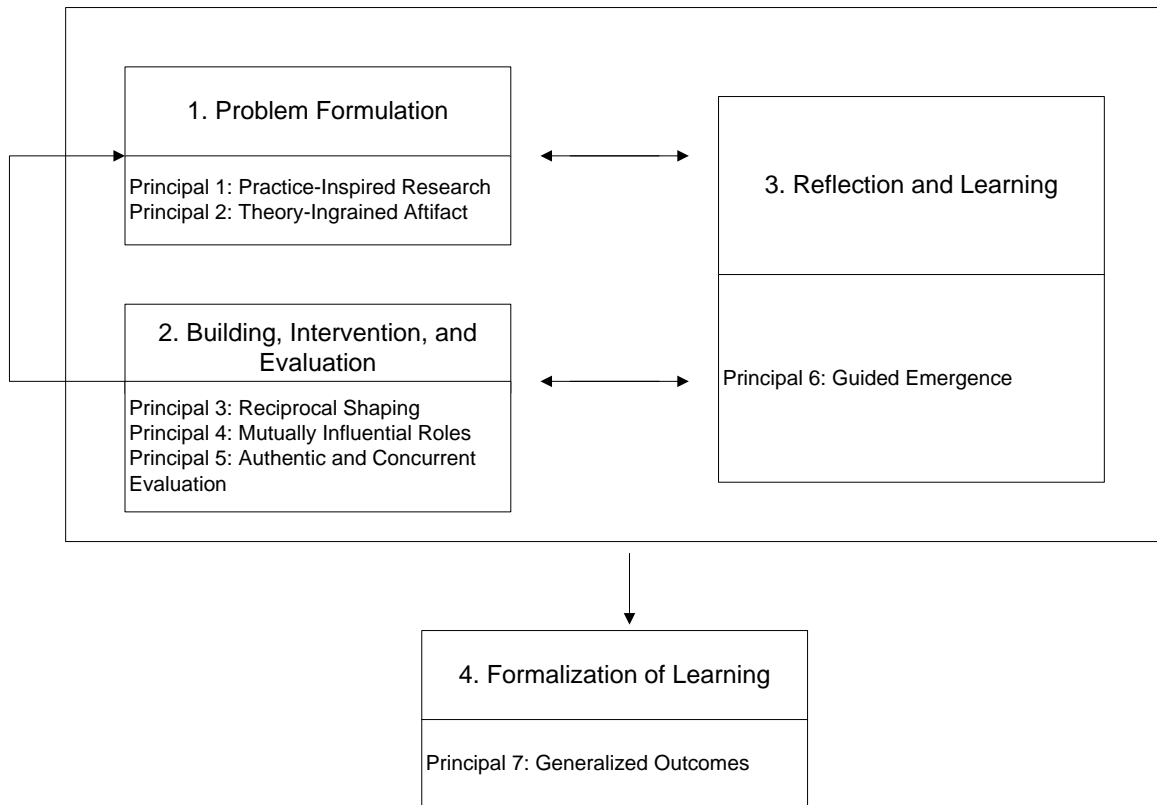


Figure 4: Action Design Research (Sein, et al., 2011)

Principal 3: Reciprocal Shaping: this principle emphasises the inseparable influence of the artefact and the organisational context. Researcher uses the design to interpret organisational context and organisational context influences the design of the artefact. In the case of this research, lack of DQ level information within the organisations helped identifying key rules for monitoring. Organisations also helped deciding on frequency of DQ monitoring for the purpose of certification. Artefact on the other hand, helped the organisation to put processes and resources in place to ensure that DQ compliance reports were acted upon so that a meaning evaluation of the artefact can take place.

Principal 4: Mutually Influential Roles: this principal recognises that the researcher brings to the table knowledge of theory and artefact and the organisation provides the knowledge of the work practice. These complementary roles enhance the research. In this research the organisations have been at the core of shaping the research problem and evaluation scenarios and my knowledge of theory in the DQ area has shaped the design and implementation of the artefact for on-going DQ certification.

Principal 5: Authentic and Concurrent Evaluation: evaluation is not separate step in ADR. Instead designing and reshaping occurs concurrent with evaluation. Earlier evaluations continue to inform later enhancement to the artefact. Number of issues identified in the evaluation of the earlier design of the artefact in this research lead to refinement of the design. For example, initial design called for continuous monitoring and identification of DQ problem. However, it was meaningless and the organisation resources were limited to act on it at such interval. So, weekly or nightly monitoring processes are implemented depending on the scenario.

Principal 6: Guidance Emerges: This principal captures the fact that the knowledge is formulated from on-going evaluation and reshaping of design based on the organisation context. In this research through evaluation, as described in Chapter 5, various findings in relation to research questions emerged.

Principal 7: Generalised outcome: ADR provides a solution to a problem. However, to generalise, a three step approach is adopted (1) generalise the problem, (2) generalise the solution and (3) derive design principals from the research outcome (Sein, et al., 2011). In this research, findings at evaluation stages are summarised and generalised in Chapter 6.

3.6 Organisational Context

In order to design and evaluate my approach, my model had to be designed and applied to enterprise level systems where data generation was linked business process. I was seeking systems where data certification in the context of the process would allow identifying and improving processes to reduce further DQ problems within the process and also subsequent processes from using poor quality data. I also needed the systems to have sufficient amount of data to be able to demonstrate my model. Two enterprise level applications were selected as candidates for certifying DQ.

SMS

SMS is an online student management system deployed in various private colleges in both the UK and Ireland. Eden Further Education based in Dublin is using SMS to manage its past and current 10000 students. There are large numbers of records that deal with various aspects of student management. Student personal details, course details, enrolment records, class attendance and participation, fees records are just to name a few. There are also processes involved in every step of the way for the records to be generated. This suited a perfect candidate for my study.

EMOS

EMOS is an online booking engine for hotel and holiday home accommodation providers. It is a central database facilitating various promoters of holiday breaks throughout Europe. In 2011, the number of holiday bookings made through EMOS exceeded 800,000. About 60% of these bookings were made online and the remainder through various call centres throughout Europe.

4 Design and Construction of the Artefact

In this chapter I provide the design and construction of the independent DQ monitor and the certification process.

4.1 Root Cause of Data Quality Problems

Investigating the root cause of certain types of DQ problems reveals why they are persistent. There are many types of DQ problem resulting from a range of issues. In this research, I look at the causes of DQ problems that arise out of system design, development, implementation and use of IMS. Figure 5 presents the layers that an IMS must go through from the initial conceptualization to the final implementation and use. There are many opportunities for inconsistency at each layer (Pham Thi & Helfert, 2007a).

At the user requirement gathering phase, user requirements can be misunderstood or incorrectly documented. These documentations may be inconsistent with the technical system specification derived from the user requirement. At the implementation layer, technical specification might be inconsistently implemented. Finally as various systems and users interact, there might be misunderstandings, miscommunication and inappropriate use of the system.

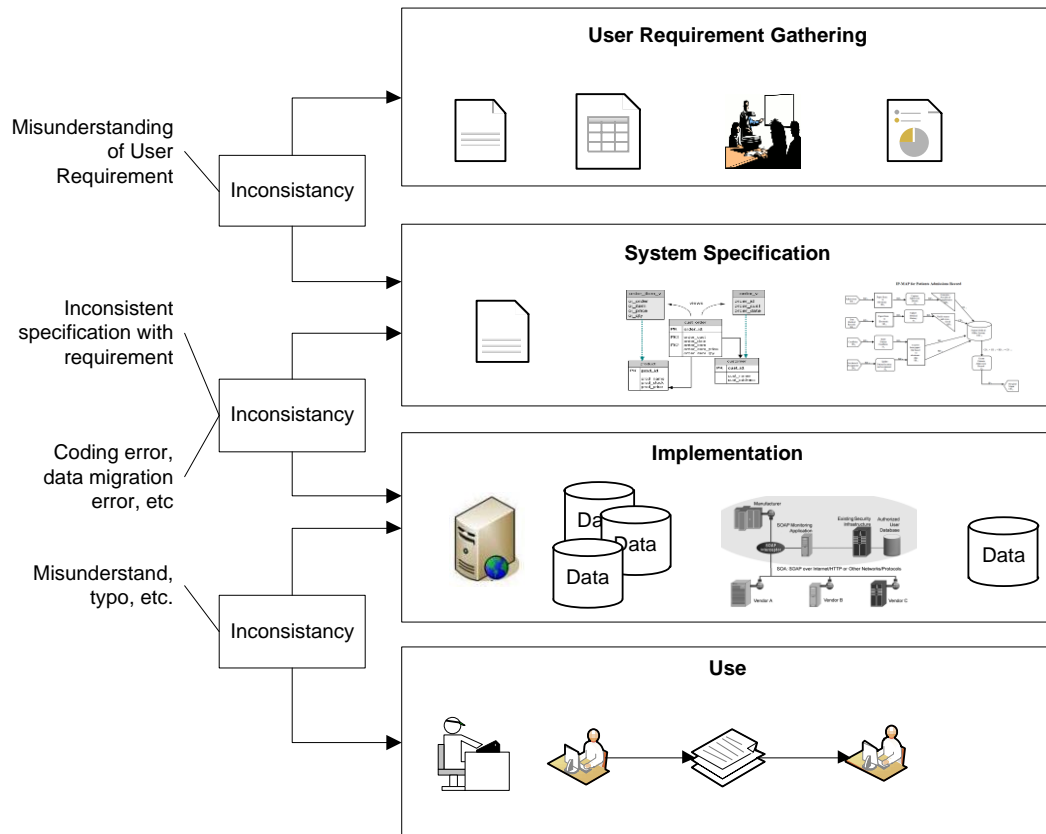


Figure 5: Cause of DQ problems

Due to the nature of the IMS, it often takes a considerable amount of time from the requirement gathering phase to implantation. The dynamic nature of modern business requires constant changes in business practices. Therefore, the system is often implemented, even if according to the design specification, inconsistently with business needs. This situation also arises if the business requirement change after the system has already been implemented (Pham Thi & Helfert, 2007b) as illustrated in Figure 6.

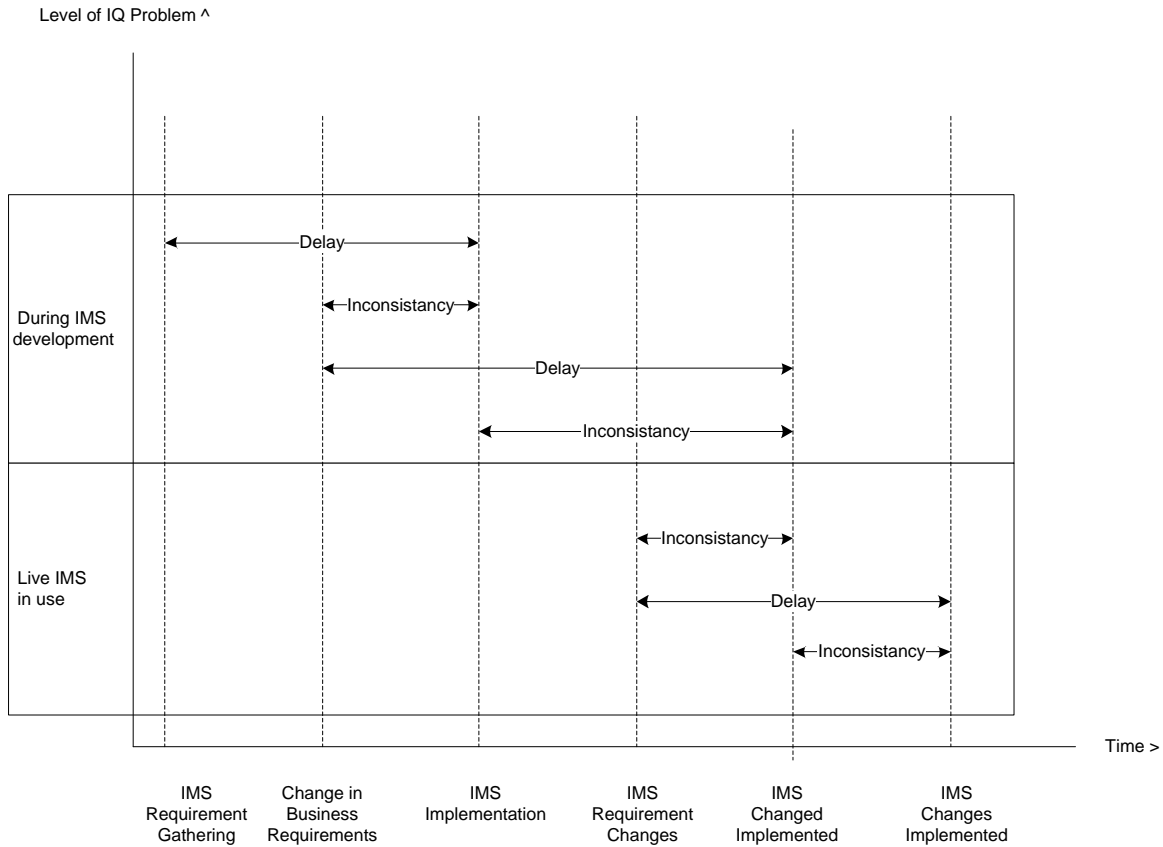


Figure 6: IMS Adoption to Business Changes

Inconsistency or error in any one or any combination of above will lead to obvious DQ problems. Some of these problems can be traced to root causes and eliminated over time. However, there are other types of causes of DQ problems that are unavoidable, for example, sudden change in a business requirement, change in personnel using the system, etc. Due to the nature of the IMS design, development, implementation and use, DQ problems are unlikely to ever be completely eliminated (Blake & Mangiameli, 2011).

4.2 Data Quality Blocks

In order to get the IMS developers to focus on other quality aspects of an IMS, researchers have proposed modelling IMS to be in various blocks. A dedicated block has been allocated to focus only on various quality aspects of data. Part of the IMS responsible for ensuring DQ is referred as “Quality Block” (Ballou, et al., 1998). This quality block is traditionally integrated with in the IMS. Despite best efforts from the IMS engineers and developers, IMS are often subject to errors. Therefore, being part of IMS, quality block itself may also contain errors. This can lead to poor DQ. To secure an independent certification of DQ, a system independent quality monitor is essential. However, to develop an IMS independent monitor will require additional time and cost which might make it prohibitive.

Another challenge within the quality block is that systems are often designed without sufficient model to define quality meta data to ensuring its on-going conformance and certification. This makes it difficult for the system to monitor and certify the DQ on an on-going basis. Of the three elements of IMS, namely, data, process and rules, last one is often neglected (Kovacic, 2004). Several approaches had been adopted to link data models and business processes (Nelson, et al., 2008) (Vasilecas & Smaizys, 2006) (Muehlen, et al., 2007) (Khan, et al., 2004). But most of the approaches have failed to provide an integrated approach for modelling quality business rules linking business process and data models. Hence quality rules are often defined in isolation from the underlying process it is trying to address or the data it is producing. This makes detecting IQ problem along with root cause of the problem with in IMS very difficult.

Motivated by these challenges, in this research, I suggested a model to configure quality rules in such a way that data in the IMS confirms with its quality requirements throughout its

lifecycle. I followed an approach proposed by (Ballou, et al., 1998) and described an approach to modelling DQ Blocks.

4.3 Extending IMS Design – Introducing Global DQ Blocks

Most of the process driven approaches suggest some form of improvement in the design of IMS (Pham & Helfart, 2007) (Ballou, et al., 1998) (Shankaranarayanan, et al., 2000). As discussed in the previous section, Ballou and Shankaranarayanan designed IMS in various blocks. DQ block generally work in the local context, i.e., works with in the context of the block previous to it and next to it. I propose further breaking down the DQ block into local and global DQ blocks. In the context they are currently used, these are local DQ blocks. However, there is a need to have global quality blocks that hold true independent of the steps with in IMS. In the system design context, I propose extending IMS design to include the global quality block to sit alongside traditional system design, as presented in Figure 7 below.

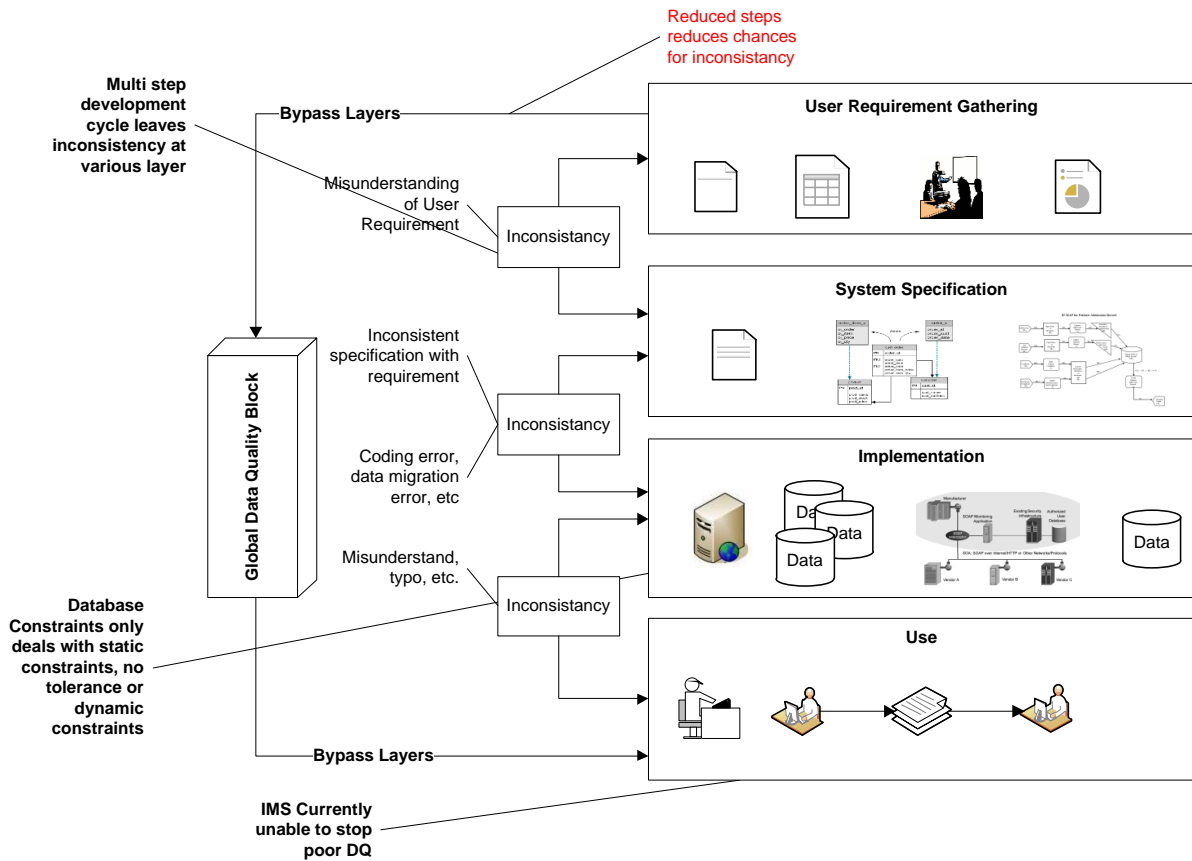


Figure 7: IMS Design Incorporating Global IQ Blocks

To demonstrate a possible implementation of the global IQ block, I illustrate my work in the context of a Hotel Reservation System. Two scenarios described here are actual scenarios that I have used for evaluation. In EMOS, booking must be made in the future, i.e. the arrival date must be after the reservation date. Once a booking is made, hotel is notified immediately of the booking. At this stage, the booking will be provisional. However, once the payment is received, the booking will become confirmed and a further email will be sent to the client confirming the booking. Eventually if the booking was cancelled, two cancellation emails will be sent, one to the client and one to the hotel.

Since extensible markup language (XML) is currently the preferred and trendy technology for data interchange (Caballero, et al., 2006), Caballero based DQXML on XML to define the quality rules. I based my work on Caballero and extended the quality rules definition to incorporate the process at the time of data inspection so that the certification can be made in the context of the process that generated the data.

In order to demonstrate the application of DQXML in developing the quality rules I refer back to my example. Let us assume that BOOKING is a data product that will be produced by the hotel reservation system. One of my assumptions is that reservation date must be before the guest's arrival date. This particular aspect of the BOOKING can be described as below in Figure 8.

```
<dq:quately_check_point>

  <dq:check_type>product_condition</dq:check_type>

  <dq:quality_dimension>accuracy</dq:quality_dimension>

  <dq:quality_user>call center agent</dq:quality_user>

  <dq:quality_process>making a booking</quality_process>

  <dq:quality_process>updating a booking</quality_process>

  <dq:condition_test>Booking.ReservationDate lt
Booking.ArrivalDate</dq:condition_test>

  <dq:quality_tolerance>0%</dq: quality_tolerance >

</dq:quately_check_point>
```

Figure 8: DQXML for Booking Reservation Date Condition

Let's examine a second aspect of my example and consider the BOOKING product as it passes through various stages in production. I can easily track and record all emails sent for a given booking. Let's assume that all of these emails are also stored in the database in a table called MailTracker. This stage based product criteria can be described in Figure 9.

```

<dq:quality_check_point>

  <dq:check_type>migrants_counter</dq:check_type>

  <dq:quality_dimension>Completeness</dq:quality_dimension>

  <dq:quality_user>call center agent</dq:quality_user>

  <dq:quality_process>making a booking</quality_process>

  <dq:quality_process>updating a booking</quality_process>

  <dq:foreign_table>MailTracker</dq:foreign_table>

  <dq:foreign_table_row_condition>MailTracker.isAcknowledged eq TRUE</dq:
foreign_table_row_condition >

  <dq:chose>

    <dq:when test="Booking.BookingStatusID eq
BOOKING_STATUS_PROVISIONAL">

      <dq:number_of_migrants>1</dq:number_of_migrants>

    </dq:when>

    <dq:when test="Booking.BookingStatusID eq BOOKING_STATUS_CONFIRMED">

```

```
<dq:number_of_migrants>2</dq:number_of_migrants>

</dq:when>

<dq:when test="Booking.BookingStatusID eq BOOKING_STATUS_CANCELLED">

    <dq:number_of_migrants>gt 2</dq:number_of_migrants>

</dq:when>

</dq:chore>

<dq:quality_tolerance>40%</dq: quality_tolerance >

</dq:quately_check_point>
```

Figure 9: DQXML for Booking Email Business Rules

Not only I am able to codify the quality rules, I am also able to tag process, user and data model along with the quality rules. This will help us determining root cause when a rule might be violated as the error can only occur either by a process, or a user or a data model.

Final part of my global DQ block, the on-going automated monitoring could be architected as shown in Figure 10 below.

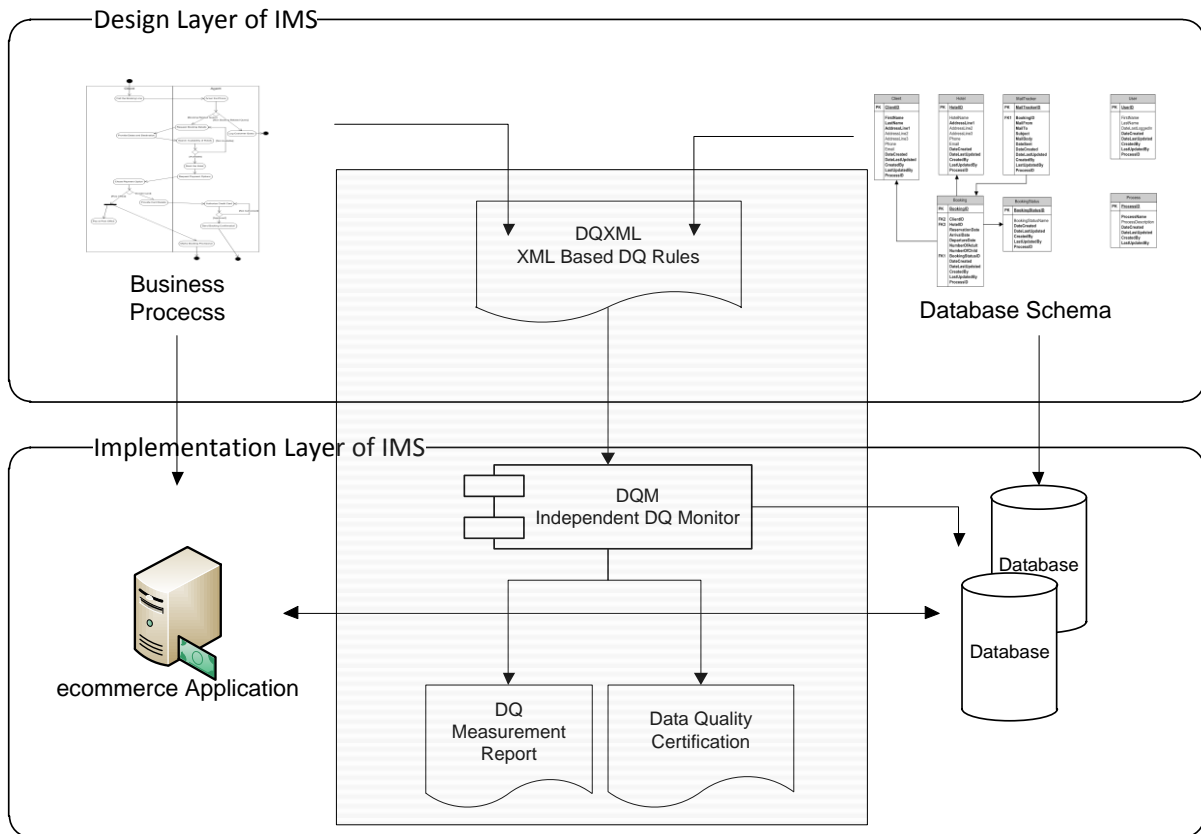


Figure 10: DQ Monitoring and Certification Model

The area outside the shaded box is usually present in a typical IMS development process. By introducing the global IQ block, I can codify the quality rules independent of system implementation and ensure on-going monitoring quality conformance.

4.4 Designing Data Quality Certificate

A DQ certificate format can be designed in many ways depending on the need of the user. Important part of the solution is its ability to provide basic information required for the DQ certificate. Once the required information is available, certificate itself may take various

forms. There are also various levels where certification might be provided. However, at top most level, the information might not be very useful. For instance, DQ certificate could merely state that “DQ is measured at 90%”. This will require summarizing every dimension and their measurement into one single dimension. Not only the matrix to develop such level summary would be complicated and complex, it would itself do not provide enough detail to the data consumer enabling them to take action for improvements.

At the next level, certificate could be issued for each DQ dimensions measured provided along with its measurement. For example, DQ certificate could state that “data is 90% complete 50% accurate, etc.” While this approach provides the consumer a sense of meaningful information, however, this will still have some of the issues raised for top level certification.

I adopt the approach where DQ certificate provides certification at each DQ rules level. This helps avoid measurement of various rules into a complex amalgamated figure, even though it might relate to the same dimension. This provides specific measurement in relation to a specific DQ rule enabling the consumer to take action when the DQ problem is certified to be outside of tolerance level. A basic construction of the DQ certificate is presented in Table 2.

DQ RULE	DQ DIMENSION	DQ MEASUREMENT	TOLERANCE – DESIRED RANGE	COMPLIANCE STATUS
...
...

Table 2: Elements of DQ Certificate

5 Evaluation

In this chapter I evaluate the utility of the artefact designed in the previous chapter. Initially I describe the strategy I selected for the evaluation. I then present the findings of my evaluation.

5.1 Evaluation Strategy

Evaluation has two parts to it. One is to develop the criteria for development and other is the assessment of the effectiveness of the artefact against the criteria. Evaluation has been emphasised as one of the most important part of the DS methodology (March & Smith, 1995) (Herver, et al., 2004). Role of evaluation is not only to determine if something worked or not, it also has to demonstrate why it worked using natural science methods so that the findings can be theorized (Pries-Heje, et al., 2008). A suitable evaluation framework will enable me to develop an evaluation strategy for the research outcome and it will also help me reach the required rigour. I have adopted a comprehensive evaluation framework proposed by Pries-Heje (Pries-Heje, et al., 2008) featuring *what* is actually evaluated, *how* the evaluation is conducted and *when* evaluation takes place.

5.1.1 What is Evaluated

. In the previous chapter I introduced a Global DQ Block, InfoGuard, incorporating DQXML and an on-going automated monitoring tool that provides DQ certification. InfoGuard has been designed keeping in mind the primary objectives for this research. In order to evaluate whether InfoGuard has reached its objectives, I will evaluate to see if InfoGuard is able to detect DQ problems, reduce DQ problem detection lead time and certify DQ on an on-going

basis. I will also see if InfoGuard is system independent, practical and adoptable. Finally, I will further observe the impact data certification has on DQ.

As described in section 2.1, various researchers have focused on different sets of DQ dimensions for their work. For the purpose of evaluating InfoGuard, I have selected the following three dimensions: *accuracy*, *completeness* and *timeliness*. The reason for the selection is that they are most common amongst most researchers. While strategy could be developed to certify on any DQ dimensions (objective or subjective), I have focused in this research on the objective dimensions. Other researchers have also adopted similar strategy for effective evaluation (Ballou & Pazer, 1985) (Scannapieco, et al., 2004).

5.1.2 How is it Evaluated

In DS, evaluation can take two forms: artificial and naturalistic evaluation. In the first situation, tool is evaluated in a laboratory simulation. In naturalistic evaluation solutions are deployed in a real world situation. In some cases, evaluation can only be done artificially. However, where possible many have favoured the natural evaluation. They argued that involving real users using real system solving real problem embraces the complexity of human practice in real organizations (Pries-Heje, et al., 2008).

In this research I have adopted the evaluation strategy to deploy InfoGuard in real world setting. I have engaged InfoGuard against two enterprise level systems to carry out the evaluation.

5.1.3 When Evaluation Takes Place

There are generally two options as to when evaluation takes place, *ex ante* and *ex post*. In *ex ante*, artefact is evaluated before the construction of the artefact. In *ex post*, artefact is evaluated after the implementation of the system (Pries-Heje, et al., 2008).

In my research I have adopted the *ex post* approach. In order to evaluate the effectiveness of InfoGuard, I had to put InfoGuard to test in real settings. Therefore, InfoGuard was first constructed and then deployed for evaluation.

5.2 Evaluation Method

As indicated in the research design in section 3.7, action research is used to carry out the evaluation of the research. I describe my active part in Susman's (Susman, 1983) five steps approach for carrying out an action research: Diagnosis, Action Planning, Action Taking, Evaluating and Specifying Learning. I identified DQ problems and lack of DQ certification as an issue from being part of the practice for over 15 years. After comprehensive literature review I diagnosed into some of the root causes for DQ problems. I then designed a solution that might be able to provide on-going DQ certification to data consumers. I provided the detail architect of the proposed solution in chapter 4. I then engaged with two suitable systems where the designed solution could be put into action. I actively took part in discussing the solution with the stake holders and identified the scenarios that are appropriate for this research. I designed the DQ rules and implemented the monitoring at appropriate interval. I then adopted a detail evaluation strategy. I report the evaluation part of the research in chapter 5. Finally I summarize the findings of my work in chapter 6 where I specify the learning from this research.

5.3 Evaluation Scenarios

Evaluating the utility of the artefact to achieve its objective remains a primary way to evaluate (Foley, 2011). I engaged with Eden Further Education (Eden), Dublin to carry out the first evaluation of InfoGuard. In the scope of the first study, I focused on InfoGuard's ability to achieve its objective, which is to identify DQ problems, certify DQ and to reduce lead time for DQ problem detection. I also observed the improvement of DQ over time.

After discussions with the stakeholders in Eden, various organisational impacts arising from DQ problems were identified. Some identified impacts were number of posts returned as undelivered, unnecessary postage cost when email could be used free of charge, student complaints, allocation of too many or too little classrooms, last minutes class cancellations, etc. First two in the list above were identified by the stakeholders as top priority for the purpose of this study.

5.3.1 Scenario 1: Up to Date Student Address

I undertook an analysis of the process involved in this scenario. A large number of students come from overseas to learn English and initially they take up temporary accommodation. This address is registered in the database when the student enrolls with Eden. Students on securing permanent accommodation often fail to notify Eden of their change of address. The same thing occurs with other students who relocate. Eden has a constant need to communicate various updates and notices with students. A large number of posts are returned undelivered. It is important that Eden keeps student contact details up to date. DQ dimension related to this process is *Timeliness*. The process is presented in Figure 11.

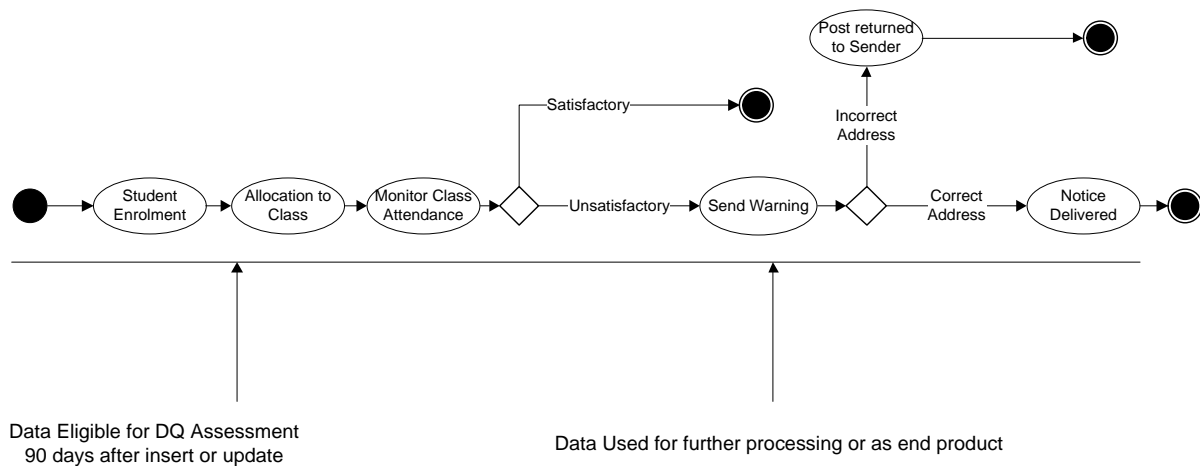


Figure 11: Notice for Unsatisfactory Attendance

This problem was detected only when post was returned undelivered. There was an average of 180 days delay in detecting out of date student address information.

5.3.2 Scenario 2: Missing Student Email Addresses

When exam results are published, students are notified by post at a substantial cost to Eden. Eden cannot refuse a student on the basis of the student not having an email address. Therefore, email provision cannot be made mandatory. However, if 90% of the students had an email addresses, email could be their primary means of publishing result and the remaining 10% by post. The IQ dimension related to this process is *Completeness*. The process is presented in Figure 12.

Incomplete email records only became visible when the method was first implemented in SMS and after 2 years of SMS in use. As exam results are sent out end of each semester, Eden only detected this IQ problem every 90 days.

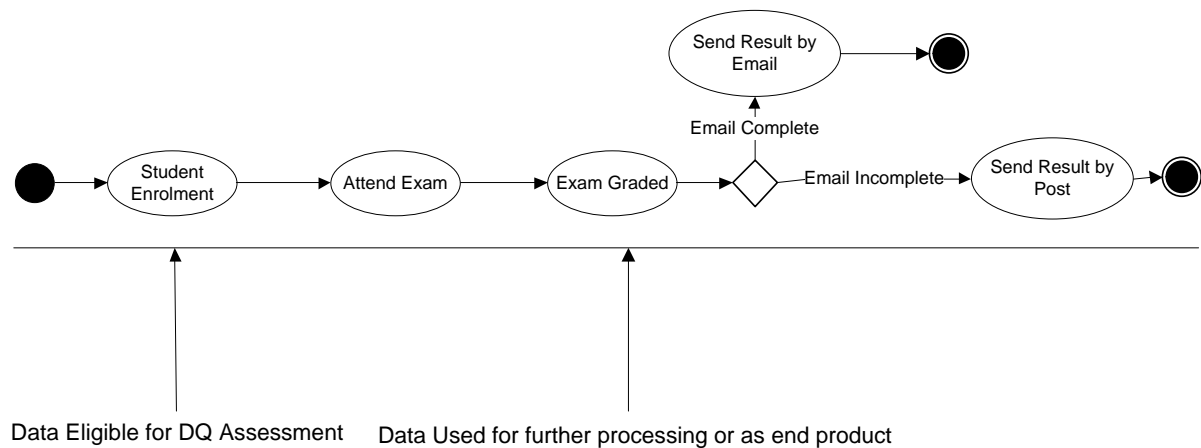


Figure 12: Process for Posting Grades

For my second study, I engaged with EMOS. After discussion with stakeholders, the most damaging problem was identified as when a guest arrives in a hotel to that there are no room booked for them. During off season, the problem is easy to solve as hotels generally have plenty of unoccupied room to facilitate the guest. However, if the hotel does not have additional room available it can result into a very unsatisfied customer with potential liability to compensate the client. There are many reasons as to why this might arise. However, for the purpose of this study, I selected to monitor two scenarios. One is when the arrival date in incorrectly recorded. And the other is when the hotel actually did not receive a notification of the guest's booking.

5.3.3 Scenario 3: Incorrect Booking Dates

After analysing the booking process, presented in Figure 13, there are several reasons why the booking dates could be incorrect. Primary reason for this mistake was booking agents putting incorrect dates while making the booking. Booking dates could be either before or after the intended travel dates. It was difficult to detect automatically when this occurred.

However, when bookings are made, arrival dates must be after the reservation date. This particular mistake accounted for almost 30% of the incorrect booking date entries. This particular process related to DQ dimension of *accuracy*.

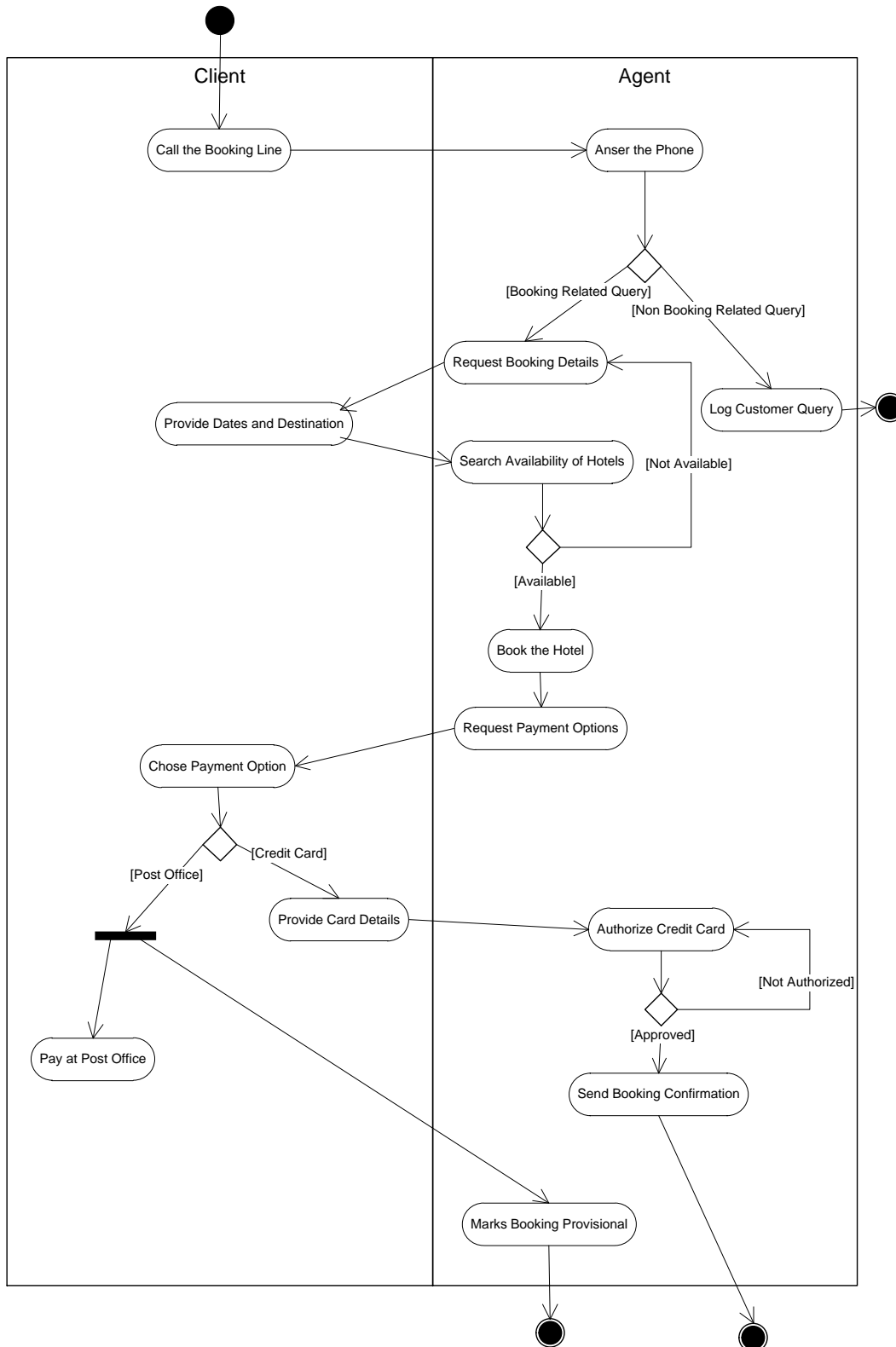


Figure 13: Process for Making a Booking

5.3.4 Scenario 4: Missing Booking Notification to Hotel

Once a booking is made, hotels are notified immediate of the booking. Depending on the state of the booking, number of notification sent to the hotel varies and these are described in detail earlier in section 4.3. This particular dimension of DQ related to *completeness*.

5.4 Implementing InfoGuard

When implementing InfoGuard, a plan has to be devised in the context of the process involved in generating the data in the first place. This is helpful for not only designing the monitoring rule; it also helps to determine the frequency at which the monitoring should occur. In determining frequency of monitoring, attention has to be paid on the availability of resources required to review the certification and take corrective action where possible.

5.4.1 Scenario 1: Up to Date Student Address

In order for the data to become out of date, I first had to determine the criteria that would make the data out of date. Consulting with the users, I conducted a survey to see how often students would relocate themselves. There were some students that move very rarely while others were frequent movers. In this context, based on the survey for average stay in a particular location, an address is up to date only for 90 days after system input. So, I constructed the quality rules that were intended to monitor student records age and detected problems with data older than 90 days. This rule is presented in the in Figure 14. This allowed us to certify as to how timely the student records were.


```

<dq:quately_check_point>

    <dq:check_type>product_condition</dq:check_type>

    <dq:quality_dimension>timeliness</dq:quality_dimension>

    <dq:quality_user>admin team member</dq:quality_user>

    <dq:quality_process>enrol student</quality_process>

    <dq:quality_process>updating student record</quality_process>

    <dq:condition_test>(Now() - Student.DateLastUpdate) lt 90</dq:condition_test>

    <dq:quality_tolerance>30%</dq: quality_tolerance >

</dq:quately_check_point>

</dq:data_product>

```

Figure 14: Quality Rules for Scenario 1

I looked into the availability of resources to review the certification and take corrective measures. Eden was able to put in place a weekly review plan for student records identified as out of date. Students were then contacted by phone to verify the student address in the system. This rule could easily be monitored and certified more frequently. However, without being able to review the results and follow up plans, certification will make little difference. InfoGuard settings for this quality rule are presented in Figure 15.

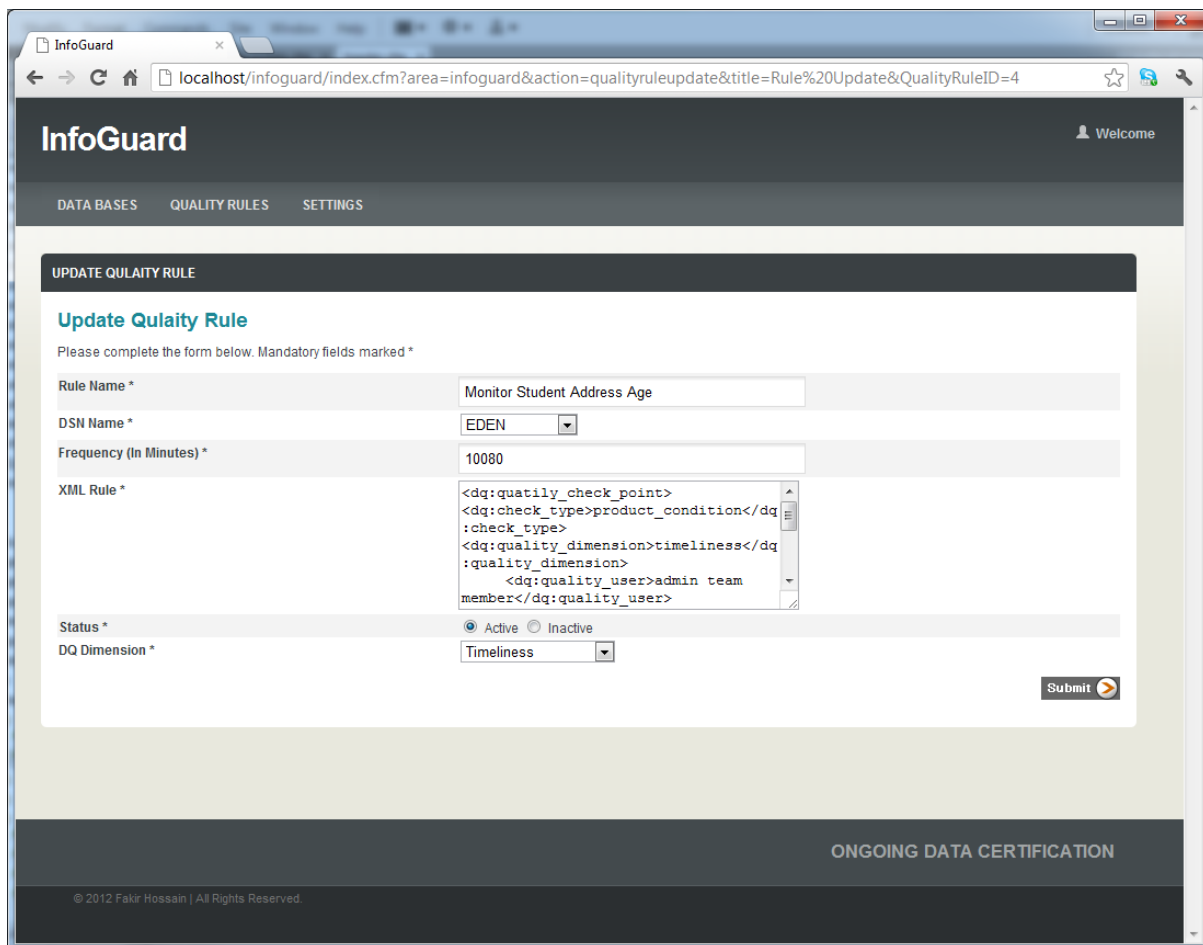


Figure 15: Quality Rules Settings in InfoGuard for Scenario 1

5.4.2 Scenario 2: Missing Student Email Addresses

After discussing the users, it transpired that the main reason for so many students missing email address in the database was an overall lack of emphasis on collecting the email in the first place. Quality rule for this scenario was that student records were monitored for email completeness. Tolerance of 10% was also introduced. This means that even though 10% student records might be missing email, this would still be considered acceptable. The quality rule for this scenario is presented in Figure 16.

```
<dq:quately_check_point>

    <dq:check_type>product_condition</dq:check_type>

    <dq:quality_dimension>completeness</dq:quality_dimension>

    <dq:quality_user>admin team member</dq:quality_user>

    <dq:quality_process>enrolling student</quality_process>

    <dq:quality_process>updating student record</quality_process>

    <dq:condition_test>Student.Email gt ''</dq:condition_test>

    <dq:quality_tolerance>10%</dq: quality_tolerance >

</dq:quately_check_point>
```

Figure 16: Quality Rules for Scenario 2

Students were serviced every day for various reasons by the admin team members. This presented an opportunity for the staff to collect student's email address if this was not already collected before. Due to the nature of this rule, it was decided that I monitor this rule every night and report the findings to ensure that a steady progress is being made to reach the desired level of quality. Settings for this quality rule are presented in Figure 17.

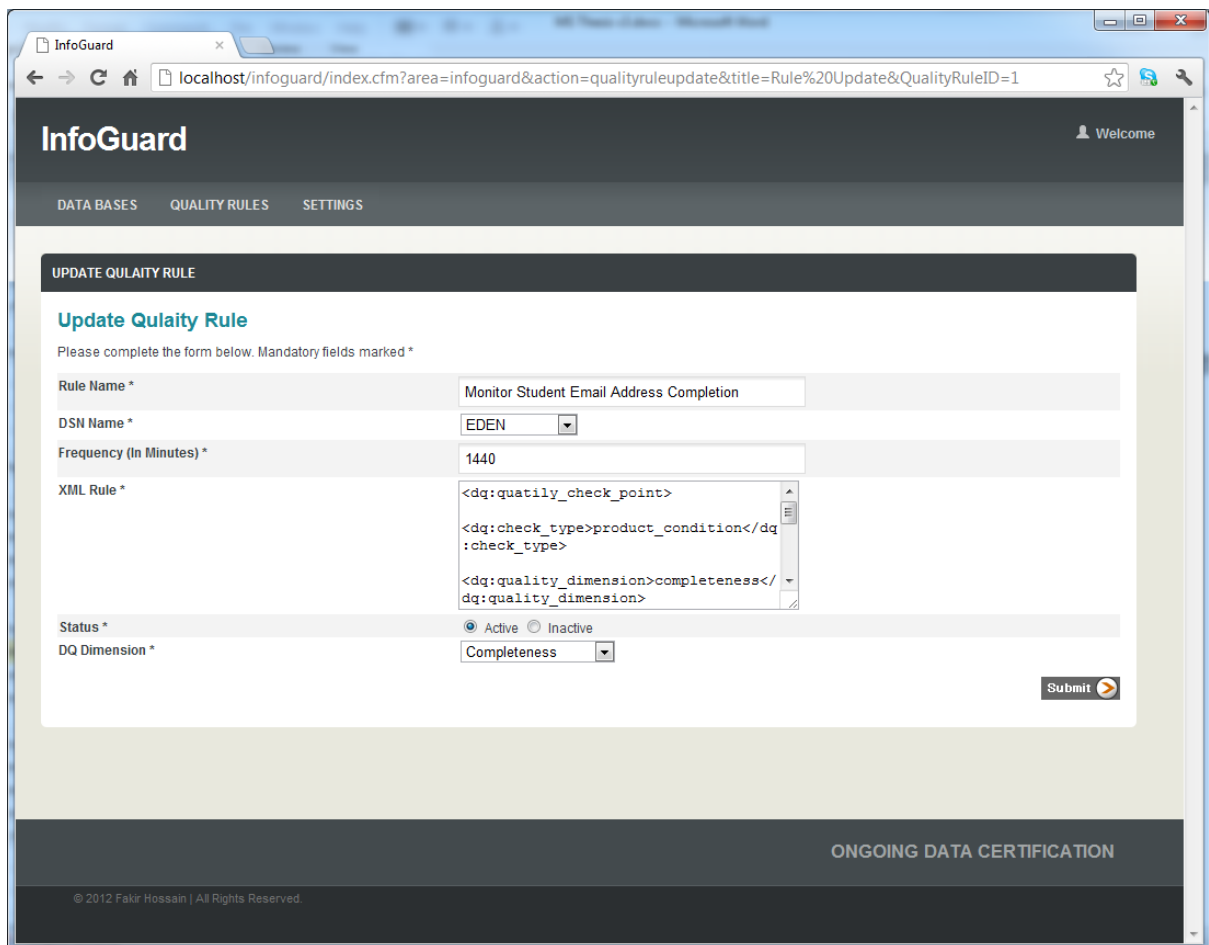


Figure 17: Quality Rules Settings in InfoGuard for Scenario 2

5.4.3 Scenario 3: Incorrect Booking Dates

Process involved when incorrect booking dates occurred are *making a booking* and *updating a booking*. Quality rule for this scenario was that the data product booking, regardless of its state, must confirm with the condition that reservation is always before the arrival date. The quality rule for this scenario is already presented in Figure 8.

This particular data problem, while occurred infrequently, had severe consequences. This could result in an unsatisfied customers potentially stranded in a hotel without a room secured for them. Users were eager to identify these bookings shortly after it occurred. They set up

a process where customer care staff will review this monitoring report every one hour and take corrective action when this error occurred. Settings for this quality rule are presented in Figure 18.

The screenshot shows a web browser window with the URL `localhost/infoguard/index.cfm?area=infoguard&action=qualityruleupdate&title=Rule%20Update&QualityRuleID=2`. The page title is "InfoGuard" and the user is logged in as "Welcome". The navigation menu includes "DATA BASES", "QUALITY RULES", and "SETTINGS". The main content area is titled "UPDATE QUALITY RULE" and contains the following form:

Update Quality Rule
Please complete the form below. Mandatory fields marked *

Rule Name *	INCORRECT BOOKING DATES
DSN Name *	EMOS
Frequency (In Minutes) *	60
XML Rule *	<pre><dq:quaitly_check_point> <dq:check_type>product_condition</dq: check_type> <dq:quality_dimension>accuracy</dq:q uality_dimension></pre>
Status *	<input checked="" type="radio"/> Active <input type="radio"/> Inactive
DQ Dimension *	Accuracy

Submit

ONGOING DATA CERTIFICATION

© 2012 Fakir Hossain | All Rights Reserved.

Figure 18: Quality Rules Settings in InfoGuard for Scenario 3

5.4.4 Scenario 4: Missing Booking Notification to Hotel

Process involved in missing booking notification to hotel are *making a booking* and *updating a booking*. There are many reasons why an email may not reach the hotel and some of these factors are outside the control of EMOS. When an email is dispatched an entry is made in the MailTracker table. Once the hotel receives the email, there is a link to confirm the receipt of

the email. This confirmation is sent to EMOS via HTTP protocol and the MailTracker table is updated with the confirmation of the receipt by the hotel. Quality rule for this scenario is already presented in Figure 8.

This particular data problem also had severe consequences. This could result in an unsatisfied customers potentially stranded in a hotel without a room secured for them. Users were eager to identify these notifications if the email confirmation was received within 24 hours of dispatch. They set up a process where customer care staff will review this monitoring report every morning and take corrective action when this error occurred. Settings for this quality rule are presented in Figure 19.

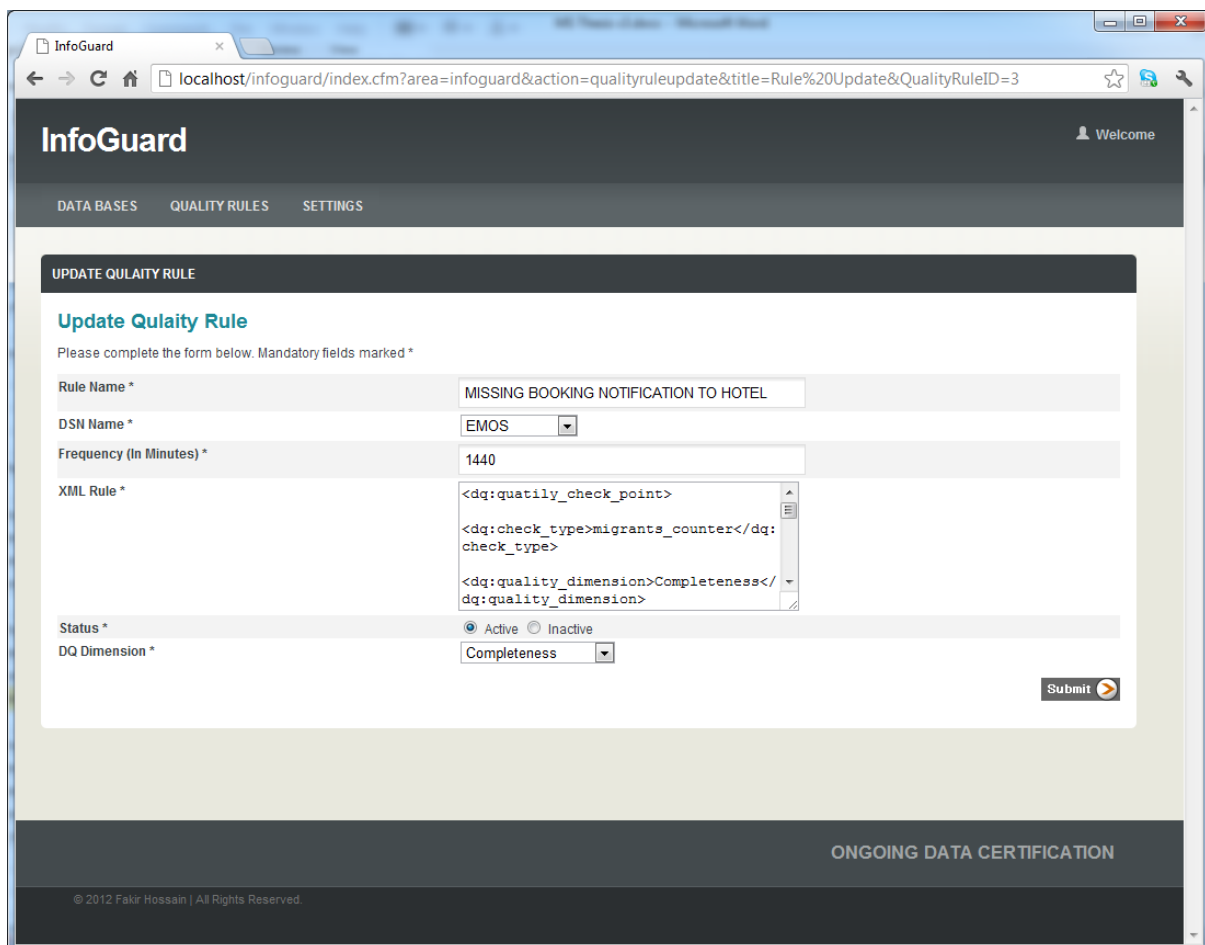


Figure 19: Quality Rules Settings in InfoGuard for Scenario 4

5.5 Findings

InfoGuard is set up to monitor the relevant quality rules for compliance and produce the required reports to evaluate my research objective. I present my findings for the various scenarios below.

5.5.1 Scenario 1: Up to Date Student Address

There were three reports I was primarily interested in. I wanted to see if InfoGuard was able to detect and report the level of DQ problem in the first place. My findings in this regard for a period of three months are presented in Figure 20. The process identified the total number of records that did not meet the 90 days rules for student address being up to date. For example in week 1, InfoGuard reported 5674 records exciding the 90 days age rule. In week 12 the number was 5072.

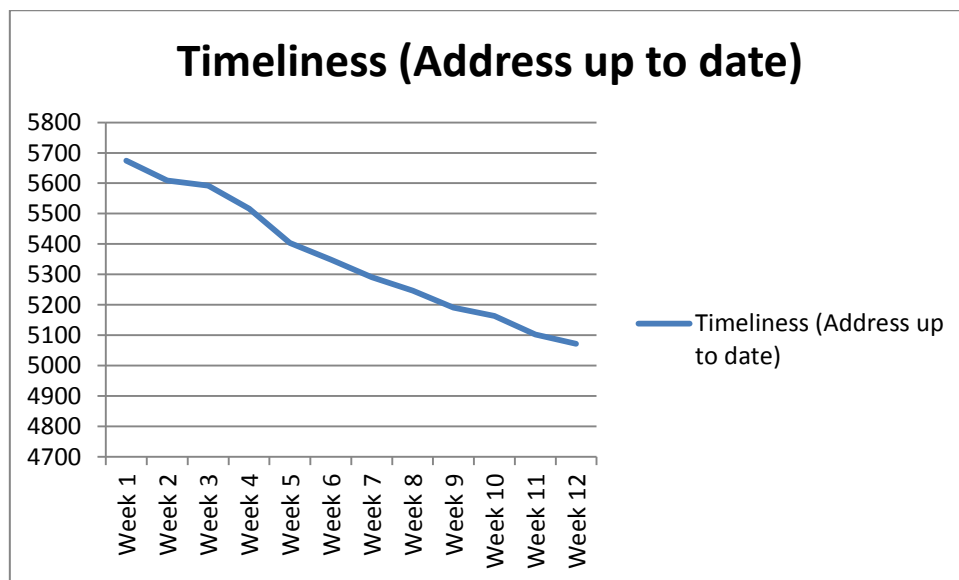


Figure 20: Ability to Detect DQ Problems for Scenario 1

Number of steps was taken to see if once the college came to know the scale of the problem with student address data being out of date. A process had been introduced to verify a student’s address upon student contact with the admission office. Each student in class was given a contact update form and had to provide their up to date information regardless of what the system contained. This was then matched for each student record and updates were made as appropriate. In this scenario, timeliness was calculated by using the following formula:

$$Timeliness = \left\{ \max \left(1 - \frac{\text{age of data (in days)}}{90 \text{ days}}, 0 \right) \right\}^1$$

My second objective was to see if InfoGuard was able to improve DQ over time. My finding is presented in Figure 21. Timeliness did improve over time. I like to clarify though InfoGuard does not in itself improve DQ. However, through on-going certification, it increase consumer awareness about the problems and if actions are put in place to reduce the problems, DQ will improve over time.

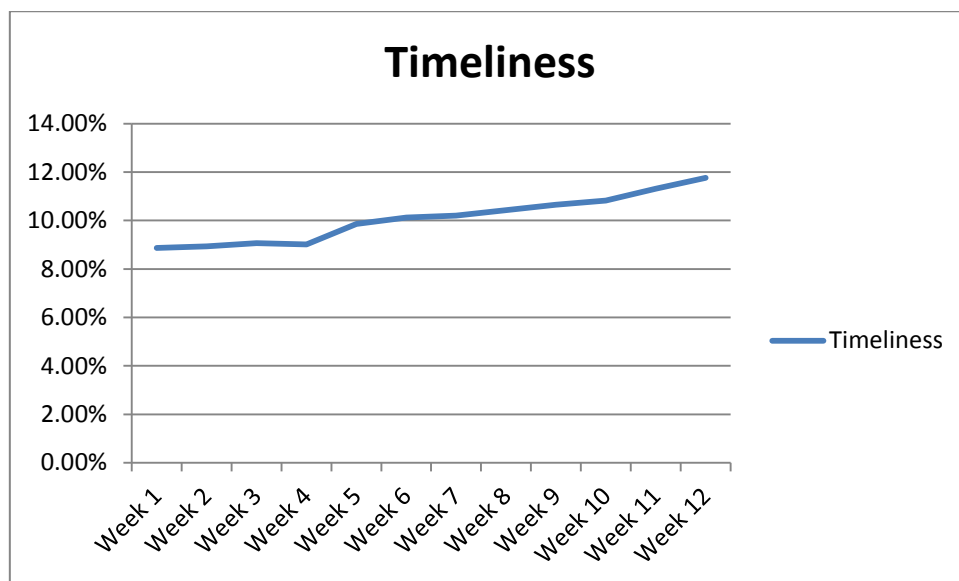


Figure 21: DQ Improvement over Time for Scenario 1

Finally I wanted see if InfoGuard is able to reduce lead time for DQ problem detection. My findings are presented in Figure 22. This was reduced form earlier average of 180 days to 7 days after implementation of InfoGuard. As mentioned before, lead time was on average 180 days. There were no proactive ways to find out if the address were out of date. Once the weekly monitoring process was implemented, the lead time for DQ problem detection was significantly reduced.

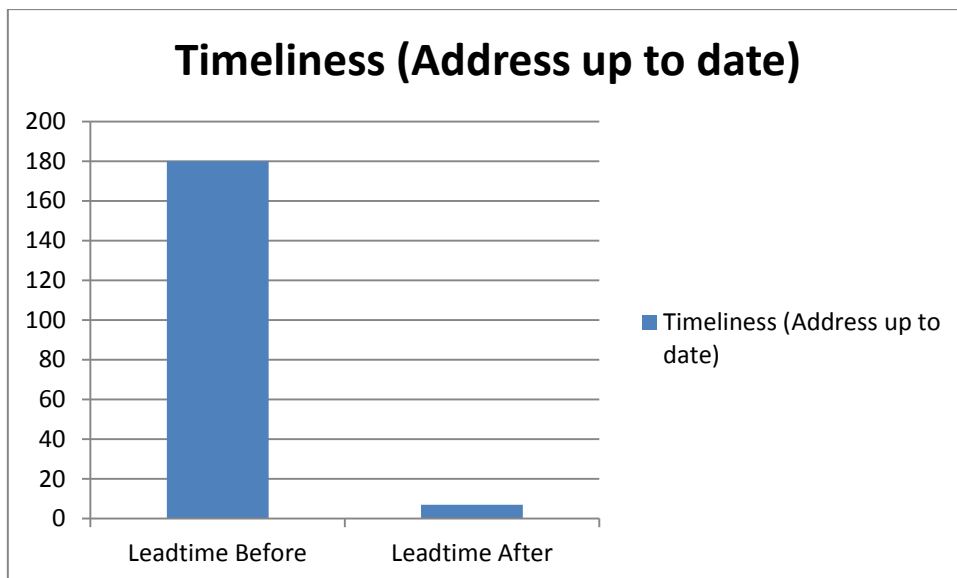


Figure 22: Ability to Reduce Lead Time for Scenario 1

5.5.2 Scenario 2: Missing Student Email Addresses

My findings in relation to InfoGuard’s ability to detect DQ problem for this scenario are presented in Figure 23. InforGuard detected the records that were missing email address and reported the numbers short for it to reach 90% completeness. For example in week 1, 3920 records were requiring email address for it to meet the quality tolerance designed by the rule.

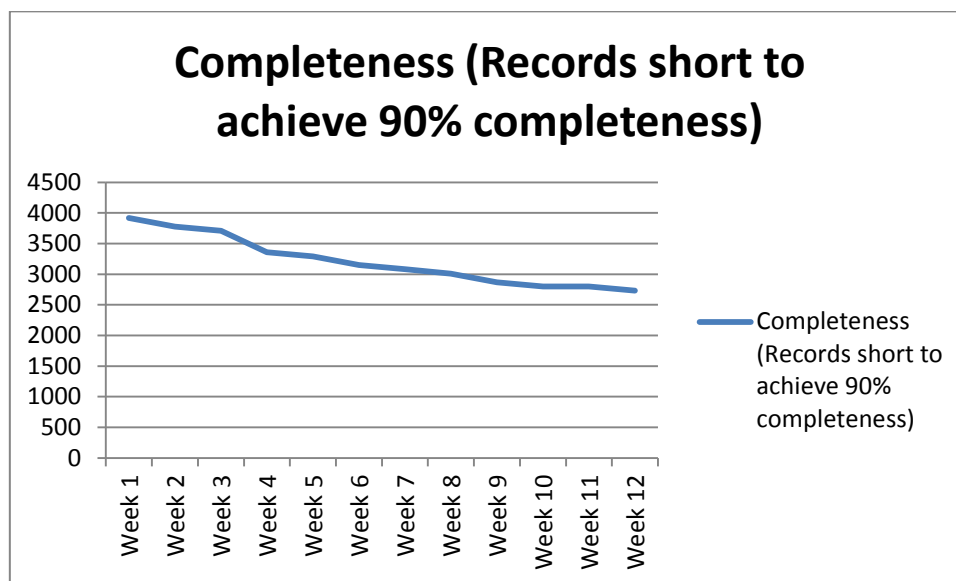


Figure 23: Ability to Detect DQ Problems for Scenario 2

Similar processes were introduced for email collection and staff training was provided to emphasise the importance of email address collection. The following formula was used to calculate the *completeness* for email records:

$$completeness = 1 - \frac{\text{number of students missing email}}{\text{total number of students}}$$

My finding on the second objective of InfoGuard to observe DQ improvement over time is presented in Figure 24. *Completeness* gradually improved as more and more records were

updated collecting emails from students. For example in week 2, data was only 36% complete. At the end of 12 weeks, data was 51% complete.



Figure 24: DQ Improvement over Time for Scenario 2

My findings in relation to the research objective to reduce lead time for DQ problem are presented in Figure 25. This was reduced from earlier average of 180 days to 1 day after implementation of InfoGuard. Lead time for scenario 2 was on average 180 days. There were no proactive ways to find out if email address were incomplete. Once the daily monitoring process was implemented, the lead time for DQ problem detection was drastically reduced.

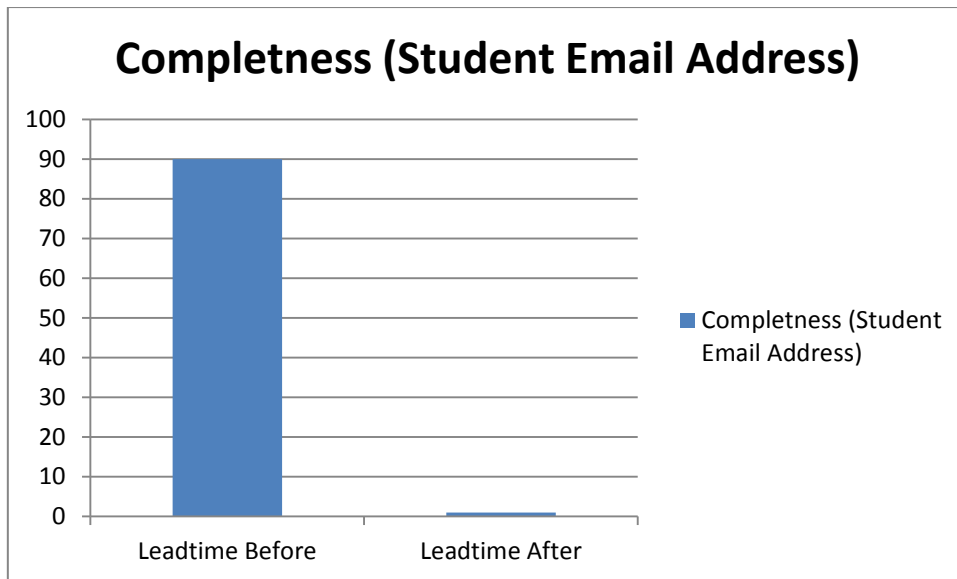


Figure 25: Ability to Reduce Lead Time for Scenario 2

5.5.3 Scenario 3: Incorrect Booking Dates

My findings in relation to InfoGuard’s ability to detect DQ problem for this scenario are presented in Figure 26. While the number of occurrence was few, InfoGruad was able to identify the bookings that failed to comply with this rule. There were weeks, for example at week 2, week 7 and week 8 when all booking were *accurate*. However, the number ok bookings picked to 5 bookings during week 5 that had reservation date after the arrival date.

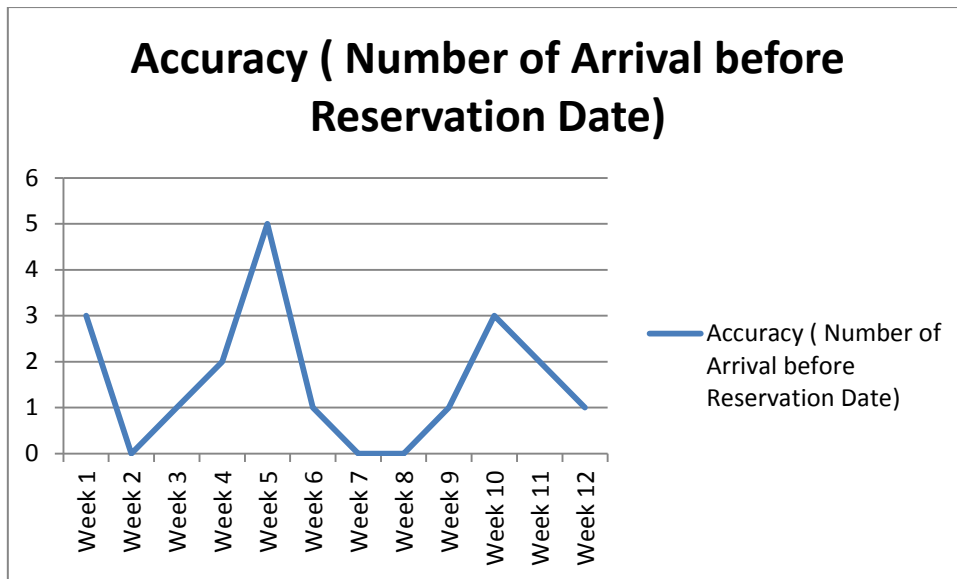


Figure 26: Ability to Detect DQ Problems for Scenario 3

Booking made for the past could not be totally banned in the system as some holiday taken by sponsor's Staff voucher were often placed in EMOS after the holiday is taken. However, staffs were alerted about this scenario and staff training emphasised the importance of paying attention to booking dates. My finding on the second objective of InfoGuard to observe DQ improvement over time is presented in Figure 27. The following formula was used to calculate the *accuracy* for the booking records:

$$Accuracy = 1 - \frac{\text{Number of boking containing error}}{\text{Total Number of booking}}$$

Given the high level of accuracy in this scenario, improvements were insignificant.

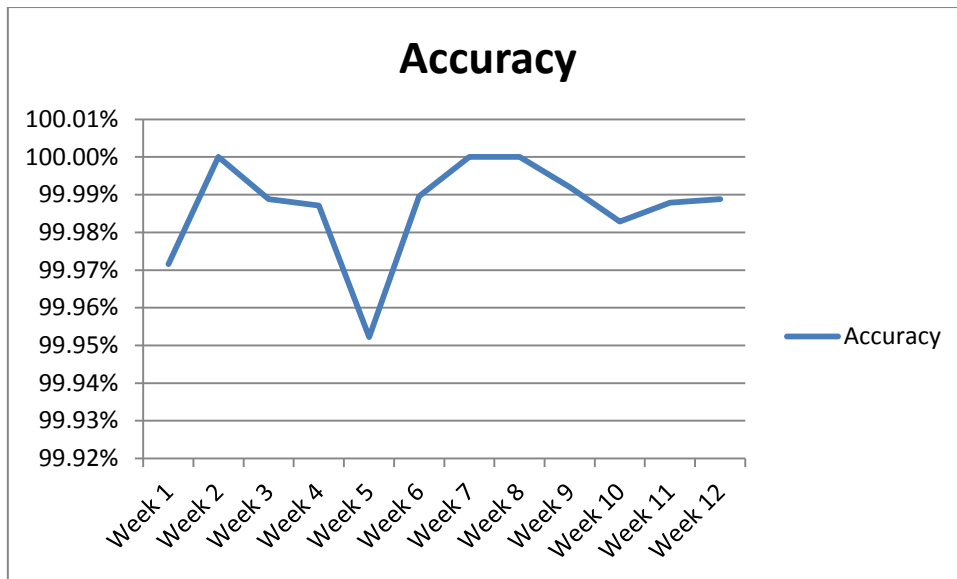


Figure 27: DQ Improvement over Time for Scenario 3

My findings in relation to the research objective to reduce lead time for DQ problem are presented in Figure 28. This was reduced from earlier average of 7 days to 45 minutes after implementation of InfoGuard. Introducing the hourly process to certify *accuracy* in this scenario ensured that the lead time for detecting inaccurate bookings was less than 1 hour.

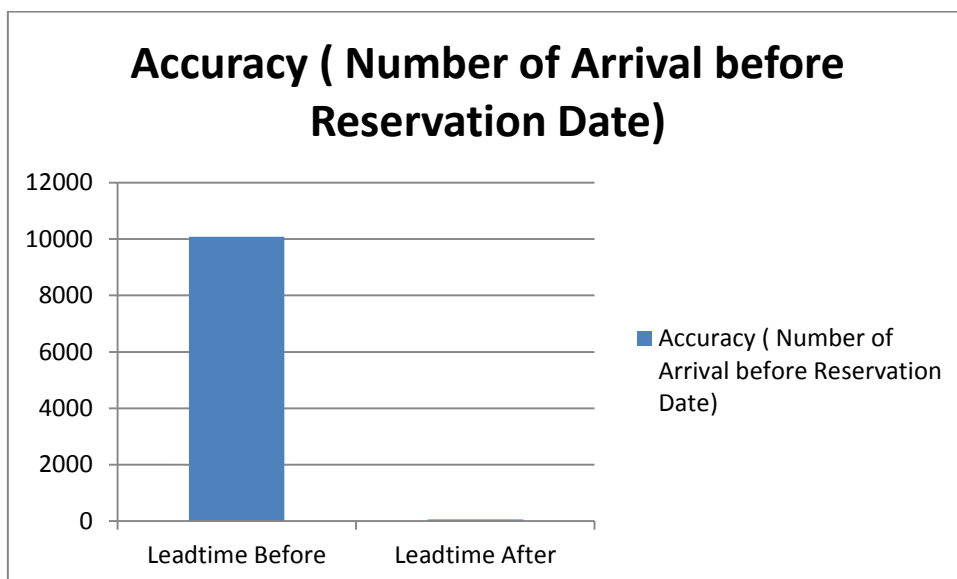


Figure 28: Ability to Reduce Lead Time for Scenario 3

5.5.4 Scenario 4: Missing Booking Notification to Hotel

My findings in relation to InfoGuard’s ability to detect DQ problem for this scenario are presented in Figure 29. Booking that did not have an acknowledgement in the MailTracker were identified by InfoGuard. There were no particular patterns to this as most of the factors were outside the control of the company. During various weeks, number of booking without acknowledgement ranged from 50 to 800 bookings.

Some problems were identified with the email not reaching the hotel. Amongst them were having incorrect email address for the hotel, lack of attention on behalf of the hotel for confirming the email, etc. However, drastic changes were difficult to implement for this scenario. My finding on the second objective of InfoGuard to observe DQ improvement over time is presented in Figure 30.

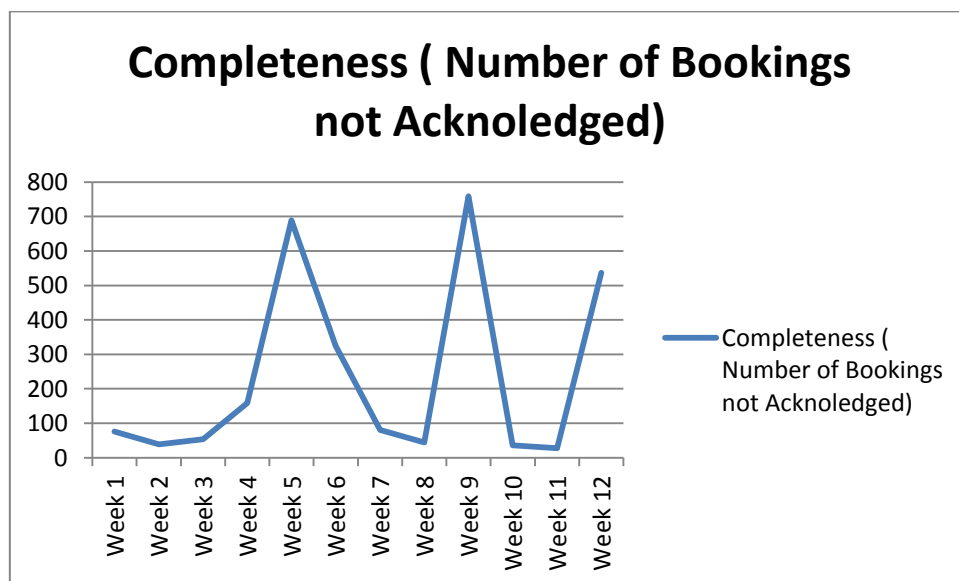


Figure 29: Ability to Detect DQ Problems for Scenario 4

The following formula was used to calculate the *completeness* for email records:

$$completeness = 1 - \frac{\text{number of bookings missing acknowledgement}}{\text{total number of bookings}}$$

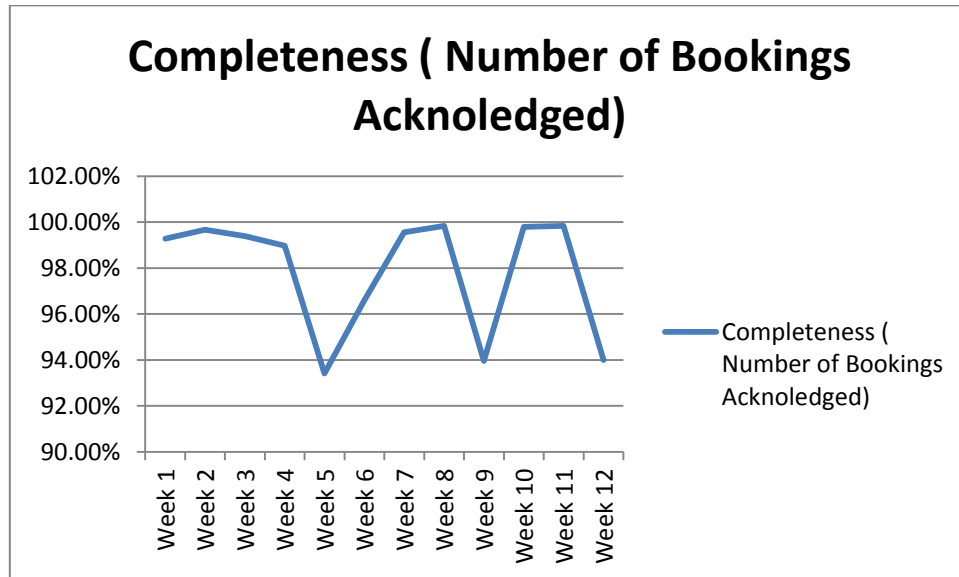


Figure 30: DQ Improvement over Time for Scenario 4

My findings in relation to the research objective to reduce lead time for DQ problem are presented in Figure 31. This was reduced form earlier average of 180 days to 1 day after implementation of InfoGuard. Introducing the daily certification process drastically cut down on the lead time for DQ problem detection.

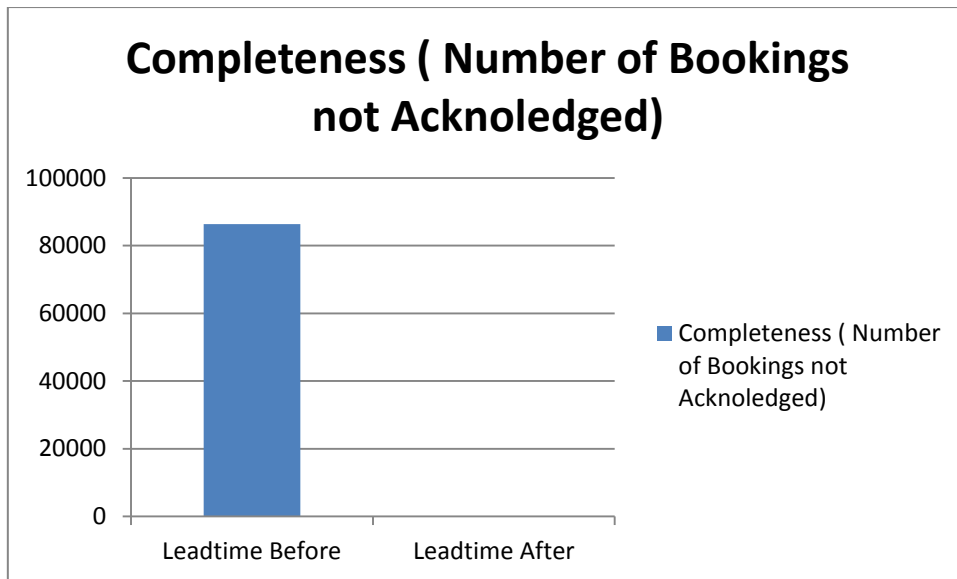
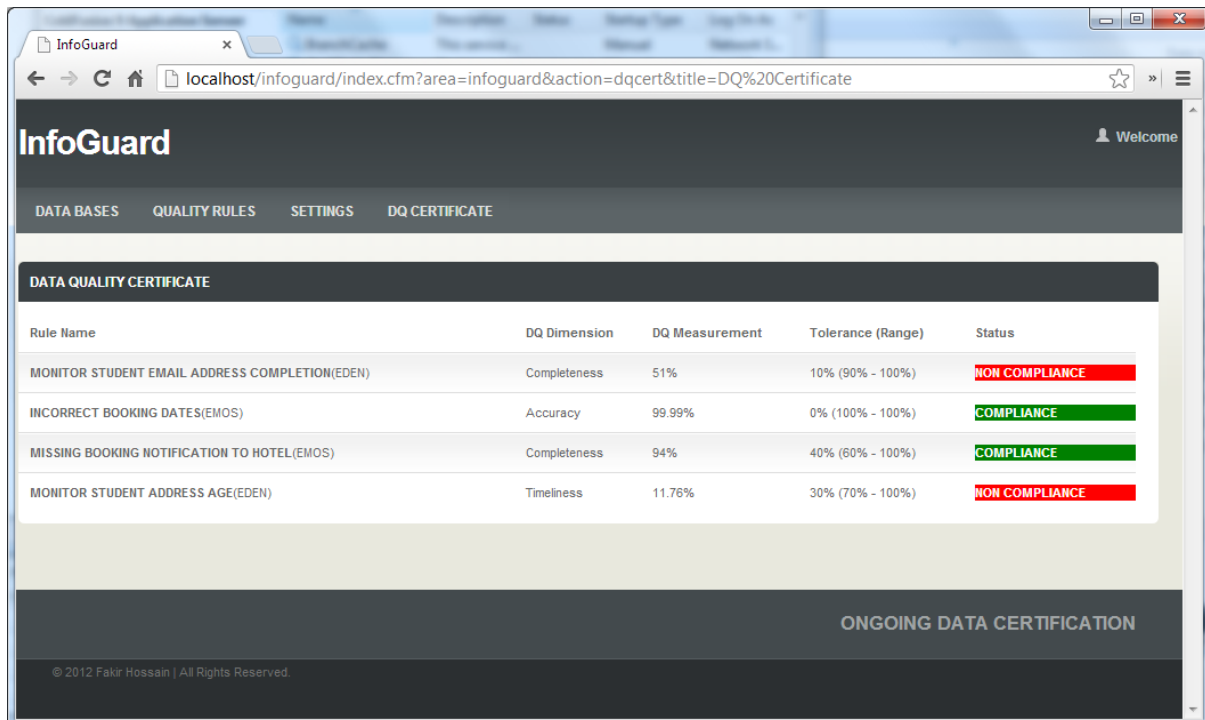


Figure 31: Ability to Reduce Lead Time for Scenario 4

5.6 Data Quality Certificate

I have already set out the format for the DQ certificate used in this research in section 4.4.

DQ certificate derived at week 12 for the scenarios discussed in the research are presented in Figure 32.



The screenshot shows the InfoGuard web application interface. The main content area displays a 'DATA QUALITY CERTIFICATE' table. The table has five columns: Rule Name, DQ Dimension, DQ Measurement, Tolerance (Range), and Status. The status column uses color-coded text: red for 'NON COMPLIANCE' and green for 'COMPLIANCE'. Below the table, there is a footer with the text 'ONGOING DATA CERTIFICATION' and a copyright notice '© 2012 Fakir Hossain | All Rights Reserved.'

Rule Name	DQ Dimension	DQ Measurement	Tolerance (Range)	Status
MONITOR STUDENT EMAIL ADDRESS COMPLETION(EDEN)	Completeness	51%	10% (90% - 100%)	NON COMPLIANCE
INCORRECT BOOKING DATES(EMOS)	Accuracy	99.99%	0% (100% - 100%)	COMPLIANCE
MISSING BOOKING NOTIFICATION TO HOTEL(EMOS)	Completeness	94%	40% (60% - 100%)	COMPLIANCE
MONITOR STUDENT ADDRESS AGE(EDEN)	Timeliness	11.76%	30% (70% - 100%)	NON COMPLIANCE

Figure 32: Data Quality Certificate at the End of 12 Weeks

DQ certificate lists each of the rules configured in InfoGuard. It reports the measurement and tolerance range for the DQ dimensions for each DQ rule. If the current measurement is outside the tolerance range, it also provides a non-compliance status so that the consumer can investigate this further.

6 Summary and Conclusion

In this chapter I provide a summary and conclusion of my research. Furthermore, I discuss the limitation of my work and provide indication of future work for my research.

6.1 Summary

I set out this research with the primary aim to provide a practical, adoptable, system independent solution for on-going DQ certification to data consumers. I also aimed to see if the solution would be able to detect DQ problems and if DQ detection lead time could be reduced as a result. I also wanted to observe if this on-going DQ certification had any impact on DQ itself. In this section I summarize how the InfoGuard addressed the research objective. I do so in exploring each of the research questions I set out to investigate to reach the objectives.

My first research question is to define DQ certification. Defining DQ certification is a very important aspect of this research as one of the primary objective of this research is to provide practical, adoptable, system independent on-going DQ certification to data consumer. In chapter 2, after providing a detail overview of related work to DQ, I adopted a definition of DQ certification in section 2.6. Furthermore, the solution offered in the research is a practical and adoptable one as it utilises DQXML rule based approach to DQ certification. This makes it non cost and time prohibitive. Also, as InfoGuard is a separate system and not a part of the IMS itself, it is independent of the IMS. This independence from the IMS offers added confidence and creditability to the data consumer.

My second research question is to see whether DQXML can be used to certify DQ. On-going monitoring and certification approach enabled me to detect the deficiencies in the data

product. It did so by reporting the data that did not meet the quality rule set out in the DQXML. This was presented in the DQ compliance report. It also identified the process responsible for the issue which helped expedite the corrective measure. As part of this on-going monitoring, InfoGuard produced data quality certificate detailing various quality aspects of data product using the comprehensive quality meta model specified in DQXML. This certification provides for confidence of owner and user of data. From the research findings, it is clear that DQXML is clearly capable of codifying the required quality metadata that can be utilized in the DQ evaluation process to provide DQ certification.

My third research question is if DQ certification is effective on an on-going basis. This issue is more evident when I address the next research question below. DQ certification is only really effective if this is done on an on-going basis. It is the on-going nature of it that informs the data consumer awareness of the DQ level. Once the consumers are aware of the DQ level, they are able to act on it. They are able to take appropriate measure to reduce DQ problems. Unless DQ certification is on-going, over time DQ level can easily slip out of acceptance level. It is not suggested here that just DQ certification alone will improve DQ. However, on-going certification will continue to keep the consumer active to prevent DQ level from falling out of quality tolerance.

My final research question is if DQ certification can reduce DQ problem detection lead time and improve DQ. As part of InfoGuard implementations, I planned how often the monitoring should take place given the context of the significance of the error and the availability of resource to review them. In the first scenario, lead time was reduced from 90 days to 1 week. In the second scenario, lead time was reduced from 90 days to 1 day. In the third scenario, 7 days lead time was reduced to 45 minutes. In the final scenario, 60 days of lead time was reduced to just 1 hour. Lead time reduction for DQ problem detection was successfully achieved by InfoGuard.

I also wanted to see the impact of InfoGuard on DQ. There is not a direct relationship between InfoGuard and DQ improvement. It all depended on various factors. Some of them were number of data effected in proportion to the entire data set, nature of the process involved in generating the data whether improvement can be made to the process itself, nature of the DQ problem whether the problem could be avoided all together. In the first scenario, the number of records that are out of date is very large in relation to the total number of records. Even though processes were introduced to update student data as soon as any opportunity arouse in any type of contact with students, DQ improvement was slow. Timeliness did improve from about 8% to 12% over three month period, however, as large number of student data related to students that were no longer studying at the college and therefore not in regular touch with the college for the data to be updated. Therefore drastic improvement in DQ was not possible. In the second scenario, circumstances were very similar to the first one. While some DQ improvement was observed, large number of missing email could not be easily collected. In the third and fourth scenario, natures of the DQ problems were such that the factors causing them were unrelated to the implementation of InfoGuard. There was little impact of InfoGuard on DQ improvement.

6.2 Comparison with other DQ Management Strategy

InfoGuard provides a comprehensive strategy for dealing with static (when quality rules are independent of state) and dynamic business rules (when quality rules are state dependent), detecting and identifying DQ problems automatically. A summary of this comparison is provided in Table 3.

	DQ Blocks with in IMS	Database Constraints	Process Management	InfoGuard On-going DQ Certification
Reduces DQ Problems	X	X	X	X (not directly)
Deal with Static Business Rules	X	X		X
Deal with Dynamic Business Rules	X			X
Detect DQ Problem			X	X
Detect DQ Problem Automatically				X
Reduces Detection Time for DQ Problems				X

Table 3: InfoGuard Evaluation

6.3 Limitation and Future Work

Some of the limitation of this approach is that legacy systems may not be able to benefit from it as quality meta model required for this might be absent. More research could be carried out developing a comprehensive Meta model for the quality and other IMS blocks to offer the benefit of the InfoGuard, yet offering required flexibility to the IMS engineers. Most literatures in DQ field are so complicated or abstract that it can hardly be used in everyday development. I expect the major contribution be the practical aspect of the InfoGuard in DQ monitoring and certification.

List of References

Aggarwal, C. & Yu, P., 2009. A survey of uncertain data algorithms and applications. 21(5).

Alavi, M. & Carlson, P., 1992. A review of MIS research and disciplinary development.

Journal of Management Information Systems, 4(8), pp. 45-62.

Ballou, D. P. & Pazer, H. L., 1985. Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science*, 31(2).

Ballou, D. P., Wang, R. Y., Pazer, H. & Tayi, G. K., 1998. Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, 44(4).

Ballou, P. & Pazer, L., 1985. Modeling Data and Process Quality in Multi-Input, Multi-Output Information Systems. *Management Science*, 31(2).

Ballou, P., Wang, Y., Pazer, H. & Tayi, K., 1998. Modeling Information Manufacturing Systems to Determine Information Product Quality. *Management Science*, 44(4).

Barab, S. & Squire, D., 2004. Design-based research: Putting our stake in the ground..

Journal of the Learning Science, 1(13), pp. 1-14.

Baskerville, R. & Wood-Harper, T., 1996. A critical perspective on action research as a method for information systems research. *Journal of Information Technology*.

Batini, C., Cappiello, C., Francalanci, C. & Maurino, A., 2009. Methodologies for data quality assessment and improvement. 41(3).

Blake, R. & Mangiameli, P., 2011. The effects and interactions of data quality and problem complexity on classification.. 2(2).

Braun, C., Wortmann, F., Hafner, M. & Winter, R., 2005. *Method construction - a core approach to organizational engineering*. New York, ACM.

Brock, E., 2000. A general treatment of dynamic integrity constraints. *Data & Knowledge Engineering*, Issue 32, pp. 223-246.

Brunson, D., 2005. *Certified Data and the Certification Process for Financial Institutions*.

[Online]

Available at: <http://searchdatamanagement.techtarget.com/news/2240111233/Certified-Data-and-the-Certification-Process-for-Financial-Institutions>

[Accessed 10 August 2013].

Buneman, O. & Clemons, E., 1979. *Efficiently Monitoring Relational Databases*. s.l., ACM Transactions on Database Systems.

Caballero, I., Verbo, E., Calero, C. & Piattini, M., 2006. *A data quality measurement information model based on iso/iec 15939*. Cambridge, USA: MIT, s.n.

Checkland, P., 1985. Achieving 'desirable and feasible' change: an application of soft system methodology. *Journal of the Operational Research Society*, 9(36), pp. 821-831.

Chengular-Smith, I., Ballou, D. & Pazer, H., 1999. *The impact of data quality information on decision making: An exploratory analysis*. s.l., IEEE Trans. Knowl. Data Eng.

Collins, J. & Hussey, R., 2002. *Business Research: A Practical Guide for Undergraduate and Postgraduate Students*. s.l.:s.n.

Decker, H., 2009. *Modelling and Monitoring the quality o data by integrity constraints and integrity checking*. s.l., International Conference on Software and Data Technology.

Delone, W. H. & McLean, E. R., 1992. Information Systems Success: The Quest for the Dependent Variable. 3(1).

DeLone, W. & McLean, E., 1992. Information Systems Success: The Quest for the Dependent Variable. 3(1).

Denzin, N. & Lincoln, Y., 2005. *The SAGE handbook of qualitative research*. s.l.:Sage Publications, Inc..

Doan, A. & Halevy, A., 2005. Semantic-Integration research in the database community: A brief survey. 26(1).

English, L., 1999. *Improving data warehouse and business information quality: methods for reducing cost and increasing profits*. 1st ed. s.l.:Wiley.

Fan, W., Lu, H., Madnick, S. & Cheung, D., 2001. *Discovering and reconciling data value conflicts for numerical data integration*. s.l., s.n.

Fellegi I, S. A., 1969. A Theory for Record Linkage. *Journal of the American Statistical Association*.

Fisher, C., Chengular-Smith, I. & Ballou, D., 2003. *The impact of experience and time on the use of data quality information in decision making*. s.l., Inf. Syst. Res.

Foley, O., 2011. *Information Quality and Diverse Information Systems Situations*, s.l.: PhD thesis, Dublin City University, Ireland.

Furber, C. & Hepp, M., 2011. *Towards a vocabulary for data quality management in semantic web architectures*. Uppsala, ACM.

- Galahards, H. et al., 2001. *Declarative data cleaning: Language, model and algorithms*. s.l., s.n.
- Galliers, R., 1992. *Choosing Information Systems Research Approaches. Information Systems Research: Issues, methods and practical guidelines*. s.l.:Blackwell Scientific Publications, pp..
- Ge, M., 2009. *Information Quality Assessment and Effects on Inventory Decision making*, s.l.: PhD thesis, Dublin City University, Ireland.
- Ge, M., Helfert, M. & Jannach, D., 2011. *Information Quality Assessment: Validating Measurement Dimensions and Processes*. Helsinki, Finland, s.n.
- Gilmore, T., Krantz, J. & Ramirez, R., 1986. *Action Based Modes of Inquiry and the Host-Researcher Relationship*, s.l.: s.n.
- Goh, C., Bressan, S., Madnick, S. & Siegel, M., 1999. Context interchange: New features and formalisms for the intelligent integration of information. 17(3).
- Goodhue, D., 1995. Understanding User Evaluations of Information Systems. 41(12).
- Goodhue, D. H., 1995. Understanding User Evaluations of Information Systems. 41(12).
- Henver, A., March, S. & Park, J., 2004. Design Science in Information System Research. *MIS Quarterly*, March.pp. 75-105.
- Herbert, K. et al., 2004. *BIO-AJAX: An extensible framework for biological data cleaning*. s.l., SIGMOD.
- Hernandez, M., 1998. Real-World data is dirty: Data cleansing and the merge/purge problem. 2(1).

- Huang, J. et al., 2010. *Lightweight problem determination in DBMSs using data stream analysis techniques*. New York, s.n.
- Iivar, J., 2007. A Paradigmatic Analysis of Information Systems as a Design Science. *Scandinavian Journal of Information System*, 19(2), pp. 39-63.
- Ivan, I., Pocatilu, P., Mihai, T. & Stanca, C., 2000. *Data Certification*. Boston, MIT.
- Jarke, M. & Vassiliou, Y., 1997. Data Warehouse Quality: A Review of the DWQ Project.
- Juha-Pekk, T., 1998. *Incremental Method Engineering with Modeling Tools: Theoretical Principles and Empirical Evidence*, s.l.: University of Jyväskylä.
- Jung, W., Olfman, L., Ryan, T. & Park, Y., 2005. *An experimental study of the effects of contextual data quality and task complexity on decision performance*. s.l., s.n.
- Keeton, K., Mehra, P. & Wilkes, J., 2009. Do you know your IQ?: a research agenda for information quality in systems. 37(3).
- Khan, B., Strong, D. & Wang, R., 1998. *Product and service performance model for information quality: an update*. s.l., s.n.
- Khan, K. M., Kapurubandara, M. & Chadha, U., 2004. *Incorporating business requirements and constraints in database conceptual models*. s.l., s.n., pp. 59 - 64.
- Klein, B., Goodhue, D. & Davis, G., 1997. Can humans detect errors in data? Impact of base rates, incentives, and goals. 21(2).
- Kovacic, A., 2004. Business renovation: business rules (still) the missing link. *Business Process Management*, pp. 158-170.

- Kovacic, A., 2004. Business renovation: business rules (still) the missing link. *Business Process Management*, pp. 158-170.
- Kumar, R., 1996. *Research Methodology*. 2nd ed. s.l.:Addison Wesley Longman Australia Pty Limited, London, UK..
- Lee, Y., 2004. *Crafting rules: Context-reflective data quality problem solving*. s.l., s.n.
- Lee, Y., Pipino, L., Fund, J. & Wang, R., 2006. *Journey to Data Quality*. s.l., The MIT Press.
- Lee, Y. & Strong, D., 2004. *Knowing-Why about data processes and data quality*. s.l., J. Manag. Inf. Syst.
- Lee, Y., Strong, D., Khan, B. & Wang, R., 2002. A methodology for information quality assessment. 40(2).
- Lee, Y. W., Strong, D. M., Khan, B. K. & Wang, R. Y., 2002. A methodology for information quality assessment. 40(2).
- Madnick, S., Wang, R., Lee, Y. & Zhu, H., 2009. Overview and Framework for Data and Information Quality Research. 1(1).
- Madnick, S. & Zhu, H., 2006. Improving data quality with effective use of data semantics. 59(2).
- Manion, L., Cohen, L. & Morrison, K., 2000. *Research Methods in Education*. s.l.:Routledge Famer.
- March, S. & Smith, G., 1995. Design and Natural science research on Information Technology. *Decision Support System*, pp. 251-266.

Martyn, S., 2008. *Quantitative research design*. [Online]

Available at: <http://www.experiment-resources.com/quantitative-research-design.html>

[Accessed 1 June 2010].

McCune, W. & Henschen, L., 1989. Maintaining State Constraints in Relational Databases: A Proof Theoretic Basis. *Journal of the Association for Computing Machinery*, pp. 46-68.

Mikkelsen, G. & Aasly, J., 2005. *Consequences of impaired data quality on information retrieval in electronic patient records*. s.l., Int. J. Med. Inf.

Missier, P., Oliaro, A. & Raffa, S., 2006. *Practical data quality certification: model, architecture, and experience*. Chicago, s.n.

Muehlen, M. z., Indulska, M. & Kamp, G., 2007. *Business process and business rule modeling languages for compliance management: a representational analysis*. s.l., s.n., pp. 127-132.

Nelson, M. L., Rariden, R. L. & Sen, R., 2008. *A Lifecycle Approach toward Business Rules Management*. s.l., s.n.

Peffer, K., Tuunanen, T., Rothenberger, M. & Chatterjee, S., 2007. A Design Science Research Methodology for Information System Research. *Journal of Management Information System*, pp. 45-77.

Pham Thi, T. & Helfert, M., 2007a. *An Information System Quality Framework based on Information System Architectures*. Galway, International Conference on Information Systems Development.

Pham Thi, T. & Helfert, M., 2007b. The IASDO Model for Information Manufacturing System Modelling. *International Journal of Information Quality*, pp. 5-21.

Pham, . T. T. T. & Helfart, M., 2007. *An Information System Quality Framework based on Information System Architectures*. Galway, International Conference on Information Systems Development.

Pipino, L., Lee, Y. & Wang, R., 2002. Data Quality Assessment. *Communications of the ACM*, 45(4).

Pipino, L. L., Lee, Y. W. & Wang, R. Y., 2002. Data Quality Assessment. *Communications of the ACM*, 45(4).

Pries-Heje, J., Baskerville, R. & Venable, J., 2008. *Strategies for Design Science Research Evaluation*. Galway, s.n.

Ragnathan, S., 1999. *Impact of information quality and decision-making quality on decision quality: A theoretical model..* s.l., Decision Support Syst.

Rahm, E. & Bernstein, P., 2001. Onmatching schemas automatically. 10(4).

Redman, T., 1996. *Data Quality for the information age*. s.l.:Artech House.

Redman, T., 1998. *The impact of poor data quality on the typical enterprise*. s.l., Commun. ACM.

Redman, T., 2001. *Data Quality: The Field Guide*. s.l.:Digital Press.

Scannapieco, M. et al., 2004. The daquincis architecture: a platform for exchanging and improving data quality in cooperative information systems. *Inf. Syst*, 29(7), pp. 551-582.

Schelp, J. & Winter, R., 2006. Method Engineering: Lesson learned from reference modeling. *Design Science in Information Systems Research*.

Sein, M. et al., 2011. Action Design Research. *MIS Quarterly*, March, 35(1), pp. 37-55.

- Shankaranarayanan, G., Wang, R. Y. & Ziad, M., 2000. *M. IP-Map: Representing the manufacture of an information product*. s.l., s.n.
- Shankaranarayanan, G., Wang, R. & Ziad, M., 2000. *M. IP-Map: Representing the manufacture of an information product*. s.l., s.n.
- Sheng, Y. & Mykytyn, P., 2002. *Information technology investment and firm performance: A perspective of data quality..* s.l., s.n.
- Slone, J., 2006. *Information quality strategy: An empirical investigation of the relationship between information quality improvements and organizational outcomes*. s.l., Ph.D. dissertation, Capella University.
- Storey, V. & Wang, R., 1998. *Modeling quality requirements in conceptual database design*. s.l., s.n.
- Strong, D., Lee, Y. & Wang, R., 1997. Data Quality in Context. 40(5).
- Susman, G., 1983. *Action Research: A Sociotechnical Systems Perspectiven*. London: Sage Publications.
- Takeda, H., Veerkamp, P., Tomiyama, T. & Yoshikawam, H., 1990. Modeling Design Processes. pp. 37-48.
- Vaishnavi, V. & Kuechler, W., 2008. On Theory Development in Design Science Research: Anatomy of a Research Project. October, 15(5), pp. 489-504.
- Vasilecas, O. & Smaizys, A., 2006. *The framework: an approach to support business rule based data analysis*. s.l., s.n., pp. 141 - 147.
- Vianu, V., 1983. Dynamic Constraints and Database Evolution. *ACM*, September. pp. 389-399.

- Wand, Y. & Wang, R., 1996. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), pp. 86 - 95.
- Wand, Y. & Wang, R. Y., 1996. Anchoring data quality dimensions in ontological foundations. *Communications of the ACM*, 39(11), pp. 86 - 95.
- Wang, R., 1998. A Product Perspective on Total Data Quality Management. *Communications of the ACM*, 41(2).
- Wang, R., Kon, H. & Madnick, S., 1993. Data Quality Requirement Analysis and Modeling. *Ninth International Conference of Data Engineering*.
- Wang, R., Lee, Y., Pipino, L. & Strong, D., 1998. Manage Your Information as a Product. *Sloan Management Review*.
- Wang, R., Lee, Y. & Ziad, M., 2001. *Data Quality*. s.l.:Springer.
- Wang, R. & Strong, D., 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, Spring, 12(4), pp. 5-33.
- Wang, R. W. & Strong, D. M., 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems*, Spring, 12(4), pp. 5-33.
- Winter, R., 1989. *Learning From Experience: Principles and Practice in Action-Research*. Philadelphia: The Falmer Press.
- Wynekoop, J. & Russo, N., 1993. *System Development Methodologies: Unanswered questions and the research-practice gap*. s.l., s.n., pp. 181-190.
- Xu, H., Nord, J., Brown, N. & Nord, G., 2002. *Data quality issues in implementing an ERP..* s.l., Industrial Manag. Data Syst..

Zmud, R., 1978. Concepts, Theories and Techniques: An Empirical Investigation of the Dimensionality of the Concept of Information. 9(2).